

基于余弦核函数在 Solve-the-Equation

王俊明¹, 茹 杨¹, 陈 瑜¹, 徐延新²

(1. 哈尔滨理工大学 应用科学学院, 黑龙江 哈尔滨 150080; 2. 哈尔滨师范大学 数学科学学院, 黑龙江 哈尔滨 150025)

摘 要: 核概率密度函数估计是对未知的概率密度函数进行估计的一种方法, 其中关键是对核函数的选取以及窗口宽度的确定. 利用 Solve-the-Equation 方法对窗宽中存在的二阶导数进行推导, 设计迭代算法得出在 Solve-the-Equation 方法下的最优窗口宽度. 在核方法确定的情况下, 利用余弦核函数对未知的概率密度函数进行估计, 并对改进核函数进行估计, 找到最优的余弦核函数, 通过验证得出改进后的核函数其估计精度和平滑度都有所提高.

关键词: 核概率密度函数估计; 核函数; 核方法; 窗口宽度

DOI: 10.15938/j.jhust.2016.01.023

中图分类号: O212.7 **文献标志码:** A **文章编号:** 1007-2683(2016)01-0114-04

WANG Jun-ming¹, RU Yang¹, CHEN Yu¹, XU Yan-xin²

(1. School of Applied Sciences, Harbin University of Science and Technology, Harbin 150080, China;

2. School of Mathematical Science, Harbin Normal University, Harbin 150025, China)

Abstract: Kernel probability density function estimation is a method to estimate unknown probability density function, and the key of kernel probability density function estimation method is the selection of kernel function and the determination of bandwidth. In this paper, the Solve-the-Equation method is used to derive the second derivative of the bandwidth, and the iterative algorithm is designed to calculate the optimal bandwidth. Under the premise of the determinate kernel method, cosine kernel is used to estimate unknown probability density function and make an improvement. Results show that the new kernel function improves the precision and gliding property of kernel density estimation.

Keywords: kernel probability density function estimation; kernel function; kernel method; bandwidth

收稿日期: 2015-01-11

基金项目: 黑龙江省自然科学基金(A201410).

作者简介: 王俊明(1970—), 男, 教授, 博士, E-mail: wangjunming@hrbust.edu.cn;

茹 杨(1991—), 男, 硕士研究生;

陈 瑜(1991—), 女, 硕士研究生.

0 引言

在统计问题中, 概率密度估计的应用范围十分广泛, 可以通过概率密度进行运算和分析相关的统计问题. 本文主要应用非参数估计的方法, 非参数估计^[1-2]也是应用较为普遍的概率密度函数估计方法. 在非参数估计中, 核密度估计法在理论上是比较完善的方法, 因为它最终可以给出概率密度函数的显式形式, 所以被人们广泛地使用. 其中 Parzen 窗核密度估计法^[3]是常用的核方法, 此方法利用核函数叠加的思想去逼近真实的概率密度函数. 因为它对某部分区间估计都可以得到较好精确度^[4], 所以适合整体估计. 但是由于 Parzen 窗方法中出现的二阶导数存在偏差, 计算复杂度较高. 为了减小误差, 利用 Solve-the-Equation 方法^[5]对存在的二阶导数进行近似推导, 以此估计概率密度函数. 这些方法在一定程度上解决了估计曲线精度较差的问题.

核函数的选取以及窗口宽度的取值都影响着核密度估计的精度. 本文在运用两种核方法的条件下, 并对余弦核函数进行改进, 减小其振幅, 使渐进积分均方误差值减小^[6], 让核函数可以适应随机的样本点, 并与常用的余弦核函数估计效果做对比实验, 从而找到最优核函数. 结果表明, 此方法可以较好地估计出真实的概率密度函数.

1 Parzen 方法及窗宽的确定

1.1 Parzen 窗方法基本思想

设 $X_1, X_2, \dots, X_n$ 是从一维总体 X 中抽出的独立同分布样本, X 具有未知的密度函数 $f(x)$, $x \in R$, 则 $f(x)$ 的核密度估计为

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \quad (1)$$

式中: $\hat{f}$ 表示概率密度函数的估计; n 表示样本数量; h 为窗宽; $K(x)$ 为核函数.

