Economics Letters 152 (2017) 112–116

Contents lists available at ScienceDirect

Economics Letters

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

Smoothed kernel conditional density estimation

Kuangyu Wen [a], Ximing Wu [b,*]

[a] International School of Economics and Management, Capital University of Economics and Business, Beijing 100070, PR China
[b] Department of Agricultural Economics, Texas A&M University, College Station, TX 77843, USA

HIGHLIGHTS

- Conditional density of Y given X with multiple Y's for each observed X.
- Kernel conditional density estimator that smooths $f(y|x)$ across x.
- Large sample properties depend on the sample size of X and that of Y at each X.
- A practical cross validation bandwidth selector.

ARTICLE INFO

Article history:
Received 25 November 2016
Received in revised form 11 January 2017
Accepted 13 January 2017
Available online 16 January 2017

JEL classification:
C14

Keywords:
Conditional density estimation
Bandwidth selection
Body mass index

ABSTRACT

Given multiple Y observations for each observed X, we propose a conditional kernel density estimator that exploits smoothing of $f(y|x)$ across x. We obtain large sample properties of the proposed estimator and present a practical cross validation bandwidth selector. An application to adult BMI densities conditional on age is provided.

1. Introduction

Consider an $\mathbb{R}^d \times \mathbb{R}$-valued random vector (X, Y). This article concerns the estimation of the conditional density of Y given X. Throughout the text, we shall refer to Y as the dependent/response variable and to X as the covariate. Denote by $g(x, y)$ the joint density of (X, Y) and by $h(x)$ the marginal density of X, both of which are assumed to exist. The conditional density of Y with $X = x$ is given by $f(y|x) = g(x, y)/h(x)$.

Given an I.I.D. sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, one can estimate g and h nonparametrically using the kernel density estimator. Plugging them into the formula above yields the classical conditional density estimator by Rosenblatt (1969):

$$\tilde{f}(y|x) = \frac{\sum_{i=1}^n G(\|x - X_i\|/a)K((y - Y_i)/b)}{b\sum_{i=1}^n G(\|x - X_i\|/a)}, \tag{1.1}$$

where K and G are univariate kernel functions, a and b are their respective bandwidths, and $\|\cdot\|$ is some suitable norm. The kernel function is usually taken to be a density with support $\mathbb{R}$ or some finite interval. The bandwidths a and b control the degree of smoothing along X and Y respectively. Detailed treatments of estimator (1.1) and its generalizations can be found in Hyndman et al. (1996), Bashtannyk and Hyndman (2001) and Hall et al. (2004). Alternatively conditional density estimation can be conducted in a regression setting; see e.g., Fan et al. (1996), Hyndman and Yao (2002), Fan and Yim (2004), De Gooijer and Zerom (2003) and Efromovich (2007).

In practice for each observed X_i, $i = 1, \ldots, n$, we may observe multiple occurrences of Y, say, $\mathcal{Y}_i = \left\{Y_{i,1}, Y_{i,2}, \ldots, Y_{i,N_i}\right\}$, $N_i > 1$, yielding an expanded sample

$$\left\{(X_i, Y_{i,k}) : i = 1, \ldots, n; k = 1, \ldots, N_i\right\}. \tag{1.2}$$

For simplicity, we assume that $Y_{i,k}$'s are I.I.D. as Y_i given X_i and $\mathcal{Y}_i$'s are independent across i. The goal of this study is to investigate the nonparametric estimation of the conditional density $f(y|x)$ given a sample like (1.2). First note that given $\mathcal{Y}_i$, we can estimate the density of $Y|X_i$ by

$$\tilde{f}(y|X_i) = (N_i b)^{-1}\sum_{k=1}^{N_i} K((y - Y_{i,k})/b). \tag{1.3}$$

* Corresponding author.
E-mail addresses: kweneco@gmail.com (K. Wen), xwu@tamu.edu (X. Wu).

