Journal of Statistical Planning and Inference 140 (2010) 139 – 152

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Nonparametric density estimation for multivariate bounded data

Taoufik Bouezmarni[a], Jeroen V.K. Rombouts[b,*]

[a] *Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale Centre-ville Montréal, Canada H3C 3J7*
[b] *Institute of Applied Economics at HEC Montréal, CIRANO, CIRPEE, Université catholique de Louvain (CORE, Belgium), CREF, 3000 Côte Sainte Catherine, Montréal (QC), Canada H3T 2A7*

ARTICLE INFO

Article history:
Received 7 April 2008
Received in revised form
25 March 2009
Accepted 2 July 2009
Available online 17 July 2006

Keywords:
Asymmetric kernels
Multivariate boundary bias
Nonparametric multivariate density estimation
Asymptotic properties
Bandwidth selection
Least squares cross-validation

ABSTRACT

We propose a new nonparametric estimator for the density function of multivariate bounded data. As frequently observed in practice, the variables may be partially bounded (e.g. nonnegative) or completely bounded (e.g. in the unit interval). In addition, the variables may have a point mass. We reduce the conditions on the underlying density to a minimum by proposing a nonparametric approach. By using a gamma, a beta, or a local linear kernel (also called boundary kernels), in a product kernel, the suggested estimator becomes simple in implementation and robust to the well known boundary bias problem. We investigate the mean integrated squared error properties, including the rate of convergence, uniform strong consistency and asymptotic normality. We establish consistency of the least squares cross-validation method to select optimal bandwidth parameters. A detailed simulation study investigates the performance of the estimators. Applications using lottery and corporate finance data are provided.

1. Introduction

Among multivariate nonparametric density estimators, the standard Gaussian kernel is the most popular. The estimator has excellent asymptotic properties; see Silverman (1986), Scott (1992), and Wand and Jones (1995) for more details. However, the estimator does not take into account the potential finite support of the variables. When the support of some variables is bounded, for example, in the case of nonnegative data, the standard kernel estimator continues to give weight outside the supports. This causes a bias in the boundary region. The boundary bias problem of the standard kernel is well documented in the univariate case. An initial solution to the boundary problem is given by Schuster (1985), who proposes the reflection method. Müller (1991), Lejeune and Sarda (1992), Jones (1993), Jones and Foster (1996), and Cheng et al. (1997) suggest the use of adaptive and boundary kernels at the edges and a fixed standard kernel in the interior region. Marron and Ruppert (1994) investigate some transformations before using the standard kernels, and Cowling and Hall (1996) propose a pseudodata method. Recently, Chen (2000), Bouezmarni and Scaillet (2003), and Bouezmarni and Rombouts (2006) study the gamma kernels for univariate nonnegative data. For data defined on the unit interval, Chen (1999) proposes to use a beta kernel.

Although the consequences of the boundary problem in multivariate dimensions are much more severe, because the boundary region increases with dimension, solutions to the problem are not well investigated. Müller and Stadtmüller (1999) propose boundary kernels for multivariate data defined on arbitrary support by selecting the kernels that minimize a variational problem. In fact, they extend the minimum variance selection principle kernel used to select the optimal kernel in the interior region, as in Epanechnikov (1969) and Granovsky and Müller (1991). Although this estimator has interesting properties, it remains complicated in practice. In addition, it requires an additional bandwidth parameter and a weighting function. In the nonparametric regression

* Corresponding author.
E-mail address: jeroen.rombouts@hec.ca (J.V.K. Rombouts).

0378-3758/$ - see front matter Crown Copyright © 2009 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.07.013

context, the problem of boundary bias is developed by Gasser et al. (1985) and Zhang et al. (1999) for the univariate case, and Fan and Gijbels (1992), Ruppert (1994), Staniswalis et al. (1993), and Staniswalis and Messer (1996) for multivariate data.

