Journal of the Korean Statistical Society 45 (2016) 557–567

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Bayesian bandwidth selection in discrete multivariate associated kernel estimators for probability mass functions

Nawal Belaid [a,*], Smail Adjabi [a], Nabil Zougab [a,b], Célestin C. Kokonendji [c]

[a] LAMOS, Laboratory of Modelling and Optimization of Systems, University of Bejaia, Algeria
[b] University of Tizi-ouzou, Algeria
[c] University of Franche-Comté, LMB UMR 6623 CNRS-UFC, Besançon Cedex, France

ARTICLE INFO

Article history:
Received 9 July 2015
Accepted 8 April 2016
Available online 30 April 2016

AMS 2000 subject classifications:
primary 62G07
secondary 62G99

Keywords:
Binomial kernel
Cross-validation
Discrete triangular kernel
MCMC
Product kernel

ABSTRACT

This paper proposed a nonparametric estimator for probability mass function of multivariate data. The estimator is based on discrete multivariate associated kernel without correlation structure. For the choice of the bandwidth diagonal matrix, we presented the Bayes global method against the likelihood cross-validation one, and we used the Bayesian Markov chain Monte Carlo (MCMC) method for deriving the global optimal bandwidth. We have compared the proposed method with the cross-validation method. The performance of both methods is evaluated under the integrated square error criterion through simulation studies based on for univariate and multivariate models. We also presented applications of the proposed methods to bivariate and trivariate real data. The obtained results show that the Bayes global method performs better than cross-validation one, even for the Poisson kernel which is the very bad discrete associated kernel among binomial, discrete triangular and Dirac discrete uniform kernels.

Contents

[*] Correspondence to: LAMOS, route de Targa-Ouzemmour, 06000 Bejaia, Algeria.
E-mail addresses: belaidnawelro@hotmail.fr (N. Belaid), adjabi@hotmail.com (S. Adjabi), nabilzougab@yahoo.fr (N. Zougab), celestin.kokonendji@univ-fcomte.fr (C.C. Kokonendji).

http://dx.doi.org/10.1016/j.jkss.2016.04.001
1226-3192/© 2016 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.

1. Introduction

Nonparametric estimation of probability density and probability mass functions (pdf and pmf) by kernel method is one of the most investigated topics in statistical inference; see, e.g., Aitchison and Aitken (1976), Kokonendji and Senga Kiessé (2011), Kokonendji, Senga Kiessé, and Zocchi (2007), Silverman (1986), Simonoff (1996), Wand and Jones (1995) and Wang and Ryzin (1981); see also Somé and Kokonendji (2015) for nonparametric multivariate regression function (rf) estimation by using the associated continuous and discrete kernel method. The discrete kernel estimation has been far less investigated in comparison with continuous kernel estimation. For estimating the pmf of discrete variable, the empirical (Dirac) estimator is used because of its good asymptotical properties. However, this Dirac type kernel estimator is not appropriate with small sample sizes (see, Kokonendji & Senga Kiessé, 2011). Aitchison and Aitken (1976) provided an extension of the Dirac kernel method. But this proposed discrete kernel is appropriate for categorical data and finite count distributions; see also Racine and Li (2004) for rf estimation with both categorical and continuous data. Therefore, recently the so-called discrete associated kernel method is largely developed by several authors; see, e.g., Kokonendji and Senga Kiessé (2011), Kokonendji, Senga Kiessé, and Balakrishnan (2009), Kokonendji et al. (2007), Wansouwé, Kokonendji, and Kolyang (2015), Zougab, Adjabi, and Kokonendji (2012) and Zougab, Adjabi, and Kokonendji (2013a). Note that the semi-parametric estimation of pmf and count rf (crf) is also investigated, see for example Abdous, Kokonendji, and Senga Kiessé (2012) and Senga Kiessé, Zougab, and Kokonendji (2015). They showed that the use of a discrete associated kernel is more appropriate than the use of a continuous kernel for both estimations of pmf and crf of a discrete variable. To our knowledge, the multivariate pmf estimation by discrete kernel method has not been investigated in the literature. This paper introduces a nonparametric discrete kernel estimator for pmf of multivariate discrete data and proposes a Bayesian approach for bandwidth matrix selection.

