Computational Statistics and Data Analysis 56 (2012) 732–740

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian adaptive bandwidth kernel density estimation of irregular multivariate distributions

Shuowen Hu, D.S. Poskitt, Xibin Zhang *

Department of Econometrics and Business Statistics, Monash University, Australia

ARTICLE INFO

Article history:
Received 21 January 2011
Received in revised form 25 July 2011
Accepted 16 September 2011
Available online 29 September 2011

Keywords:
Marginal likelihood
Markov chain Monte Carlo
S&P 500 index
Value-at-risk

ABSTRACT

In this paper, we propose a new methodology for multivariate kernel density estimation in which data are categorized into low- and high-density regions as an underlying mechanism for assigning adaptive bandwidths. We derive the posterior density of the bandwidth parameters via the Kullback–Leibler divergence criterion and use a Markov chain Monte Carlo (MCMC) sampling algorithm to estimate the adaptive bandwidths. The resulting estimator is referred to as the tail-adaptive density estimator. Monte Carlo simulation results show that the tail-adaptive density estimator outperforms the global-bandwidth density estimators implemented using different global bandwidth selection rules. The inferential potential of the tail-adaptive density estimator is demonstrated by employing the estimator to estimate the bivariate density of daily index returns observed from the USA and Australian stock markets.

1. Introduction

Kernel density estimation is one of the most important techniques for understanding the distributional properties of data. It is understood that the effectiveness of such approach depends on the choice of a kernel function and the choice of a smoothing parameter or bandwidth (see for example, Izenman, 1991, for a discussion). Although the two issues cannot be treated separately, it is widely accepted that the performance of a kernel density estimator is mainly determined by the bandwidth (see for example, Scott, 1992; Wand and Jones, 1995), while the impact of kernel choices on the performance of the resulting density estimator was examined by Marron and Nolan (1988), Vieu (1999) and Horová et al. (2002). Most investigations have aimed at choosing a fixed or global bandwidth for a full sample of data (see Jones et al., 1996, for a survey). Terrell and Scott (1992) and Sain and Scott (1996) proposed the idea of data-driven adaptive bandwidth density estimation, which allows the bandwidth to vary at different data points. In this situation, bandwidth selection remains as an important issue and has been extensively investigated for univariate data. However, less attention has been paid to investigations of data-driven methods for estimating adaptive bandwidths in multivariate data. This motivates the investigation of this paper.

Our investigation is also empirically motivated. Most financial analysts believe that during the course of the global financial crisis caused by the fall-out of the USA sub-prime mortgage crisis, the USA stock market has had a leading effect on most other stock markets world wide. Using a kernel density estimator of bivariate stock-index returns we can derive the conditional distribution of the stock-index return in one market for a given value of the stock-index return in the USA market, and therefore we can better understand how the former market was associated with the USA market. However, the marginal density of stock-index returns often exhibits leptokurtosis. Consequently, the kernel estimation of the bivariate density of stock-index returns may require different bandwidths to different groups of observed returns.

* Correspondence to: Department of Econometrics and Business Statistics, Monash University, 900 Dandenong Road, Caulfield East, Victoria 3145, Australia. Tel.: +61 3 99032130; fax: +61 3 99032007.
E-mail address: xibin.zhang@monash.edu (X. Zhang).
URL: http://users.monash.edu.au/~xzhang/ (X. Zhang).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.09.022

1.1. Some background

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)'$ denote a d-dimensional random vector with its density function $f(\boldsymbol{x})$ defined on $\boldsymbol{R}^d$ (see Jácome et al., 2008, for an example of censored data). Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ be a random sample drawn from $f(\boldsymbol{x})$. The kernel density estimator of $f(\boldsymbol{x})$ is (Wand and Jones, 1995)

$$\hat{f}_H(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_H(\boldsymbol{x} - \boldsymbol{x_i}) = \frac{1}{n|H|^{1/2}}\sum_{i=1}^{n} K(H^{-1/2}(\boldsymbol{x} - \boldsymbol{x_i})), \tag{1}$$

where $K(\cdot)$ is a multivariate kernel, and H is a symmetric and positive definite $d \times d$ matrix known as the bandwidth matrix. To choose an optimal H, several methods have been discussed in literature (see for example, Devroye and Györfi, 1985; Marron, 1987). Marron (1992) proposed a bootstrapping approach to the approximation of the mean integrated squared error (MISE), and the normal reference rule (NRR) or equivalently the rule-of-thumb that minimizes the asymptotic MISE was discussed by Scott (1992). The least squares cross-validation was discussed by Sain et al. (1994) and Duong and Hazelton (2005), and a plug-in method was suggested by Wand and Jones (1994) and improved by Duong and Hazelton (2003). Moreover, Zhang et al. (2006) proposed a Bayesian approach to the estimation of the bandwidth matrix based on Kullback–Leibler information criterion.