选取的核函数要同时满足以下 3 条性质:

- 1) 非负性: $k(x) \geq 0, x \in R$;
- 2) 对称性: $K(x) = K(-x)$, $x \in R$;
- 3) 归一性: $\int_{-\infty}^{+\infty} K(x) dx = 1$.

在核密度估计中, 对于给定的样本, 窗宽 h 是一个重要的参数^[7-12]. 所以只要确定合适的窗口宽度值, 核方法就可以较好的去逼近真实的概率密度函

数. 从式(1)可以看出, 如果窗宽 h 值太大, 那么就会淹没了密度的细节部分, 使得密度估计曲线过于平稳, 就会导致结果的分辨率太低; 反之, 如果窗宽 h 值太小, 随机性的影响就会增加, 那么就会掩盖密度的重要特性, 使得密度估计曲线波动较大, 结果就会变得很不稳定^[13-15]. 因此, 在有限样本个数的约束下, 确定一个最佳窗宽对核概率密度估计是至关重要的.

1.2 窗宽确定

关于窗宽的优化算法主要是基于积分均方误差 MISE (mean integrated square error) 最小化原则.

$$\begin{aligned} \text{MISE}(\hat{f}(x)) &= E \int [\hat{f}(x) - f(x)]^2 dx = \\ &\int E[\hat{f}(x) - f(x)]^2 dx + \int \text{var} \hat{f}(x) dx \end{aligned} \quad (2)$$

式(2)中分解后的第一项为 $\hat{f}$ 的期望值与真实值之间的偏差平方的积分, 简称偏差, 将 $E[\hat{f}(x) - f(x)]$ 记为 $\text{Bias}(\hat{f}(x))$, 第二项为估计值的方差积分, 记为方差. 对上面式子进行求解可得到以下结果^[16-19]:

$$\text{Bias}(\hat{f}(x))^2 \approx \frac{1}{4} h^4 [f''(x)]^2 \int t^2 K(t) dt \quad (3)$$

$$\text{var} \hat{f}(x) \approx \frac{1}{nh} f(x) \int K^2(t) dt \quad (4)$$

$$\begin{aligned} \text{AMISE}(\hat{f}(x)) &= \frac{1}{4} h^4 \left[\int t^2 K(t) dt \right]^2 [f''(x)]^2 + \\ &\frac{1}{nh} f(x) \int K^2(t) dt \end{aligned} \quad (5)$$

而 $\text{MISE}(\hat{f}(X))$ 的最小值可以由 $\text{AMISE}(\hat{f}(X))$ 的最小值近似求得. 所以对式(5)中的 h 进行求导, 并令其一阶导数等于零便得到最优窗宽^[20]

$$h = \left(\frac{\int K^2(t) dt}{n \left(\int t^2 K(t) dt \right)^2 \int [f''(x)]^2 dx} \right)^{\frac{1}{5}} \quad (6)$$

2 Solve-the-Equation 方法和 $\int [f''(x)]^2 dx$ 的推导

在所得的窗宽 h 表达式中, 存在 $f(x)$ 的二阶导数 $f''(x)$ 表示密度估计函数各点的凹凸程度, 如果 $f''(x)$ 直接用上述算法求得的 $\hat{f}(x)$ 二次求导来代

替,会有较大的偏差.因为一般情况下,上述算法求得的 $\hat{f}(x)$ 的密度图像不一定具有良好的光滑性,曲线局部起伏程度可能会很大,虽然 $\hat{f}(x)$ 和 $f(x)$ 的偏差可能不大,但是 $\hat{f}(x)$ 与 $f''(x)$ 则会出现较大的偏差.所以如果能对 $\hat{f}(x)$ 进行适当的处理,就有利于减小 $\hat{f}(x)$ 与 $f''(x)$ 的偏差,使其变得更加平滑.

因此,为了减小 $\hat{f}(x)$ 与 $f''(x)$ 的偏差,给出了一种最优窗口宽度确定的方法,即 Solve-the-Equation 方法.对 $\int [f''(x)]^2 dx$ 进行适当的推导.此方法为直接计算 $f(x)$ 的二阶导数提供了方便,而且在理论和实现上均较为简单.