Similarly to the univariate case, the selection of bandwidth matrix is a serious problem in nonparametric pdf and pmf estimation with multivariate associated kernel method. This method depends on the bandwidth and the associated kernel function. The choice of the kernel should be appropriate with respect to the support of the unknown function to be estimated. However, the choice of the bandwidth is substantive because, in the univariate case, the bandwidth is a scalar parameter h strictly positive which controls the degree of smoothing; and, in the multivariate case, the bandwidth is a symmetric and positive definite matrix H, that controls both the degree of smoothing and the form of orientation of the kernel. The bandwidth matrix has $d(d+1)/2$ independent elements to be chosen in $\mathbb{R}^d$, $d \geq 1$. For this, the choice of the bandwidth in this context is more difficult than that of the univariate case. A simplification can be obtained by imposing H as a diagonal matrix $H = Diag_d(h_j)_{j=1,2,\ldots,d}$. But for certain target densities, some authors have shown the importance of full bandwidth matrices, see Chacón and Duong (2011). Several methods have been proposed and studied to select bandwidth matrices for various processings. For example, the classical ones based on the plug-in methods that adopt as criterion the mean integrated square error ($MISE$) using symmetric kernels, such as the Gaussian kernel, see Chacón and Duong (2010), Duong and Hazelton (2003), and Wand and Jones (1994); and, cross-validation methods including the unbiased, biased, smoothed and least squares cross validation approaches for which we can refer to Bouezmarni and Roumbouts (2010), Chacón and Duong (2011), Duong and Hazelton (2005) and Sain, Baggerly, and Scott (1994). Note that the kernel estimations of density and probability mass functions are fundamental for other studies as kernel regression estimation and kernel discriminant analysis.

The Bayesian approach is a good alternative to the classical methods. This approach has received an attention in the literature, particularly in the univariate context for symmetric kernels. We can see, Brewer (1998, 2000) who proposed the global and adaptive Bayesian approach. Gangopadhyay and Cheung (2002), Kulasekera and Padgett (2006) and Kuruwita, Kulasekera, and Padgett (2010) proposed a Bayesian local bandwidth selection. For asymmetric kernels, we can consult Zougab et al. (2012, 2013a); Zougab, Adjabi, and Kokonendji (2013b) who used discrete associated kernels. In the multivariate context, Zhang, King, and Hyndman (2006) presented MCMC method for the global bandwidth matrix selection using the symmetric Gaussian kernel; see also Zhang, King, and Shang (2013) for bandwidth selection in nonparametric regression model with mixed types of regressors. Always for symmetric kernels, we can also refer to De Lima and Atuncar (2010) for the Bayesian local approach for which they have extended to the multivariate case the method described in Gangopadhyay and Cheung (2002), and Hu, Poskitt, and Zhang (2012) for Bayesian adaptive approach with MCMC method. Recently Zougab, Adjabi, and Kokonendji (2014) presented a Bayesian adaptive for the choice of the bandwidth matrix using the Gaussian kernel.

As mentioned earlier, there is now a study that used discrete multivariate associated kernels to estimate pmf of multivariate data. Therefore, the main contribution of this paper is to propose a nonparametric estimator for pmf of multivariate data by using discrete associated kernel and develop a Bayesian approach to bandwidth matrix selection. Our study is motivated by several points. First, similar to discrete univariate case, the use of multivariate discrete kernel are more suitable than the use of multivariate continuous kernel for multivariate data. Second, the discrete multivariate associated kernels method is useful in various domains such as in actuarial science, demography, economics, environmental sciences and sport. As third motivation, the classical methods to bandwidth matrix selection such as cross-validation technique do not provide a good estimator, then the Bayesian approach which is considered as a good alternative is presented.

The rest of this paper is organized as follows. We give a complete definition of discrete multivariate associated kernels which includes the product ones, and we illustrate some examples in Section 2. In Section 3, we present Bayesian global methods for estimating the bandwidth matrix by using MCMC algorithms through the likelihood cross-validation criterion,