Physica A 454 (2016) 94–98

Contents lists available at ScienceDirect

Physica A

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

Novel and simple non-parametric methods of estimating the joint and marginal densities

Moawia Alghalith

Department of Economics, University of the West Indies, St Augustine, Trinidad and Tobago

HIGHLIGHTS

- Simple non-parametric methods that overcome key limitations of the existing literature on both the joint and marginal density estimation.
- Does not assume any form of the marginal distribution or joint distribution a priori.
- The method circumvents the bandwidth selection problems.
- Compare our method to the kernel density method.

ARTICLE INFO

Article history:
Received 23 November 2015
Received in revised form 13 January 2016
Available online 2 March 2016

Keywords:
Joint density
Marginal density
Distribution
Semi-martingale
Non-parametric density estimation

ABSTRACT

We introduce very simple non-parametric methods that overcome key limitations of the existing literature on both the joint and marginal density estimation. In doing so, we do not assume any form of the marginal distribution or joint distribution a priori. Furthermore, our method circumvents the bandwidth selection problems. We compare our method to the kernel density method.

1. Introduction

Each of the existing methods of density estimation (parametric, non-parametric and semi-parametric methods) suffers serious limitations. For example, one of the limitations of the parametric method is that the distributions need to be known (see, for example, Refs. [1,2]). The existing non-parametric methods in particular suffer key limitations such as the bandwidth selection problems and the very high computational cost, among others. Moreover, the non-parametric methods still require some assumptions about the form of the distribution, since they require a kernel specification. Recent examples of the non-parametric approach include Chen [3], Zhang [4], Jones et al. [5], Ruppert et al. [6], Scricciolo [7], Shen et al. [8], Rousseau [9], Durante and Okhren [10], Hazlett [11] and Weib [12], among many others. Examples of empirical studies include Sheikhpour et al. [13], Talamakrouni et al. [14], Siddharth and Taylor [15] and Xu et al. [16]. Other studies used copulas (for a discussion of the copula method and its limitations, see, for example, Ref. [17]).

In this paper, we introduce very simple (yet accurate) non-parametric methods that overcome key limitations of the existing literature on both the joint and marginal density estimation. In doing so, we do not assume any form of marginal distribution or joint distribution a priori. That is, we do not need to have any prior knowledge of the distributions. Moreover,

E-mail address: malghalith@gmail.com.

http://dx.doi.org/10.1016/j.physa.2016.02.034

our method circumvents the bandwidth selection problems. Furthermore, our methods, compared to the existing methods, are exceedingly simple and fast. We compare our method to the kernel density method. This method can be applied to numerous areas of physics. Examples include high energy physics, spectroscopy and artificial neural networks (see, for example, Refs. [18–20]).

2. The joint density estimation

Let $F(x,y)$ be the cumulative joint density of X and Y, then we have

$$\mathrm{d}F(x,y)=\frac{\partial F(x,y)}{\partial x}\mathrm{d}x+\frac{\partial F(x,y)}{\partial y}\mathrm{d}y;\qquad \frac{\partial^2 F(x,y)}{\partial x\partial y}=f(x,y), \tag{1}$$

where $f(x,y)$ is the joint density. It is also well known that

$$\frac{\partial F(x,y)}{\partial x}=\int f(x,y)\,\mathrm{d}y=f_X(x);\qquad \frac{\partial F(x,y)}{\partial y}=\int f(x,y)\,\mathrm{d}x=f_Y(y), \tag{2}$$

$$\frac{\partial^2 F(x,y)}{\partial x^2}=\frac{\mathrm{d}^2F_X(x)}{\mathrm{d}x^2}=\frac{\mathrm{d}f_X(x)}{\mathrm{d}x};\qquad \frac{\partial^2 F(x,y)}{\partial y^2}=\frac{\mathrm{d}^2F_Y(y)}{\mathrm{d}y^2}=\frac{\mathrm{d}f_Y(y)}{\mathrm{d}y}, \tag{3}$$

where $f_X(x)$ is the marginal density of X, $f_Y(y)$ is the marginal density of Y, $F_X(x)$ and $F_Y(y)$ are the cumulative densities of X and Y, respectively. Substituting (2) and (3) into (1) yields

$$\mathrm{d}F(x,y)=f_X(x)\,\mathrm{d}x+f_Y(y)\,\mathrm{d}y. \tag{4}$$

Thus,

$$\begin{aligned}\mathrm{d}^2F(x,y)&=\frac{\partial^2 F(x,y)}{\partial x^2}(\mathrm{d}x)^2+\frac{\partial^2 F(x,y)}{\partial y^2}(\mathrm{d}y)^2+2f(x,y)\,\mathrm{d}x\mathrm{d}y\\&=df_X(x)\,\mathrm{d}x+df_Y(y)\,\mathrm{d}y+2f(x,y)\,\mathrm{d}x\mathrm{d}y.\end{aligned} \tag{5}$$

In practical applications (using empirical data), we use the first difference and the second difference instead of the differential as follows

$$\Delta F(x,y)=f_X(x)\,\Delta x+f_Y(y)\,\Delta y, \tag{6}$$

$$\Delta^2F(x,y)=\Delta f_X(x)\,\Delta x+\Delta f_Y(y)\,\Delta y+2f(x,y)\,\Delta x\Delta y. \tag{7}$$

Therefore, the joint density can be calculated (for each observation of the data) as follows

$$\hat{f}(x,y)=\frac{\Delta^2F(x,y)-\Delta f_X(x)\,\Delta x-\Delta f_Y(y)\,\Delta y}{2\Delta x\Delta y}. \tag{8}$$

Clearly, x and y are observed data; $\Delta F(x,y)$ can be calculated for each observation using (6) if the marginal densities are known or estimated, and it is needless to say that $\Delta^2F(x,y)$ is the first difference of $\Delta F(x,y)$. In sum, the joint density can be easily calculated if the marginal densities are known. However, we can easily estimate the marginal densities (see the next section).

If we use the finite difference method to estimate (8), the estimation error $o(.)$ is given by

$$\hat{f}(x,y)=f(x,y)+o\left(\Delta x^2,\Delta y^2\right). \tag{9}$$

Therefore

$$Bias\left(\hat{f}(x,y)\right)=E\hat{f}(x,y)-f(x,y)=Eo\left(\Delta x^2,\Delta y^2\right),$$

$$MSE=E\left(\hat{f}(x,y)-f(x,y)\right)^2=Eo\left(\Delta x^2,\Delta y^2\right)^2,$$

where MSE is the mean squared error.

$$\mathrm{Var}\left(\hat{f}(x,y)\right)=Eo\left(\Delta x^2,\Delta y^2\right)^2-\left(Eo\left(\Delta x^2,\Delta y^2\right)\right)^2,$$

$$MISE=E\iint\left(\hat{f}(x,y)-f(x,y)\right)^2\mathrm{d}x\mathrm{d}y=E\iint o\left(\Delta x^2,\Delta y^2\right)^2\mathrm{d}x\mathrm{d}y,$$

where MISE is the mean integrated squared error.

$$AMISE=0,$$