Computational Statistics and Data Analysis 57 (2013) 364–376

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Full bandwidth matrix selectors for gradient kernel density estimate

Ivana Horová [a,*], Jan Koláček [a], Kamila Vopatová [b]

[a] Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic

[b] Department of Econometrics, University of Defence, Brno, Czech Republic

ARTICLE INFO

Article history:
Received 4 July 2011
Received in revised form 2 July 2012
Accepted 5 July 2012
Available online 10 July 2012

Keywords:
Asymptotic mean integrated square error
Multivariate kernel density
Unconstrained bandwidth matrix

ABSTRACT

The most important factor in multivariate kernel density estimation is a choice of a bandwidth matrix. This choice is particularly important, because of its role in controlling both the amount and the direction of multivariate smoothing. Considerable attention has been paid to constrained parameterization of the bandwidth matrix such as a diagonal matrix or a pre-transformation of the data. A general multivariate kernel density derivative estimator has been investigated. Data-driven selectors of full bandwidth matrices for a density and its gradient are considered. The proposed method is based on an optimally balanced relation between the integrated variance and the integrated squared bias. The analysis of statistical properties shows the rationale of the proposed method. In order to compare this method with cross-validation and plug-in methods the relative rate of convergence is determined. The utility of the method is illustrated through a simulation study and real data applications.

1. Introduction

Kernel density estimates are one of the most popular nonparametric estimates. In a univariate case, these estimates depend on a bandwidth, which is a smoothing parameter controlling smoothness of an estimated curve and a kernel which is considered as a weight function. The choice of the smoothing parameter is a crucial problem in the kernel density estimation. The literature on bandwidth selection is quite extensive, e.g., monographs Wand and Jones (1995), Silverman (1986) and Simonoff (1996), papers Marron and Ruppert (1994), Park and Marron (1990), Scott and Terrell (1987), Jones and Kappenman (1991) and Cao et al. (1994). As far as the kernel estimate of density derivatives is concerned, this problem has received significantly less attention. In paper Härdle et al. (1990), an adaptation of the least squares cross-validation method is proposed for the bandwidth choice in the kernel density derivative estimation. In paper Horová et al. (2002), the automatic procedure of simultaneous choice of the bandwidth, the kernel and its order for kernel density and its derivative estimates was proposed. But this procedure can be only applied in case that the explicit minimum of the Asymptotic Mean Integrated Square Error of the estimate is available. It is known that this minimum exists only for $d = 2$ and the diagonal matrix H. In paper Horová et al. (2012), the basic formula for the corresponding procedure is given.

The need for nonparametric density estimates for recovering structure in multivariate data is greater since a parametric modeling is more difficult than in the univariate case. The extension of the univariate kernel methodology is not without its problems. The most general smoothing parameterization of the kernel estimator in d dimensions requires the specification entries of $d \times d$ positive definite bandwidth matrix. The multivariate kernel density estimator we are going to deal with is a direct extension of the univariate estimator (see, e.g., Wand and Jones (1995)).

Successful approaches to the univariate bandwidth selection can be transferred to the multivariate settings. The least squares cross-validation and plug-in methods in the multivariate case have been developed and widely discussed in papers

* Correspondence to: Department of Mathematics and Statistics, Kotlářská 2, 61137, Brno, Czech Republic. Tel.: +420 549494429; fax: +420 549491421.
E-mail addresses: horova@math.muni.cz (I. Horová), kolacek@math.muni.cz (J. Koláček), 63985@mail.muni.cz (K. Vopatová).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.07.006

Chacón and Duong (2010), Duong and Hazelton (2005b,a), Sain et al. (1994) and Duong (2004). Some papers (e.g., Horová et al. (2008, 2012) and Vopatová et al. (2010)) have been focused on constrained parameterization of the bandwidth matrix such as a diagonal matrix. It is well-known fact that a visualization is an important component of the nonparametric data analysis. In paper Horová et al. (2012), this effective strategy was used to clarify the process of the bandwidth matrix choice using bivariate functional surfaces. The paper Horová and Vopatová (2011) brings a short communication on a kernel gradient estimator. Tarn Duong's PhD thesis (Duong, 2004) provides a comprehensive survey of bandwidth matrix selection methods for kernel density estimation. Papers Chacón et al. (2011) and Duong et al. (2008) investigated general density derivative estimators, i.e., kernel estimators of multivariate density derivatives using general (or unconstrained) bandwidth matrix selectors. They defined the kernel estimator of the multivariate density derivative and provided results for the Mean Integrated Square Error convergence asymptotically and for finite samples. Moreover, the relationship between the convergence rate and the bandwidth matrix has been established here. They also developed estimates for the class of normal mixture densities.