通过计算可以得出在渐近积分均方误差达到最小的最优 Parzen 窗口宽度值.式(6)中的 $\int K^2(t) dt$ 和 $\int t^2 K(t) dt$ 是可以计算得到的,对于 $\int [f''(x)]^2 dx$ 计算比较复杂,所以本文利用 Solve-the-Equation 方法来对其进行求解.

式子 $\int [f''(x)]^2 dx$ 中的 $f(x)$ 是待估计的真实概率密度函数,其具体形式是未知的,对于 $\int [f''(x)]^2 dx$ 的计算不能依赖 $f(x)$,在 Solve-the-Equation 方法中将式(6)代入式(1)得到 $\hat{f}_h(x)$ 代替 $\int [f''(x)]^2 dx$ 中的 $f(x)$,因此 $\int [f''(x)]^2 dx \approx \int [\hat{f}_h''(x)]^2 dx$.

通过 $\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)$,所以可以得到

$$\hat{f}_h''(x) = \frac{1}{nh^3} \sum_{i=1}^n K''\left(\frac{x-X_i}{h}\right). \quad (7)$$

将式(7)代入 $\int [f''(x)]^2 dx$ 可得

$$\int [f''(x)]^2 dx = \int \left[\frac{1}{nh^3} \sum_{i=1}^n K''\left(\frac{x-X_i}{h}\right) \right]^2 dx = \frac{1}{n^2 h^6} \int \left[\sum_{i=1}^n K''\left(\frac{x-X_i}{h}\right) \right]^2 dx. \quad (8)$$

表达式(8)即为 Solve-the-Equation 方法公式,在此对其进行化简,由于

$$\left[\sum_{i=1}^n a_i \right]^2 = 2 \sum_{i=1}^n \sum_{j<i}^n a_i a_j + \sum_{i=1}^n a_i^2.$$

所以,

$$\frac{1}{n^2 h^6} \int \left[\sum_{i=1}^n K''\left(\frac{x-X_i}{h}\right) \right]^2 dx =$$

$$\frac{1}{n^2 h^6} \left\{ \sum_{i=1}^n \int \left[K''\left(\frac{x-X_i}{h}\right) \right]^2 dx + 2 \sum_{i=1}^n \sum_{j<i}^n \int \left[K''\left(\frac{x-X_i}{h}\right) K''\left(\frac{x-X_j}{h}\right) \right] dx \right\}.$$

其中令

$$R_i = \int \left[K''\left(\frac{x-X_i}{h}\right) \right]^2 dx,$$

$$R_{ij} = \int \left[K''\left(\frac{x-X_i}{h}\right) K''\left(\frac{x-X_j}{h}\right) \right] dx.$$

选取 $K(x)$ 为余弦核函数,其表达式通式为

$$K(x) = \begin{cases} A \cos \alpha x, & x \in [-1, 1] \\ 0, & \text{其他.} \end{cases}$$

其中 A 和 α 为待定常数,当 $A = \pi/4$ 和 $\alpha\pi/2$ 时常用的余弦核函数.因为核函数具有归一性,所以得到 A 和 α 的关系表达式

$$2 \frac{A}{\alpha} \sin \alpha = 1.$$

由关系式可知, α 的取值影响着 A 的值,为了让估计值更加精确,图像更加光滑,本文对余弦核函数进行改进.考虑减小振幅 A 的值,使峰值降低,同时减小渐近积分均方误差值.因此,设这一类余弦核函数在 Solve-the-Equation 方法下窗宽为 h_g ,其表达式为

$$h_g = \left\{ \frac{(\alpha^6 + \alpha^5 \sin \alpha \cos \alpha)}{4(\alpha^2 \sin \alpha + 2\alpha \cos \alpha - 2 \sin \alpha)^2 M} \right\}^{\frac{1}{5}} \quad (9)$$

式中:

$$M = \frac{1}{nh^6} \left\{ 2 \sum_{i=1}^n \sum_{j<i}^n [R_{ij}] + \sum_{i=1}^n [R_i] \right\}.$$

3 设计迭代并验证算法

由于在 Solve-the-Equation 方法下最优窗宽是无法直接得到的,因此要设计迭代算法寻找最优窗宽值.构造函数 $h_g = F(h)$,将式(9)代入构造的函数中,

计算最优窗口宽度值得迭代算法如下:

Step1: 设 $h_g = F(h)$,给定一个初值,其 $h_0 = s$,对于 s 的初值确定之后会有阐述,然后初始化一个极小值 κ .

