DOI: 10.13451/j.cnki.shanxi.univ(nat.sci.).2001.02.005

山西大学学报(自然科学版)24(2): 107~110, 2001
Journal of Shanxi University(Nat. Sci. Ed.)

文章编号: 0253-2395(2001)02-0107-04

二维概率密度核窗估计的快速算法

夏春华

(山西行政学院 计算机中心,山西 太原 030006)

摘 要: 概率密度核窗估计法,在窗宽选取最佳时,具有很高的精度。在各种确定最佳窗宽的方法中,最小二乘相关确认法是性能较好的一种,但因运算量大使其实际应用受到限制。本文在保证核窗估计高精度的前提下,给出了二维概率密度核窗估计的快速算法,并通过仿真运算,验证了该算法的有效性。

关键词: 快速算法;核窗估计;概率密度

中图分类号: O211 文献标识码: A

从随机向量 $\vec{X}$ 的 n 个样本 $\{\vec{X}_i\}_{i=1}^n$ 出发,用非参数的方法估计其未知的概率分布密度函数 $f(\vec{X})$,在统计分析、通信理论和模式识别等领域中有着广泛的应用。概率密度非参数估计的传统方法是直方图法,其特点是简单,但精度差,满足不了某些实际应用的要求。用核窗法估计概率密度函数,在窗宽选取恰当的前提下,可以得到精度很高的数值结果。但要得到一个恰当的窗宽并不容易,在各种确定最佳窗宽的方法中,对各种分布均有良好性能的是最小二乘相关确认(Lcast square cross validation)的方法,然而该方法确定最佳窗宽时所需的运算量大。不仅如此,在最佳窗宽确认后,要计算密度函数在一组点上的取值,其运算量也不小。因此,使用核窗法估计概率密度函数时,寻求快速算法是十分必要的。Silverman在1985年报道了使用最小二乘相关确认法确定最佳窗宽的一维概率密度函数核窗估计的快速算法[1].本文通过对该快速算法原理的推广,给出了二维概率密度核窗估计的快速算法,使得高精度核窗估计在某些场合的实际应用成为可能。

1 二维概率密度核窗估计的基本原理

设 $\vec{X}$ 是一具有连续分布的 d 维随机变量,$f(\vec{X})$ 是其 d 维联合概率分布密度,用 $f_n(\vec{X})$ 记由其 n 个样本估计得到的概率分布密度函数,则随机向量 $\vec{X}$ 的密度函数的核窗估计可由下式定义:

$$\hat{f}_n(\vec{X}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}\cdot k\left(\frac{\vec{X}-\vec{X}_i}{h}\right), \tag{1}$$

其中 $\{\vec{X}_i\}_{i=1}^n$ 是随机向量 $\vec{X}$ 的 n 个样本,$k(\cdot)$ 是 d 维核窗函数,满足

$$\int k(\vec{X})\,\mathrm{d}\vec{X} = 1. \tag{2}$$

取 $d=2$,即为二维概率密度的核窗估计,它可表示为:

$$\hat{f}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^2}\cdot k\left(\frac{x-x_i}{h},\frac{y-y_i}{h}\right) \tag{3}$$

估计结果与真实密度函数的差异,是用积分均方误差来度量的,简记为 MISE(f).

$$MISE(\hat{f}_n) = E\iint[f(x,y)-\hat{f}_n(x,y)]^2\,\mathrm{d}x\,\mathrm{d}y. \tag{4}$$

记

* 收稿日期: 2000-06-07

作者简介: 夏春华(1958-),男,山西运城人,1982年毕业于哈尔滨船舶工程学院,硕士,现任山西行政学院计算机中心讲师。研究方向: 数字信号处理与计算机应用。

$$T=\iint x^2\cdot k(x,y)\,dx\,dy,\quad U=\iint k^2(x,y)\,dx\,dy, \tag{5}$$

则在 $K(x,y)$旋转圆对称的条件下可推得:

$$MISE(\hat{f}_n)=\frac{1}{4}h^4T^2\iint[\nabla^2 f(x,y)]^2\,dx\,dy+n^{-1}h^{-1}U, \tag{6}$$

对其求导,可求得使积分均方误差达到最小的最佳窗宽 h_{opt}.

$$h_{opt}=\left\{2UT^{-2}\left[\iint[\nabla^2 f(x,y)]^2dx\,dy\right]^{-1}\cdot n^{-1}\right\}^{1/6}. \tag{7}$$

