Computational Statistics and Data Analysis 63 (2013) 1–15

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Generalized Birnbaum–Saunders kernel density estimators and an analysis of financial data

Carolina Marchant [a], Karine Bertin [b], Víctor Leiva [a,*], Helton Saulo [c]

[a] Departamento de Estadística, Universidad de Valparaíso, Chile
[b] CIMFAV, Universidad de Valparaíso, Chile
[c] Departamento de Economia, Universidade Federal do Rio Grande do Sul, Brazil

ARTICLE INFO

Article history:
Received 4 June 2012
Received in revised form 16 January 2013
Accepted 16 January 2013
Available online 23 January 2013

Keywords:
High frequency data
Kernel nonparametric method
Kurtosis
Monte Carlo method
R statistical computation language

ABSTRACT

The kernel method is a nonparametric procedure used to estimate densities with support in $\mathbb{R}$. When nonnegative data are modeled, the classical kernel density estimator presents a bias problem in the neighborhood of zero. Several methods have been developed to reduce this bias, which include the boundary kernel, data transformation and reflection methods. An alternative proposal is to use kernel estimators based on distributions with nonnegative support, as is the case of the Birnbaum–Saunders (BS), gamma, inverse Gaussian and log-normal models. Generalized BS (GBS) distributions have received considerable attention, due to their properties and their flexibility in modeling different types of data. In this paper, we propose, characterize and implement the kernel method based on GBS distributions to estimate densities with nonnegative support. In addition, we provide a simple method to choose the corresponding bandwidth. In order to evaluate the performance of these new estimators, we conduct a Monte Carlo simulation study. The obtained results are illustrated by analyzing financial real data.

1. Introduction

The distribution of a continuous random variable can be characterized through (i) its probability density function (or simply density), (ii) its cumulative distribution function or (iii) its quantile function, among others, being the density the most used. Then, a relevant problem in statistics is to estimate this density, which can be performed by using parametric or nonparametric methods.

The present paper deals with nonparametric estimation of a density function $f: [0, +\infty) \to \mathbb{R}_+$, based on a random sample (that is, independently and identically distributed random variables) $X_1, \ldots, X_n$. Rosenblatt (1956) and Parzen (1962) introduced nonparametric kernel density estimators, that, in the sequel, we call classical kernel estimators. When nonnegative data are modeled, these estimators lead to a bias problem in the neighborhood of zero. Several methods have been proposed to solve this bias problem, such as the boundary kernel method (see Jones, 1993; Gasser et al., 1985; Zhang and Karunamuni, 2000), the local linear method (see Lejeune and Sarda, 1992; Cheng, 1997; Cheng et al., 1997; Zhang and Karunamuni, 1998), the local renormalization method (see Diggle, 1985; Härdle, 1990), the pseudo-data method (see Cowling and Hall, 1996), the reflection method (see Schuster, 1985; Cline and Hart, 1991), and the transformation

* Correspondence to: Departamento de Estadística, Universidad de Valparaíso, Gran Bretaña 1111, Playa Ancha, Valparaíso, Chile. Tel.: +56 322508320.
E-mail addresses: carolina.marchant@uv.cl (C. Marchant), karine.bertin@uv.cl (K. Bertin), victor.leiva@uv.cl, victor.leiva@yahoo.com (V. Leiva), heltonsaulo@gmail.com (H. Saulo).
URL: http://www.deuv.cl/leiva (V. Leiva).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2013.01.013

method (see Wand et al., 1991; Marron and Ruppert, 1994). For other references and methods about this topic, see Hall and Wehrly (1991), Müller (1991), Silverman (1992), Müller and Wang (1994), Zhang et al. (1999), Hall and Park (2002) and Karunamuni and Alberts (2005, 2006). The reflection method aims to recapture the missing weight by reflecting the estimate in the boundary. Renormalized Gaussian kernels are used in the renormalization method. In the transformation method, classical kernel estimators are applied to a transformation of the data; see Bouezmarni and Rombouts (2010). Recently, another approach has been introduced by Comte and Genon-Catalot (2012). The authors proposed kernel density estimators free of boundary effects near zero, by using densities of empirical means of independent, identically distributed, nonnegative random variables. These estimators can be adapted to any order of regularity of the density. The authors used data-driven choice of bandwidth. Kernel methods with local or global adaptive bandwidth selection have been also proposed; see Tsybakov (2009) and Goldenshluger and Lepski (2011).