http://dx.doi.org/10.1016/j.econlet.2017.01.008
0165-1765/

Next we estimate the general conditional density $f(y|x)$ as a weighted average of $\tilde{f}(y|X_i)$ with kernel smoothed weights across X_i's. We define this estimator as

$$\begin{aligned}\hat{f}(y|x) &= \frac{\sum_{i=1}^{n} N_i G(\|x-X_i\|/a)\tilde{f}(y|X_i)}{\sum_{i=1}^{n} N_i G(\|x-X_i\|/a)} \\ &= \frac{\sum_{i=1}^{n} G(\|x-X_i\|/a)\sum_{k=1}^{N_i} K((y-Y_{i,k})/b)}{b\sum_{i=1}^{n} N_i G(\|x-X_i\|/a)},\end{aligned} \tag{1.4}$$

where K and G are kernel functions with their respective bandwidths a and b. Although similar in form to Rosenblatt (1969)'s estimator (1.1), the present estimator (1.4) is motived by local smoothing of conditional densities, which are estimated based on multiple occurrences of Y_i associated with each X_i. This sampling scheme deviates from the classical I.I.D. setting considered in Rosenblatt (1969), and warrants a close examination of the corresponding estimator (1.4). In this article we establish the large sample properties of (1.4), which are shown to depend on the interplay of sample size n and N_i, the number of Y_i associated with each X_i. We then present a practical method of bandwidth selection. We illustrate the proposed estimator with an application to the distribution of Body Mass Index, for adult male and female separately, conditional on age.

2. Main results

For simplicity and ease of presentation, we focus on the case where X is univariate and the size of $\mathcal{Y}_i$ is identical such that $N_i = N$, $i = 1, \ldots, n$. Estimator (1.4) is then simplified to

$$\hat{f}(y|x) = \frac{\sum_{i=1}^{n} G((x-X_i)/a)\sum_{k=1}^{N} K((y-Y_{i,k})/b)}{Nb\sum_{i=1}^{n} G((x-X_i)/a)}.$$

We first derive the asymptotic bias and variance of $\hat{f}(y|x)$. Some assumptions are in order. Most importantly, we suppose that n and N both go to infinity and the bandwidths $a, b \to 0$, $naNb \to \infty$ as $n, N \to \infty$. We also assume that the joint density $g(x, y)$ and the marginal density $h(x)$ are such that their second order derivatives are continuous and square integrable. The kernel functions G and K are symmetric and square integrable density functions with mean zero and finite variance. The leading asymptotic bias and variance of $\hat{f}(y|x)$ are then given by

$$\text{abias}\left\{\hat{f}(y|x)\right\} = \frac{a^2\sigma_G^2}{2}\left[f_{(2)}(y|x) + 2f_{(1)}(y|x)\frac{h'(x)}{h(x)}\right] + \frac{b^2\sigma_K^2}{2}f^{(2)}(y|x) \tag{2.1}$$

$$\text{avar}\left\{\hat{f}(y|x)\right\} = \frac{R(G)R(K)f(y|x)}{naNbh(x)} + \frac{aT(G)}{n}\frac{[f_{(1)}(y|x)]^2}{h(x)}, \tag{2.2}$$

where $\sigma_K^2 = \int z^2K(z)dz$, $\sigma_G^2 = \int z^2G(z)dz$, $R(K) = \int K^2(z)dz$, $R(G) = \int G^2(z)dz$, $T(G) = \int z^2G^2(z)dz$, $f^{(s)}(y|x) = \partial^s f(y|x)/\partial y^s$, $f_{(t)}(y|x) = \partial^t f(y|x)/\partial x^t$, $h'(x)$ and $h''(x)$ are the first and second derivative of $h(x)$ with respect to x. The derivations of (2.1) and (2.2) are given in an online supplementary appendix.

We note that the asymptotic bias (2.1) of our estimator $\hat{f}(y|x)$ is the same as that of Rosenblatt (1969)'s estimator (1.1). This is intuitively understood: the presence of multiple Y_i's to each X_i does not affect the bias of the estimate, but generally influences its variance. The asymptotic variance (2.2) consists of two terms; the first term approximates $\mathrm{E}\{\text{var}\{\hat{f}(y|x)|\mathcal{X}\}\}$ while the second approximates $\text{var}\{\mathrm{E}\{\hat{f}(y|x)|\mathcal{X}\}\}$, where $\mathcal{X} = \{X_1, \ldots, X_n\}$. When a single Y_i is observed for each X_i, i.e. $N = 1$, our estimator $\hat{f}(y|x)$ reduces to Rosenblatt (1969)'s estimator (1.1) and clearly the first term in (2.2) dominates. In fact, this first term is exactly the asymptotic variance of Rosenblatt (1969)'s estimator. However in the presence of multiple Y_i's, either term in (2.2) may dominate, depending on how fast n and N go to infinity.