在样本数目确定的前提下,影响估计精度的参数主要是核窗宽 h.相对来讲,核窗函数的形状对估计精度的影响并不大,其选取通常是从运算和推导的简单方便等方面考虑的。由于(7)式中 h_{opt}依赖于未知的分布密度函数,因而由(7)式并不能直接求得最佳窗宽。在各种近似逼近最佳窗宽的方法中,对各种不同特征的未知分布均表现出良好性能的是最小二乘相关确认方法[1],该方法是通过如下统计量来实现的。

$$M_1(h)=n^{-2}h^{-2}\sum_i\sum_j k^*\left(\frac{x-x_i}{h},\frac{y-y_i}{h}\right)+2n^{-1}h^{-1}k(0,0), \tag{8}$$

其中

$$k^*(x,y)=k^2(x,y)-2k(x,y).$$

可以证明

$$EM_1(h)\doteq MISE(\hat{f}_n)-\iint f^2(x,y)\,dx\,dy,$$

因而,使 $EM_1(h)$取最小的 h_{opt}同样可使 $MISE(\hat{f}_n)$取最小,在实际确定最佳窗宽时,将使 $M_1(h)$取最小的 h 作为 h_{opt}.

2 快速算法的实现原理

核窗估计的实现分为两个部分:第一部分要确定最佳窗宽;第二部分要估计密度函数 $\hat{f}_n(x,y)$.我们注意到密度函数每一点的估值都要对 $K(\cdot,\cdot)$作 N次函数运算,然后再将 N次运算的结果相加,运算量很大,不仅如此,在使用最小二乘相关确认法确定最佳窗宽时,对应于每一 h值,要计算 $M_1(h)$的值,又要作出 $N\times N$次 $K(\cdot,\cdot)$的函数运算,然后再将这 $N\times N$次运算结果相加,运算量更大。下面将通过 Fourier变换来实现核窗估计的快速算法,其原理如下:

定义

$$\tilde{g}(s_1,s_2)=(2\pi)^{-1}\iint e^{j(s_1t_1+s_2t_2)}\cdot g(t_1,t_2)\,dt_1dt_2$$

为函数 $g(t_1,t_2)$的二维 Fourier变换。

定义

$$\_(s_1,s_2)=(2\pi)^{-1}\cdot n^{-1}\sum_{j=1}^{n}\exp\{i(s_1x_j+s_2y_j)\}$$

为数据 $\{(x_i,y_i)\}_{i=1}^n$的 Fourier变换。

则由(3)式定义的二维核窗估计的 Fourier变换为

$$\tilde{f}_n(s_1,s_2)=(2\pi)\cdot\tilde{k}(hs_1,hs_2)\cdot\_(s_1,s_2). \tag{9}$$

定义

$$j(t_1,t_2)=n^{-2}\sum_i\sum_j h^{-2}\cdot k^*\left((x_i-x_j-t_1)h^{-1},(y_i-y_j-t_2)h^{-1}\right), \tag{10}$$

则

$$M_1(h)=j(0,0)+2n^{-1}h^{-2}k(0,0),\quad \tilde{j}(s_1,s_2)=(2\pi)\cdot\tilde{k}^*(hs_1,hs_2)\cdot|\_(s_1,s_2)|^2. \tag{11}$$

为方便起见,取二维核窗函数为标准高斯窗:

$$k(t_1,t_2)=(2\pi)^{-1}\cdot\exp\{-\frac{1}{2}(t_1^2+t_2^2)\},$$

则有

$$\tilde{k}(s_1,s_2)=(2\pi)^{-1}\cdot\exp\{-\frac{1}{2}(s_1^2+s_2^2)\}.$$

此时

$$\tilde{k}^*(s_1,s_2)=(2\pi)^{-1}\cdot[\tilde{k}(s_1,s_2)]^2-2\tilde{k}(s_1,s_2)=(2\pi)^{-1}\cdot\exp\{-(s_1^2+s_2^2)\}-2\cdot(2\pi)^{-1}\cdot\exp\{-\frac{1}{2}(s_1^2+s_2^2)\},$$

于是

$$\hat{J}(0,0) = (2\pi)^{-1}\cdot\iint\hat{J}(s_1,s_2)\,ds_1ds_2$$
$$= \iint\{\exp[-h^2(s_1^2+s_2^2)] - 2\exp[-\frac{1}{2}h^2(s_1^2+s_2^2)]\}\cdot|\hat{\varphi}(s_1,s_2)|^2\,ds_1ds_2. \quad (12)$$

将(12)式代入(11)式可得 $M_1(h)$,进而通过扫描或求解可确定 h_{opt},将 h_{opt}代入(9)式可求得 $\hat{f}_n(s_1,s_2)$然后对 $\hat{f}_n(s_1,s_2)$作Fourier逆变换便得到核窗估计结果 $\hat{f}_n(s_1,s_2)$.

以下将具体地通过 FFT实现上述算法.