The paper is organized as follows: In Section 2 we describe kernel estimates of a density and its gradient and give a form of the Mean Integrated Square Error and the exact MISE calculation for a d-variate normal kernel as well. The next sections are devoted to a data-driven bandwidth matrix selection method. This method is based on an optimally balanced relation between the integrated variance and the integrated squared bias, see Horová and Zelinka (2007a). Similar ideas were applied to kernel estimates of hazard functions (see Horová et al. (2006) or Horová and Zelinka (2007b)). It seems that the basic idea can be also extended to a kernel regression and we are going to investigate this possibility. We discuss the statistical properties and relative rates of convergence of the proposed method as well. Section 5 brings a simulation study and in the last section the developed theory is applied to real data sets.

2. Estimates of a density and its gradient

Let a d-variate random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be drawn from a density f. The kernel density estimator $\hat{f}$ at the point $\mathbf{x} \in \mathbb{R}^d$ is defined as

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

where K is a kernel function, which is often taken to be a d-variate symmetric probability function, $\mathbf{H}$ is a $d \times d$ symmetric positive definite matrix and $K_{\mathbf{H}}$ is the scaled kernel function

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$$

with $|\mathbf{H}|$ the determinant of the matrix $\mathbf{H}$.

The kernel estimator of the gradient Df at the point $\mathbf{x} \in \mathbb{R}^d$ is

$$\widehat{Df}(\mathbf{x}, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} DK_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i), \tag{2}$$

where $DK_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} \mathbf{H}^{-1/2} DK(\mathbf{H}^{-1/2}\mathbf{x})$ and DK is the column vector of the partial derivatives of K.

Since we aim to investigate both density itself and its gradient in a similar way, we introduce the notation

$$\widehat{D^r f}(\mathbf{x}, \mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} D^r K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i), \quad r = 0, 1, \tag{3}$$

where $D^0 f = f,\ D^1 f = Df$.

We make some additional assumptions and notations:

(A_1) The kernel function K satisfies the moment conditions $\int K(\mathbf{x})d\mathbf{x} = 1$, $\int \mathbf{x}K(\mathbf{x})d\mathbf{x} = \mathbf{0}$, $\int \mathbf{x}\mathbf{x}^T K(\mathbf{x})d\mathbf{x} = \beta_2 \mathbf{I}_d$, $\mathbf{I}_d$ is the $d \times d$ identity matrix.

(A_2) $\mathbf{H} = \mathbf{H}_n$ is a sequence of bandwidth matrices such that $n^{-1/2}|\mathbf{H}|^{-1/2}(\mathbf{H}^{-1})^r$, $r = 0, 1$, and entries of $\mathbf{H}$ approach zero ($(\mathbf{H}^{-1})^0$ is considered as equal to 1).

(A_3) Each partial density derivative of order $r + 2$, $r = 0, 1$, is continuous and square integrable.

(N_1) $\mathcal{H}$ is a class of $d \times d$ symmetric positive definite matrices.

(N_2) $V(\rho) = \int_{\mathbb{R}} \rho^2(x)dx$ for any square integrable scalar valued function ρ.

(N_3) $V(g) = \int_{\mathbb{R}^d} g(\mathbf{x})g^T(\mathbf{x})d\mathbf{x}$ for any square integrable vector valued function g. In the rest of the text, $\int$ stands for $\int_{\mathbb{R}^d}$ unless it is stated otherwise.

(N_4) $DD^T = D^2$ is a Hessian operator. Expressions like $DD^T = D^2$ involve "multiplications" of differentials in the sense that

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} = \frac{\partial^2}{\partial x_i \partial x_j}.$$

This means that $(D^2)^m$, $m \in \mathbb{N}$, is a matrix of the $2m$-th order partial differential operators.