An alternative way to solve the aforementioned boundary bias problem is to use nonparametric estimators based on kernels of asymmetric distributions with nonnegative support. This type of estimators has been first introduced by Chen (1999) for densities with support on the interval $[0, 1]$. Unlike the classical kernel estimators, the asymmetric kernel estimators do not engender boundary bias, because the asymmetric kernel nonparametric method never assigns weight outside the density support, which coincides with the support of the kernel, so that it produces better estimates; see Fernandes and Monteiro (2005). Also, the shape of the asymmetric kernels varies, allowing different degrees of smoothing to be obtained, according to the position where the estimation is performed. Other estimators of densities with support on the interval $[0, \infty)$ have been proposed by using kernels based on the Birnbaum–Saunders (BS), gamma (GAM), inverse Gaussian (IG), lognormal (LN) and reciprocal IG (RIG) distributions; see Chen (2000), Jin and Kawczak (2003) and Scaillet (2004). GAM and IG kernel estimators are particular cases of convolution power kernels estimators (see Comte and Genon-Catalot, 2012), but it is not the case of the BS kernel estimator, because a sum of independent random variables with such a distribution is no longer a BS distribution; see Volodin and Dzhungurova (2000) and Leiva et al. (2011). The rate of convergence of the density estimators based on the BS, GAM, IG, LN and RIG kernels is the same as that from the classical kernel estimators, under the assumption that the unknown true density is twice differentiable.

As aforementioned, a good reason for using kernel estimators based on an asymmetric density with nonnegative support is that they solve the boundary bias problem. A further reason is that these estimators usually inherit the characteristics of their kernels. For instance, if a density is highly skewed, then its estimation may be sensitive to the selected asymmetric kernel. As a consequence of it, some benefits can be obtained, for example, when the mode and quantiles located at the tails of such a density are estimated. Also, it is important to locate the mode of possibly asymmetric densities with precision, when one wants to maximize a smoothed profile likelihood function with respect to a parameter of interest; see Silverman et al. (1990) and Abadir and Lawford (2004).

As also aforementioned, one of the asymmetric distributions, with nonnegative support, used by Jin and Kawczak (2003) as kernel for estimating nonparametrically the density, is the BS distribution, which has been widely studied in recent times; see Birnbaum and Saunders (1969), Johnson et al. (1995) and Leiva et al. (2007). Díaz-García and Leiva (2005) generalized the BS distribution, obtaining a wider class of positively skewed densities with nonnegative support that possesses lighter and heavier tails than the BS distribution. Thus, generalized BS (GBS) distributions are essentially flexible in the kurtosis level; see also Sanhueza et al. (2008). The GBS family has, as particular cases, the BS-classical, BS-power-exponential (BS-PE) and BS-Student-t (BS-t) distributions.

Some applications of BS models to financial data are attributed to Jin and Kawczak (2003), Bhatti (2010) and Paula et al. (2012). Particularly, Jin and Kawczak (2003) and Bhatti (2010) used the BS distribution for analyzing high frequency financial (HFF) data, which have unique features absent in data with low frequencies. These features are, among others, (A1) the big number of observations of this type of data (e.g., the average daily number of quotes in European spot markets could easily exceed 20 000 and this number can be even higher in the New York Stock Exchange (NYSE)); (A2) the time between transactions on trades and quotes are irregularly spaced with random daily numbers of observations; (A3) high flows of transactions when markets are beginning or when these are closing; (A4) the failure rate of HFF data is unimodal; and (A5) intraday price volatility, which is serially correlated; see McInish and Wood (1991) and McInish (1992). Then, analysis of HFF data raise several econometric and statistical challenges; see Engle (2000). Such as Jin and Kawczak (2003) and Bhatti (2010) did with the BS distribution, we believe that GBS kernel methods can be appropriate for estimating densities of HFF data, because of their flexibility and properties, some of them shared by this type of data, such as heavy-tails and unimodal failure rates. In this article, for illustrative purposes, we analyze real data obtained from the Trades and Quotes (TAQ) database of the NYSE, which contains all the transactions listed in NYSE of the National Association of Securities Dealers Automated Quotation (NASDAQ) and of the American Stock Exchange (AMEX).

The objectives of this work are (i) to propose kernel estimators for densities with nonnegative support based on GBS distributions; (ii) to determine the statistical properties of these estimators; (iii) to provide a first and simple method to choose the corresponding bandwidth; (iv) to computationally implement the proposed estimation and bandwidth selection methods; (v) to numerically evaluate and compare the performance of the proposed estimators by a simulation study; and (vi) to perform a HFF data analysis based on the proposed methods.

This article is organized as follows. In Section 2, we present a description of the GBS distributions, propose the kernel estimators based on these distributions, determine the statistical properties of these estimators, and provide a bandwidth selection method. In Section 3, we computationally evaluate the performance of the postulated nonparametric estimators by means of a Monte Carlo (MC) simulation study. In Section 4, we illustrate the proposed methods by analyzing two real