3 快速算法的实现方法

选定密度函数的估计区域 $[a_1,b_1]\times[a_2,b_2]$,将该区域按 $M=2^r$(r是正整数)分别沿 X轴和 Y轴方向 M等分($M\geq 16$). 记

$$W_1=(b_1-a_1)/M,\quad W_2=(b_2-a_2)/M,\quad t_{1k}=a_1+kW_1,\quad t_{2k}=a_2+kW_2,\quad k=0,1,\cdots,M-1,$$

若二维随机变量的某样本点(x,y)落入区域 $D_{kj}=[t_{1,k},t_{1,k+1}]\times[t_{2,j},t_{2,j+1}]$,则在该区域的四个顶点分别按权数 $W(t_{1,k},t_{2,j})$, $W(t_{1,k},t_{2,j+1})$, $W(t_{1,k+1},t_{2,j})$, $W(t_{1,k+1},t_{2,j+1})$进行累加,其中

$$W(t_{1,k},t_{2,j})=\frac{(t_{1,k+1}-x)\cdot(t_{2,j+1}-y)}{nW_1W_2},\quad W(t_{1,k},t_{2,j+1})=\frac{(t_{1,k+1}-x)\cdot(y-t_{2,j})}{nW_1W_2},$$
$$W(t_{1,k+1},t_{2,j})=\frac{(x-t_{1,k})\cdot(t_{2,j+1}-y)}{nW_1W_2},\quad W(t_{1,k+1},t_{2,j+1})=\frac{(x-t_{1,k})\cdot(y-t_{2,j})}{nW_1W_2}.$$

将所有样本$\{(x_i,y_i)\}_{i=1}^n$的权系数累加后所得二维序列记为$\{W_{kl}\}_{k,l=0}^{M-1}$.记

$$V_{p,q}=M^{-2}\sum_{k=0}^{M-1}\sum_{l=0}^{M-1}W_{kl}\cdot\exp\{i2\pi(kp+lq)/M\},$$
$$S_{1p}=(2\pi p)\cdot(b_1-a_1)^{-1}=\frac{2\pi p}{MW_1},\quad S_{2p}=(2\pi p)\cdot(b_2-a_2)^{-1}=\frac{2\pi p}{MW_2},$$

其中,$-M/2\leq p,q\leq M/2$,则

$$V_{p,q}=M^{-2}\sum_{k=0}^{M-1}\sum_{l=0}^{M-1}W_{kl}\cdot\exp\{i(t_{1k}S_{1p}+t_{2k}S_{2q})\}$$
$$=n^{-1}M^{-2}\cdot W_1^{-1}W_2^{-1}\sum_{j=1}^{n}\cdot\exp\{i(S_{1p}x_j+S_{2q}y_j)\} \quad (13)$$
$$=(2\pi)(b_1-a_1)^{-1}(b_2-a_2)^{-1}\cdot\hat{\varphi}(S_{1p},S_{2q})$$

再记

$$\tilde{V}_{pq}=\exp\{-\frac{1}{2}h^2(S_{1p}^2+S_{2q}^2)\}\cdot V_{pq}, \quad (14)$$

其则 FFT逆变换为

$$\sum_{p=-\frac{M}{2}}^{M/2}\sum_{q=-\frac{M}{2}}^{M/2}\exp\{-2\pi i(kp+lq)/M\}\cdot\tilde{V}_{pq}$$
$$=\sum_p\sum_q\exp\{-i(t_{1k}S_{1p}+t_{2l}S_{2q})\}\cdot 2\pi(b_1-a_1)^{-1}(b_2-a_2)^{-1}\cdot\exp\{-\frac{1}{2}h^2(S_{1p}^2+S_{2q}^2)\}\cdot\hat{\varphi}(s_{1p},s_{2q})$$
$$=(2\pi)\iint\exp\{-i(t_{1k}s_{1p}+t_{2l}s_{2q})\}\exp\{-\frac{1}{2}h^2(s_{1p}^2+s_{2q}^2)\}\hat{\varphi}(s_1,s_2)\,ds_1ds_2$$
$$=\hat{f}_n(t_{1k},t_{2l}).$$

对(2)式近似求和有:

$$\hat{J}(0,0)=(b_1-a_1)(b_2-a_2)\cdot\sum_{p=-\frac{M}{2}}^{M/2}\sum_{q=-\frac{M}{2}}^{M/2}\{\exp[-h^2(s_{1p}^2+S_{2q}^2)]-2\cdot\exp[-\frac{1}{2}h^2(s_{1p}^2+s_{2q}^2)]\}\cdot\tilde{V}_{pq}^2$$
$$=-1+2(b_1-a_1)(b_2-a_2)\cdot\sum_{p=1}^{M/2}\sum_{q=1}^{M/2}\{\exp[-h^2(s_{1p}^2+S_{2q}^2)]-2\cdot\exp[-\frac{1}{2}h^2(s_{1p}^2+s_{2q}^2)]\}\cdot\tilde{V}_{pq}^2. \quad (15)$$

