Computational Statistics and Data Analysis 106 (2017) 27–45

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

FFT-based fast bandwidth selector for multivariate kernel density estimation[☆]

Artur Gramacki[a,*], Jarosław Gramacki[b]

[a] *Institute of Control and Computation Engineering, University of Zielona Góra, ul. Licealna 9, Zielona Góra 65-417, Poland*

[b] *Computer Center, University of Zielona Góra, ul. Licealna 9, Zielona Góra 65-417, Poland*

ARTICLE INFO

Article history:
Received 13 December 2015
Received in revised form 13 August 2016
Accepted 1 September 2016
Available online 11 September 2016

Keywords:
Multivariate kernel density estimation
Density derivative functionals
Bandwidth selection
Fast Fourier Transform
Nonparametric estimation

ABSTRACT

The performance of multivariate kernel density estimation (KDE) depends strongly on the choice of bandwidth matrix. The high computational cost required for its estimation provides a big motivation to develop fast and accurate methods. One of such methods is based on the Fast Fourier Transform. However, the currently available implementation works very well only for the univariate KDE and its multivariate extension suffers from a very serious limitation as it can accurately operate only with diagonal bandwidth matrices. A more general solution is presented where the above mentioned limitation is relaxed. Moreover, the presented solution can be easily adopted also for the task of efficient computation of integrated density derivative functionals involving an arbitrary derivative order. Consequently, bandwidth selection for kernel density derivative estimation is also supported. The practical usability of the new solution is demonstrated by comprehensive numerical simulations.

1. Introduction

Kernel density estimation (KDE) is a very important statistical technique with many practical applications. It has been applied successfully to both univariate and multivariate problems. There exists extensive literature on this issue, including several classical monographs, see Silverman (1986), Scott (1992) and Wand and Jones (1995).

A general form of the d-dimensional multivariate kernel density estimator is

$$\hat{f}(\boldsymbol{x}, \boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}\left(\boldsymbol{x} - \boldsymbol{X}_i\right), \tag{1}$$

where

$$K_{\boldsymbol{H}}(u) = |\boldsymbol{H}|^{-1/2} K\left(\boldsymbol{H}^{-1/2} u\right), \tag{2}$$

and $\boldsymbol{H}$ is the $d \times d$ *bandwidth* or *smoothing* matrix, d is the problem dimensionality, $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$, and $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T$, $i = 1, 2, \ldots, n$ is a sequence of independent identically distributed (i.i.d.) d-variate random variables

[☆] The up-to-date R source codes are included as a supplementary material. The figures and all the data shown in the tables can be replicated (see Appendix A).

[*] Corresponding author.
E-mail addresses: a.gramacki@issi.uz.zgora.pl (A. Gramacki), j.gramacki@ck.uz.zgora.pl (J. Gramacki).

http://dx.doi.org/10.1016/j.csda.2016.09.001

drawn from a (usually unknown) density function f. Here K and $K_{\boldsymbol{H}}$ are the unscaled and scaled kernels, respectively. In most cases the kernel has the form of a standard multivariate normal density.

The univariate kernel density estimator for a random sample $X_1, X_2, \ldots, X_n$ drawn from a common and usually unknown density function f is given by

$$\hat{f}(x, h) = \frac{1}{n} \sum_{i=1}^{n} K_h (x - X_i), \tag{3}$$

where

$$K_h(u) = h^{-1} K \left(h^{-1} u \right), \tag{4}$$

and h is the bandwidth which is a positive integer. The scaled (K_h) and unscaled (K) kernels are related in Eq. (4). Note that the notation used in Eq. (1) is not a direct extension of the univariate notation in Eq. (3), since in the one-dimensional case the bandwidth is $\boldsymbol{H} = h^2$, so we are dealing with 'squared bandwidths' here.