Step2: 当 $|h_1 - h_0| > \kappa$ 时,循环执行下面步骤:

将 h_1 的值赋予 h_0 ,执行 $h_1 = h_0$;令 $h_{k+1} = \frac{F(h_k) + h_k}{2}$; $k = 1, 2, 3 \dots h$.直到 $|h_{k+1} - h_k| \leq \kappa$ 终止.

Step3: 返回窗口宽度的最优值 $h_g = h_{k+1}$.

窗宽表达式是依赖于函数 $f(x)$ 的,但是不知道它的具体形式,通常都假设它服从正态分布,经过计算得 $h = 1.60\sigma n^{-\frac{1}{5}}$,其中 σ 是样本观察值的标准差,

n 是样本观察的个数,对于 s 的确定通常认为

$$h = s = 1.06\sigma n^{-\frac{1}{5}}.$$

实验中随机生成符合标准正态分布的样本 1000 个,利用余弦核函数进行试验,得到如下结果

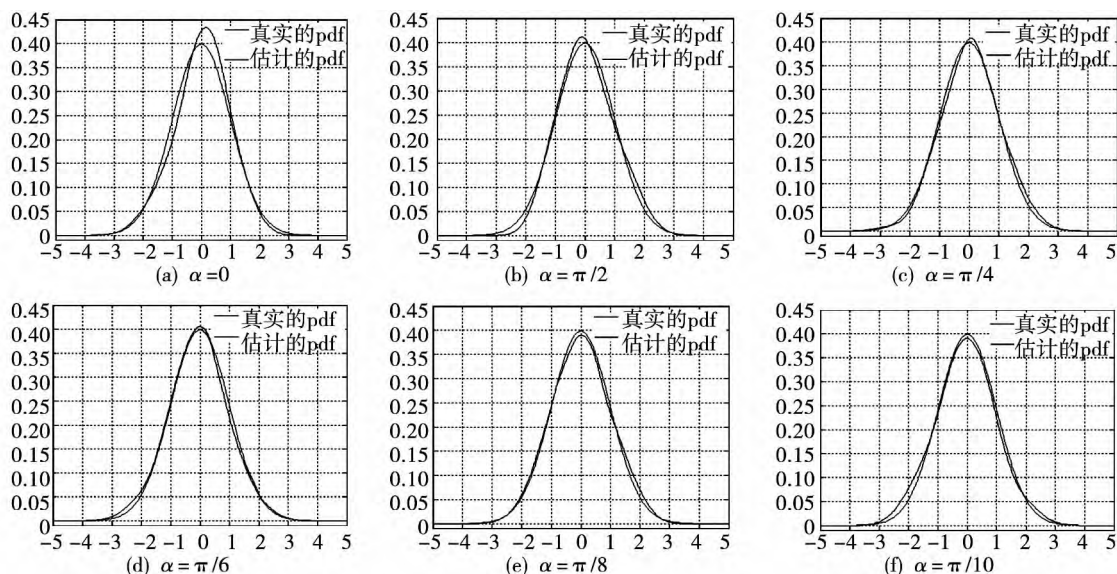


图 1 取不同 α 值的估计效果对比

当 $\alpha = 0$ 时,余弦核函数表示为均匀核函数,所以均匀核函数是这类余弦核函数的一种特殊情况,但是由图 1(a) 可以看出均匀核函数的估计效果并不好,偏差也是最大的. 其中 $\alpha = \pi/2$ 为常用的余弦核函数,图 1(b) 可以看出其估计的曲线与真实的概率密度曲线有差异,波峰较为突出. 其中图 1(c)、图 1(e) 和图 1(f) 的估计效果都要好于图 1(b). 但是,图 1(d) 的估计效果最好,估计曲线与真实的概率密度函数曲线也最为接近. 因此,当样本量确定的情况下, $\alpha = \pi/6$ 时为最优的余弦核函数.