将(15)式代入(11)式有:

$$-\frac{1}{2}(1+M_1(h))=(b_1-a_1)(b_2-a_2)\cdot\sum_{p=1}^{M/2}\sum_{q=1}^{M/2}\{\exp[-h^2(s_{1p}^2+S_{2q}^2)]$$
$$-2\exp[-\frac{1}{2}h^2(s_{1p}^2+s_{2q}^2)]\}\cdot\tilde{V}_{pq}^2+n^{-1}h^{-2}(2\pi)^{-1}, \quad (16)$$

由(16)式可确定最佳窗宽 h_{opt}.

综上所述,二维概率密度核窗估计的快速算法分下列五个步骤:

(1) 将样本$\{(x_i,y_i)\}_{i=1}^n$量化为序列$\{W_{kl}\}_{k,l=0}^{M-1}$;

(2) 对 $\{W_{kl}\}_{k,l=0}^{M-1}$作二维 FFT得到 $\{V_{pq}\}_{p,q=-\frac{M}{2}}^{M/2}$;

(3) 由(16)式确定最佳窗宽 h_{opt};

(4) 由(14)式求得 $\{\hat{a}_{pq}\}_{p,q=-M/2}^{M/2}$;

(5) 对 $\{\hat{a}_{pq}\}_{p,q=-M/2}^{M/2}$作二维逆 FFT即得 $\hat{f}_n(t_{1k},t_{2l})$.

完成上述各步的运算量大致为:

(1) $8N$ 次乘法, $12N$ 次加法, $4N\times M^2$ 次判断;

(2) $2M$次 M点的 FFT运算;

(3) 对应每一 h值,需作 $2(M/2)^2$次 exp指数运算, $5(M/2)^2$次乘法, $3(M/2)^2$次加法;

(4) M^2次乘法及指数运算;

(5) $2M$次 M点 FFT运算.

在 $N\geq 100, M\leq 64$时,将上述各步运算合并为乘法运算后,其运算量大致为 $(4N+10\triangle)M^2$次乘法运算,其中$\triangle$为确定最佳窗宽时扫描与迭代的次数.

4 误差分析

为验证算法的有效性与实际误差,本文使用一组二维标准高斯仿真数据进行验算. 根据核窗估计理论,在未知分布为标准高斯分布,窗宽选取最佳的前提下,其积分均方误差可由(6)式推得.

取样本数目 $N=512$,核窗函数为标准高斯窗,则对高斯仿真数据由(7)式其最佳窗宽应为

$$h_{opt}=0.3406,$$

由(6)式其积分均方误差应为

$$MISE(\hat{f}_n)_{h_{opt}}=2.68\times 10^{-3}.$$

从二维标准高斯仿真数据中取 512个二维样本,在 $[-3,3]\times[-3,3]$区域上估算其二维概率分布密度,估算结果与二维标准高斯密度的实际积分均方误差:

$$MISE[\hat{f}_n(x,y)]=2.28\times 10^{-3}.$$

该结果比理论预算的误差还要小,这主要是因为该结果仅计算了 $[-3,3]\times[-3,3]$区域上标准高斯密度与估算结果间的误差. 重要的是该结果与理论误差保持在同一数量级内,这表明该算法是有效的.

参考文献:

[1] SILVERMAN B.W. Density for statistics and data analysis[M]. copyright ©, by John wiley & sons Inc. Canada, 1985.

[2] POSTARE J G. VASSEAR C. A fast algorithm for nonparametric probability estimation [J]. *IEEE Trans-PAMI*, **4** (6): 663-666, 1982.

[3] 何振亚.数字信号处理的理论与应用(下册)[M].北京:人民邮电出版社, 1981.

Fast Algorithm of Window-Estimation in Two-dimensional Probability Density

XIA Chun-hua

(*Shanxi Administration College*, *Computer Centre*, *Taiyuan* 030006, *China*)

Abstract The method of window-estimation of probability density has an extremely high precision when the window-wide is chosen optimally. Among the various methods of choosing the most appropriate window-wide, the least square cross validation is one of the better in performances, whereas it is limited in practical applicationon on account of the complex operations of arithmetic. Based on ensuring the high precision of window-extimation, the fast algorithm of window-estimation in two-dimensional probability density is provided, and the efficiency of this method is proved by means of computer simulation.

Key words fast algorithm; window-estimation; probability density