However, a problem in using a global bandwidth is that the kernel methods often produce unsatisfactory results for complex or irregular densities. Sain and Scott (1996) presented a classic example, in which a bimodal mixture with two modes of equal height but different levels of variation, where an optimal global bandwidth tends to under-smooth the mode with a large variation and over-smooth the mode with a small variation. Therefore, it is necessary to let the bandwidth be adaptive to local observations; a relatively small bandwidth is needed where observations are densely distributed, and a large bandwidth is required where observations are sparsely distributed (see Jones, 1990; Wand and Jones, 1995; Sain, 2002, for similar arguments).

Several versions of the adaptive bandwidth kernel density estimator have been studied in literature. Mielniczuk et al. (1989) proposed to use a weighting function on data point $\boldsymbol{x}$ in a global bandwidth estimator. A popular method is to make the bandwidth as a function of data points, and Nolan and Marron (1989) discussed this issue from a general Delta-sequence estimator perspective. Loftsgaarden and Quesenberry (1965) suggested to replace H in (1) is by $H(\boldsymbol{x})$, which is the bandwidth at data point $\boldsymbol{x}$. This estimator was studied by Cao (2001) and Sain and Scott (2002) and is also called the balloon estimator. However, the balloon estimator does not integrate to one and therefore is not a good choice for density estimation (Terrell and Scott, 1992; Izenman, 1991). The other estimator proposed by Breiman et al. (1977) is called the sample-point estimator, which employs $H(\boldsymbol{x_i})$ as the bandwidth associated with sample point $\boldsymbol{x_i}$. Abramson (1982a,b) suggested to choose bandwidth as the inverse square root of $f(\boldsymbol{x_i})$. As the sample-point estimator always integrate to one, we consider this estimator throughout this paper.

The difficulty of using sample-point estimator is that the number of bandwidths exceeds the sample size in multivariate data, and this makes difficulties in estimating bandwidths. Sain and Scott (1996) and Sain (2002) suggested grouping data into different bins and using a constant bandwidth for each bin. Although this method reduces the number of bandwidths, the number of bandwidths still grows exponentially with the dimension.

1.2. The tail-adaptive density estimator

A simple method to control the number of bandwidths to be estimated for the sample-point kernel density estimator is to put the data into a small number of groups. In this paper, we propose dividing the observations into two regions, namely the low-density region (LDR) and high-density region (HDR), and assigning two different bandwidth matrices to these two regions. When the true density is unimodal, the low-density region is the tail area and should be assigned larger bandwidths than the high-density region. We call this type of kernel density estimator the *tail-adaptive density estimator*.

It is not new to group the observations into the low- and high-density regions. Hartigan (1975, 1987) proposed clustering data into different regions with different density values, and Hyndman (1996) presented an algorithm for computing and graphing data in different density regions. A comprehensive review of applications related to the issue of low-and high-density regions was given in Mason and Polonik (2009). In terms of bandwidth selection, Samworth and Wand (2010) considered a bandwidth selection method for univariate high-density region estimation. To our knowledge, there is yet any other investigation on bandwidth selection for a general multivariate kernel density estimation, where two different bandwidth matrices are assigned to observations in the low- and high-density regions.

In this paper, we treat the elements of the two bandwidth matrices as parameters, whose posterior can be approximately derived through the Kullback–Leibler information. Therefore, bandwidths can be estimated through a posterior simulator. During the past decade, there have been several investigations on Bayesian approaches to bandwidth estimation for kernel density estimation (see for example, Brewer, 2000; Gangopadhyay and Cheung, 2002; Kulasekera and Padgett, 2006; de Lima and Atuncar, 2010). In particular, Zhang et al. (2006) derived the posterior of bandwidths for multivariate kernel density estimation with a global bandwidth matrix. Their Monte Carlo simulation results reveal the advantage of this Bayesian approach over its competitors including the plug-in method of Duong and Hazelton (2003) and the NRR. However, Hall (1987) showed that the use of a global bandwidth for a long-tailed distribution can mislead Kullback–Leibler information. Therefore, in this paper, we extent the sampling algorithm proposed by Zhang et al. (2006) by incorporating tail-adaptive bandwidth matrices into the multivariate kernel density estimation.