This paper proposes a nonparametric product kernel estimator for density functions of multivariate bounded data. Estimation is based on a gamma kernel or a local linear kernel when the support of the variable is nonnegative and a beta kernel when the support is a compact set. Based on these kernels, the density estimators are robust to the boundary problem. The method is easy in conception and implementation. We provide the asymptotic properties of these estimators and show that the optimal rate of convergence of the mean integrated squared error is obtained. For the multivariate uniform density, we show that the estimator we propose using beta kernels is asymptotically unbiased. We examine the finite sample performance in several simulations. As for any nonparametric kernel estimator, the performance is sensitive to the choice of the bandwidth parameters. We suggest the application of the least squares cross-validation method to select these parameters. We prove the consistency of this method for the proposed estimators and investigate its performance in the simulations.

The rest of the paper is organized as follows. We introduce the multivariate nonparametric estimator for multivariate bounded data in Section 2. Section 3 provides convergence properties. The consistency of the least squares cross-validation bandwidth selection method is established in Section 4. In Section 5 we investigate the finite sample properties of several kernel estimators for nonnegative bivariate data. Section 6 contains two applications, one with lottery data and another with corporate finance data. Section 7 concludes. The proofs of the theorems are presented in the Appendix.

2. Nonparametric estimator

Let $\{(X_{i1},\dots,X_{id}), i=1,\dots,n\}$ be a sample of independent and identically distributed random variables with an unknown density function f. The general multivariate nonparametric density estimator is given by

$$\hat{f}(x)=\hat{f}(x_1,\dots,x_d)=\frac{1}{nh_1,\dots,h_d}\sum_{i=1}^{n}\mathbf{K}\left(\frac{x_1-X_{i1}}{h_1},\dots,\frac{x_d-X_{id}}{h_d}\right),$$

where $\mathbf{K}$ denotes a multivariate kernel function and $(h_1,\dots,h_d)$ the vector of bandwidth parameters.

In practice the choice of $\mathbf{K}$ is especially difficult when the supports of the random variables are potentially unequal. Therefore, we propose to use the product kernel estimator with adapted and flexible kernels in order to solve the boundary bias problem. The estimator is defined as

$$\hat{f}(x_1,\dots,x_d)=\frac{1}{n}\sum_{i=1}^{n}\prod_{s=1}^{d}K^s(h_s,X_{is})(x_s), \tag{1}$$

where $h_1,\dots,h_d$ are the bandwidth parameters and the kernel K^s is a kernel for variable s. Throughout the paper, this superscript s will be omitted for notational convenience. As described in the Introduction, the kernel for each variable is chosen to be the Gaussian kernel, which is indeed optimal when the total support is $\mathbb{R}^d$. We consider two cases for random variables with bounded support.

First, when the support of the variable is nonnegative, we propose the use of either the local linear kernel denoted by K_L or one of the two gamma kernels K_G, K_{NG} as shown below. Thus,

$$K_L(h,t)(z)=K_l\left(z,h,\frac{z-t}{h}\right),$$

where

$$K_l(z,h,y)=\frac{a_2(z,h)-a_1(z,h)y}{a_0(z,h)a_2(z,h)-a_1^2(z,h)}K(y),$$

K is any symmetric kernel with a compact support $[-1,1]$. Throughout the paper we take the Epanechnikov kernel and

$$a_s(z,h)=\int_{-1}^{z/h}z^sK(z)\,dz.$$

For more details on the local linear kernel, see Lejeune and Sarda (1992), Jones (1993), and Neilsen (1999).

The kernels K_G and K_{NG}, for which we denote the bandwidth by b, are respectively defined as

$$K_G(b,t)(z)=\frac{t^{z/b}\exp(-t/b)}{b^{z/b+1}\Gamma(z/b+1)}$$

and

$$K_{NG}(b,t)(z)=\frac{t^{\rho(z)-1}\exp(-t/b)}{b^{\rho(z)}\Gamma(\rho(z))},$$