Journal of the Korean Statistical Society 47 (2018) 112–126

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

On multivariate associated kernels to estimate general density functions

Célestin C. Kokonendji, Sobom M. Somé *

Laboratoire de Mathématiques de Besançon, Université Bourgogne Franche-Comté, Besançon, France

ARTICLE INFO

Article history:
Received 11 January 2017
Accepted 14 October 2017
Available online 21 November 2017

AMS 2000 subject classifications:
primary 62G07(08)
secondary 62H12

Keywords:
Asymmetric kernel
Boundary bias
Correlation structure
Bandwidth matrix
Nonparametric estimation
Mode-dispersion

ABSTRACT

Multivariate associated kernel estimators, which depend on both target point and bandwidth matrix, are appropriate for distributions with partially or totally bounded supports and generalize the classical ones such as the Gaussian. Previous studies on multivariate associated kernels have been restricted to products of univariate associated kernels, also considered having diagonal bandwidth matrices. However, it has been shown in classical cases that, for certain forms of target density such as multimodal ones, the use of full bandwidth matrices offers the potential for significantly improved density estimation. In this paper, general associated kernel estimators with correlation structure are introduced. Asymptotic properties of these estimators are presented; in particular, the boundary bias is investigated. Generalized bivariate beta kernels are handled in more details. The associated kernel with a correlation structure is built with a variant of the mode-dispersion method and two families of bandwidth matrices are discussed using the least squared cross validation method. Simulation studies are done. In the particular situation of bivariate beta kernels, a very good performance of associated kernel estimators with correlation structure is observed compared to the diagonal case. Finally, an illustration on a real dataset of paired rates in a framework of political elections is presented.

1. Introduction

Nonparametric estimation of unknown densities on partially or totally bounded supports, with or without correlation in their multivariate components, is a recurrent practical problem. Because of symmetry, the multivariate classical or symmetric kernels, not depending on any parameter, are not appropriate for these densities. In fact, these estimators give weights outside the support causing a bias in boundary regions. In order to reduce the boundary problem with multivariate symmetric kernels, Sain (2002) and recently Zougab, Adjabi, and Kokonendji (2014) have proposed adaptive full bandwidth matrix selection; but it does not remove the bias completely. Chen (1999, 2000) is one of the pioneers who has proposed, in the univariate continuous case, some asymmetric kernels (i.e. beta and gamma) whose supports coincide with those of the densities to be estimated. Also recently, Libengué (2013) investigated several families of these univariate continuous kernels that he called univariate associated kernels; see also Kokonendji, Kiéssé, and Zocchi (2007), Kokonendji and Kiéssé (2011) and Zougab, Adjabi, and Kokonendji (2012, 2013) for univariate discrete situations. This procedure cancels of course the boundary bias; however, it creates a quantity in the bias of the estimator which needs reduction; see, for instance, Malec and Schienle (2014) and Hirukawa and Sakudo (2014).

Several approaches to multivariate kernel estimation have been proposed for various purposes. Garcìa-Portugués, González-Rodríguez, Crujeiras, and González-Manteiga (2013) used product of kernels for estimating the different nature of

* Correspondence to: Laboratoire de Mathématiques de Besançon—UMR 6623 CNRS-UFC, 16 route de Gray, 25030 Besançon cedex, France.
E-mail addresses: celestin.kokonendji@univ-fcomte.fr (C.C. Kokonendji), sobom.some@univ-fcomte.fr (S.M. Somé).

https://doi.org/10.1016/j.jkss.2017.10.002
1226-3192/

both directional and linear components of a random vector. Hielscher (2013) investigated classical kernel density estimation using rotation group and adapted for crystallographic texture analysis. Symmetric kernel smoothers with univariate local bandwidth have been studied by González-Manteiga, Lombardía, Martínez-Miranda, and Sperlich (2013) for semiparametric mixed effect models. Girard, Guillou, and Stupfler (2013) presented frontier estimation with classical kernel regression on high order moments. In the discrete case, Aitchison and Aitken (1976) provided kernel estimators for binary data while Racine and Li (2004) proposed the product of them with classical continuous ones for smoothing regression on both categorical and continuous data. In the same spirit as Racine and Li (2004), Bouezmarni and Rombouts (2010) considered some products of different univariate associated kernels in the continuous case; i.e. the bandwidth matrix obtained was diagonal. In the classical kernels case, Chacón and Duong (2011) and Chacón, Duong, and Wand (2011) have shown the importance of full bandwidth matrices for certain target densities. See also Hazelton and Marshall (2009) for a support with arbitrary shape.