表 1 标准正态分布下,此类余弦核函数各参数比较

α 的等值	A 的等值	h_g 的值	AMISE(10^{-6})
0	0.500 0	0.658 4	9.849 4
$\frac{\pi}{2}$	0.785 4	0.525 1	8.824 0
$\frac{\pi}{4}$	0.555 4	0.514 6	8.712 6
$\frac{\pi}{6}$	0.523 6	0.490 1	8.484 9
$\frac{\pi}{8}$	0.513 1	0.509 2	8.563 8
$\frac{\pi}{10}$	0.508 3	0.510 5	8.591 3

从表 1 中也可以看出当 $\alpha = \pi/6$ 时,此时的余弦核函数的窗宽最佳,并且渐进积分均方误差的值也是最小的,进一步验证了此时的核函数为最优核函数.

4 结 论

由上述可知,在运用 Solve-the-Equation 方法所确定的最优窗宽解决了传统窗宽中二阶导数存在的误差,并且对未知的密度函数的估计效果较好. 在核方法确定的情况下,对余弦核函数进行改进实验,使其适应随机的样本数据,也得到了更好的估计效果. 因此,在核概率密度函数估计中,使得核函数的选择范围变得更加广泛.

参 考 文 献:

- [1] 董晓芳, 张良勇. 基于排序集抽样的分布估计 [J]. 统计与决策, 2010, 18: 30-31.
- [2] 盛骤, 谢式千, 潘承毅. 概率论与数理统计 [M]. 3 版. 北京: 中国高等教育出版社, 2007: 286-306.
- [3] EMANUEL PARZEN. On Estimation of a Probability Density Function and Mode [J]. Annals of Mathematical Statistics, 1962, 33(3): 1065-1076.
- [4] 叶彩园, 吴群英, 刘振. 随机删失数据下核密度估计相合性 [J]. 华侨大学学报, 2013, 34(5): 591-595.
- [5] LUS A. ALEXANDER. A Solve-the-Equation Approach for Unidimensional Date Kernel Bandwidth [J]. Selection, 2008(5): 1057-1067.

(下转第 122 页)

参考文献:

- [1] POLOVINKIN E S. Convex Bodies of Constant Width [J]. Dokl. Math., 2004, 70: 560–562.
- [2] HOWARD R. Convex Bodies of Constant Width and Constant Brightness [J]. Adv. Math., 2006, 204: 241–261.
- [3] VRECICA S. A Note on Sets of Constant Width [J]. Nouvelle série. Tome., 1981, 43(29): 289–291.
- [4] LACHAND – ROBERT T, OUDET É. Bodies of Constant Width in Arbitrary Dimension [J]. Math. Nachr., 2007, 280(7): 740–750.
- [5] MALUTA E, PAPINI P L. Diametrically Complete Sets and Normal Structure [J]. J. Math. Anal. Appl., 2015, 424: 1335–1347.
- [6] MORENO J P. Porosity and Unique Completion in Strictly Convex Spaces [J]. Math. Z., 2011, 267: 173–184.
- [7] POLOVINKIN E S, SIDENKO S V. The Addition of Subsets to Constant Width Bodies, Physics and Mathematics [J]. Kazan. Gos. Univ. Uchen. Zap. Ser. Fiz. – Mat. Nauki., 2006, 148(2): 132–143.
- [8] 吴森林, 计东海. 正交性和内积空间的一些特征 [J]. 哈尔滨理工大学学报, 2004, 9(6): 113–115.
- [9] JI DONGHAI, WU SENLIN. Quantitative Characterization of the Difference between Birkhoff Orthogonality and Isosceles Orthogonality [J]. J. Math. Anal. Appl., 2006, 323(1): 1–7.
- [10] JAMES R C. Orthogonality in Normed Linear Spaces [J]. Duke Math. J., 1945, 12(2): 291–302.
- [11] ALONSO J, MARTINI H, WU SENLIN. On Birkhoff Orthogonality and Isosceles Orthogonality in Normed Linear Spaces [J]. Aequationes Math., 2012, 83: 153–189.
- [12] ALONSO J, BENÍTEZ C. Orthogonality in Normed Linear Spaces: a survey, II, Relations between Main Orthogonalities [J]. Extracta Math., 1989, 4(3): 121–131.
- [13] JAMES R C. Orthogonality and Linear Functionals in Normed Linear Spaces [J]. Trans. Amer. Math. Soc., 1947, 61: 265–292.
- [14] 计东海, 陶冶. 度量椭圆与欧氏平面的一个特征性质 [J]. 哈尔滨理工大学学报, 2012, 17(5): 84–86.
- [15] 倪柏竹, 何婵, 计东海. 一致非方空间的新特征性质 [J]. 哈尔滨理工大学学报, 2011, 16(1): 107–109.
- [16] ROBERTS B D. On the Geometry of Abstract Vector Spaces [J]. Tohoku Math. J., 1934, 39: 42–59.
- [17] MARTINI H, SWANEPOEL K J, WEIB G. The Geometry of Minkowski Spaces—a Survey. Part I [J]. Expo. Math., 2001, 19(2): 97–142.
- [18] 定光桂. 巴拿赫空间引论 [M]. 北京: 科学出版社, 2008: 160–161.