Next we examine the global properties of our estimator in terms of the mean integrated squared error (MISE)

$$\mathrm{E}\int_x\int_y\left[\hat{f}(y|x) - f(y|x)\right]^2 dy\,w(x)dx, \tag{2.3}$$

where the integration is taken with respect to both x and y, and $w(x)$ is an appropriate weight function. A common choice of the weight function in the kernel smoothing literature is $h(x)$, the marginal density of the covariate X; see e.g. Wand and Jones (1995). When the weight function is set to be $h(x)$, we usually assume that X is defined on a bounded support to ensure integrability. We note that Bott and Kohler (2015) studied the consistency and convergence of a similar estimator in terms of the L_1-norm.

It is well known that the MISE can be written as the sum of integrated squared bias and integrated variance. Equipped with the asymptotic bias and variance above, we can show the asymptotic MISE of $\hat{f}(y|x)$ takes the form

$$c_1a^4 + c_2a^2b^2 + c_3b^4 + c_4\frac{1}{naNb} + c_5\frac{a}{n}, \tag{2.4}$$

where the constants c_1, c_2, c_3, c_4 and c_5 depend on the kernels G, K, the conditional density $f(y|x)$ and marginal density $h(x)$. To ease our presentation, the explicit expressions of these constants are omitted; they are available from the authors upon request. We then seek to minimize (2.4) with respect to the bandwidths a and b, yielding the following first order conditions

$$4c_1a^5b + 2c_2a^3b^3 - \frac{c_4}{nN} + \frac{c_5a^2b}{n} = 0 \tag{2.5}$$

$$4c_3ab^5 + 2c_2a^3b^3 - \frac{c_4}{nN} = 0. \tag{2.6}$$

It follows that the optimal bandwidths a and b satisfy the following relationship

$$b = \left[\frac{c_1a^4}{c_3} + \frac{c_5a}{4c_3n}\right]^{1/4}. \tag{2.7}$$

Next we show that the optimal asymptotic MISE depends on the relative magnitude of c_1a^4/c_3 and $c_5a/(4c_3n)$. There exist three possibilities.

(i) Suppose that $n/N \to \infty$ and $na^3 \to \infty$. In this case, a^4 dominates a/n. Dropping the dominated term from (2.7), we have $b \approx (c_1/c_3)^{1/4}a$. Plugging this into (2.5) and (2.6) yields

$$\left[4(c_1^5/c_3)^{1/4} + 2c_2(c_1/c_3)^{3/4}\right]a^6 - \frac{c_4}{nN} = 0,$$

which gives the optimal bandwidth

$$a^* = c_4^{1/6}\left[4(c_1^5/c_3)^{1/4} + 2c_2(c_1/c_3)^{3/4}\right]^{-1/6}(nN)^{-1/6}$$

and subsequently $b^* = (c_1/c_3)^{1/4}a^*$. It is easy to verify that the derived a^* is compatible with the conditions $n/N \to \infty$ and $na^3 \to \infty$. Therefore when n grows faster than N, the optimal asymptotic MISE is of order $O\left([nN]^{-2/3}\right)$.

(ii) Suppose that $n/N \to m_1$ and $na^3 \to m_2$ where $0 < m_1, m_2 < \infty$. We then have $a^* \sim n^{-1/3}$ and $b^* \sim n^{-1/3}$ according to (2.7). Let $a^* = \kappa_1 n^{-1/3}$ and $b^* = \kappa_2 n^{-1/3}$. Plugging them into (2.5) and (2.6) yields

$$\kappa_2 = \left[\frac{c_1}{c_3}\kappa_1^4 + \frac{c_5}{4c_3}\kappa_1\right]^{1/4},$$

$$4c_3\kappa_1\left[\frac{c_1}{c_3}\kappa_1^4 + \frac{c_5}{4c_3}\kappa_1\right]^{5/4} + 2c_2\kappa_1^3\left[\frac{c_1}{c_3}\kappa_1^4 + \frac{c_5}{4c_3}\kappa_1\right]^{3/4} = \frac{c_4n}{N}.$$