The main goal of this work is to introduce the multivariate associated kernels with the most general bandwidth matrix. In other words, the support of the suggested associated kernels coincides with the support of the densities to be estimated; also, the full bandwidth matrices take into account different correlation structures in the sample. Note that a full bandwidth matrix significantly improves estimation of some complex target densities (e.g. multimodal); see Sain (2002). In high dimensions, the computational choice of this full bandwidth matrix needs some special techniques. We can refer to Chacón and Duong (2010, 2011) and Chacón et al. (2011) for classical (symmetric) kernels. For illustrations in the present paper, we focus on the bivariate case using a beta kernel with correlation structure introduced by Sarmanov (1966); see also Lee (1996). A motivation to investigate the estimation of densities on $[0, 1] \times [0, 1]$ comes from the joint distribution of two comparable proportions. Many datasets in $[0, 1] \times [0, 1]$ can be found in statistical problems, for example, for comparing two rates. We shall examine the theoretical bias reduction and practical performances of the full bandwidth selection and two other bandwidth matrix parametrizations using least squares or unbiased cross validation; see, e.g., Wand and Jones (1993).

The rest of the paper is organized as follows. Section 2 presents a complete definition of multivariate associated kernels which includes both the product and the classical symmetric ones. A method to construct any multivariate associated kernel from a parametric probability density function (pdf) is then provided. Some pointwise properties of the corresponding estimator are investigated; in particular, we show the convergence in the sense of the mean integrated squared error (MISE) and an algorithm for the bias reduction. Section 3 provides a particular study of a bivariate beta kernel with a correlation structure introduced by Sarmanov (1966). Also, some algorithms for the choice of the optimal bandwidth matrix by the unbiased cross validation method are presented. This is followed, in Section 4, by simulation studies and a real data analysis of electoral behavior of a population with regard to a candidate. Especially, the role of forms of bandwidth matrices is explored in detail. Section 5 concludes with summary and final remarks. Proofs of propositions and other details are given in Supplemental material.

2. Multivariate associated kernel estimators

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent and identically distributed (i.i.d.) random vectors with an unknown density function f on $\mathbb{T}_d$, a subset of $\mathbb{R}^d$ ($d \geq 1$). As frequently observed in practice, the subset $\mathbb{T}_d$ might be unbounded, partially bounded or totally bounded as:

$$\mathbb{T}_d = \mathbb{R}^{d_\infty} \times [z, \infty)^{d_z} \times [u, w]^{d_{uw}} \tag{2.1}$$

for given reals $u < w$ and z with nonnegative values of d_∞, d_z and d_{uw} in $\{0, 1, \ldots, d\}$ such that $d = d_\infty + d_z + d_{uw}$. A multivariate associated kernel estimator $\widehat{f}_n$ of f is defined by

$$\widehat{f}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{x},\mathbf{H}}(\mathbf{X}_i), \quad \forall \mathbf{x} \in \mathbb{T}_d \subseteq \mathbb{R}^d, \tag{2.2}$$

where $\mathbf{H}$ is a $d \times d$ bandwidth matrix (i.e. symmetric and positive definite) such that $\mathbf{H} \equiv \mathbf{H}_n \to \mathbf{0}_d$ (the $d \times d$ null matrix) as $n \to \infty$, and $K_{\mathbf{x},\mathbf{H}}(\cdot)$ is the so-called associated kernel, parametrized by $\mathbf{x}$ and $\mathbf{H}$, and precisely defined as follows.

Definition 2.1. Let $\mathbb{T}_d \left(\subseteq \mathbb{R}^d\right)$ be the support of the pdf to be estimated, $\mathbf{x} \in \mathbb{T}_d$ a target vector and $\mathbf{H}$ a bandwidth matrix. A parametrized pdf $K_{\mathbf{x},\mathbf{H}}(\cdot)$ on support $\mathbb{S}_{\mathbf{x},\mathbf{H}} \left(\subseteq \mathbb{R}^d\right)$ is called "multivariate (or general) associated kernel" if the following conditions are satisfied:

$$\mathbf{x} \in \mathbb{S}_{\mathbf{x},\mathbf{H}}, \tag{2.3}$$

$$\mathbb{E}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{x} + \mathbf{a}(\mathbf{x}, \mathbf{H}), \tag{2.4}$$

$$\mathrm{Cov}\left(\mathcal{Z}_{\mathbf{x},\mathbf{H}}\right) = \mathbf{B}(\mathbf{x}, \mathbf{H}), \tag{2.5}$$

where $\mathcal{Z}_{\mathbf{x},\mathbf{H}}$ denotes the random vector with pdf $K_{\mathbf{x},\mathbf{H}}$ and both $\mathbf{a}(\mathbf{x}, \mathbf{H}) = (a_1(\mathbf{x}, \mathbf{H}), \ldots, a_d(\mathbf{x}, \mathbf{H}))^\top$ and $\mathbf{B}(\mathbf{x}, \mathbf{H}) = \left(b_{ij}(\mathbf{x}, \mathbf{H})\right)_{i,j=1,\ldots,d}$ tend, respectively, to the null vector $\mathbf{0}$ and the null matrix $\mathbf{0}_d$ as $\mathbf{H}$ goes to $\mathbf{0}_d$.