It seems that both uni- and multivariate KDE techniques have reached maturity and recent developments in this field are primarily focused on computational problems. There are two main computational problems related to KDE: (a) the fast evaluation of the kernel density estimates $\hat{f}$, and (b) the fast estimation (under certain criteria) of the optimal bandwidth matrix $\boldsymbol{H}$ (or scalar h in the univariate case). As for the first problem, a number of methods have been proposed, see for example Raykar et al. (2010) for a comprehensive review. As for the second problem, relatively less attention has been paid in the literature. An attempt of using the Message Passing Interface (MPI) was presented in Łukasik (2007). In Raykar and Duraiswami (2006) the authors give an ϵ-*exact* approximation algorithm, where the constant ϵ controls the desired arbitrary accuracy. Other techniques, like for example usage of Graphics Processing Units (GPUs), have also been used (Andrzejewski et al., 2013). In this paper we are concerned with fast estimation of the optimal bandwidth and are interested in the multivariate case only. However, our results can be easily adapted also to the univariate case.

It is obvious from Eq. (1) that the naive direct evaluation of the KDE at m evaluation points for n data points requires $O(mn)$ kernel evaluations. Evaluation points can be of course the same as data points and then the computational complexity is $O(n^2)$ making it very expensive, especially for large datasets and higher dimensions.

As for finding of the optimal bandwidth, the computational problems are even more evident. Typically, to complete all the required computations for this task a sort of numerical optimization is needed. Usually, the computational complexity of evaluating typical objective function is $O(n^2)$. During the optimization process the objective function must be evaluated many times (often more that a hundred or so), making the problem of finding the optimal bandwidth very expensive, even for moderate data dimensionalities and sizes.

In this paper we are concerned with an FFT-based method that was originally described in Wand (1994). In Wand and Jones (1995, appendix D) an interesting illustrative toy example has been presented. From now on this method will be called *Wand's algorithm*. It can be used for the KDE evaluation as well as for bandwidth selection problem and it works very well for the univariate case given in Eq. (3). Unfortunately, its multivariate extension does not support *unconstrained* bandwidth matrices (that is, if $\boldsymbol{H} \in \mathcal{F}$, where $\mathcal{F}$ is the set of all symmetric, positive definite $d \times d$ matrices). The method supports only more restricted *constrained* bandwidth matrices (that is, if $\boldsymbol{H} \in \mathcal{D}$, where $\mathcal{D}$ is the set of all positive definite diagonal matrices of the form $\boldsymbol{H} = diag(h_1^2, \ldots, h_d^2)$). This limitation was successfully overcome by the authors and the main results are presented in Gramacki and Gramacki (in press). In this paper we extend those results to the problem of fast (FFT-based) estimation of unconstrained bandwidth matrices.

To the best of our knowledge, our paper is the first where this problem is presented and successfully solved using an FFT-based approach. In this work we use excellent results presented in Chacón and Duong (2015), clearly the ones which significantly simplifies computations of integrated density derivative functionals (IDDF) involving an arbitrary derivative order (for details see Section 5). IDDFs are crucial elements in almost every modern bandwidth selection algorithm.

The remainder of the paper is organized as follows: in Section 3 we give an overview of the most popular and the most frequently used bandwidth selectors. In Section 2, based on a simple example, we demonstrate the problem. In Section 4 we give details of a complete FFT-based algorithm for fast estimation of unconstrained bandwidth matrices. To make the presentation of our algorithm clear, we do it on the basis of one of the simplest bandwidth selection algorithm, namely least square cross validation (LSCV). In Section 5 we extend our results also for the IDDFs. In Section 6 we give results from some numerical experiments based on both synthetic and real data sets. In Section 7 we conclude our paper.

2. Problem demonstration

As was mentioned in Section 1, Wand's algorithm does not support unconstrained bandwidth matrices, which considerably limits its practical usability. In this short demonstration we use a sample dataset *Unicef* presented in more detail in Section 6.2. In Fig. 1(a) we show the reference density when the bandwidth was obtained by direct (i.e., non-FFT-based) implementation of the LSCV algorithm, briefly presented in Section 3. After numerical minimization of the resulting objective function we get the sought bandwidth. In Fig. 1(b) one can see the behavior of Wand's original algorithm (i.e., FFT-based) when the minimization of the objective function proceeds over $\boldsymbol{H} \in \mathcal{F}$. The density is totally corrupted. In Fig. 1(c) we show the reference density when the bandwidth was obtained by direct (i.e., non-FFT-based) implementation of the