(编辑: 王 萍)

(上接第117页)

- [6] SANTANU DUTTA. Estimation of the MISE and the Optimal Bandwidth Vector of a Product Kernel Density Estimate [J]. Journal of Statistical Planning and Inference, 2010, 141(5): 1817–1831.
- [7] 叶菲. 改进核密度估计确定最优分组方法研究 [J]. 计算机应用与软件, 2013, 6: 292–294.
- [8] 宇世航, 赵世舜. 核实数据下的递归核密度估计 [J]. 吉林大学学报: 理学版, 2012, 50(5): 924–930.
- [9] 张敏爱, 李路亮, 张晓琴. 基于核密度估计的实验室能力验证 [J]. 中北大学学报: 自然科学版, 2013, 34(5): 586–589.
- [10] SANTANU DUTTA. Estimation of the MISE and the Optimal Bandwidth Vector of a Product Kernel Density Estimate [J]. Journal of Statistical Planning and Inference, 2010, 141(5): 1817–1831.
- [11] 尹训福, 郝志峰, 杨晓峰. 快速核密度估计算法研究进展 [J]. 计算机工程与应用, 2007, 43(31): 1–4.
- [12] 任文军, 宋向东. 核密度估计中的递归方法选择窗宽及其应用 [J]. 长春大学学报, 2009, 19(1): 23–26.
- [13] 黎运发, 黄明辉. 核密度估计逐点最优窗宽选择的改进 [J]. 统计与决策, 2009, 30(2): 231–235.
- [14] WANG SHITONG, WANG JUN. Kernel Density Estimation, Kernel Methods and Fast Learning in Large Data Sets [J]. IEEE transactions on cybernetics, 2013, 4(2): 43–44.
- [15] JOHAN LIM, MIN CHEN, SANGUN PARK, et. al. Kernel Density Estimator From Ranked Set Samples [J]. Communications in Statistics-Theory and Methods, 2014, 43(12): 2156–2168.
- [16] 于传强, 郭晓松. 基于估计点的滑动窗宽核密度估计算法 [J]. 兵工学报, 2009, 30(2): 231–235.
- [17] 邓飏, 于传强. 基于估计点双窗宽核密度估计算法 [J]. 仪器仪表学报, 2006, 32(3): 31–33.
- [18] 田玉刚, 杜渊会, 何义丰. 变窗宽核加权估计下的变形趋势拟合方法 [J]. 测绘科学技术学报, 2012, 29(1): 12–15.
- [19] DAI PEHNGJIE, SUN ZHIHUA, WANG PENG. Model Checking for General Linear Error-in-Covariates Model with Validation Data [J]. Acta Mathematica Sinica, 2010, 23(6): 1153–1166.
- [20] 乔俊峰, 朱虹, 史静. 一种快速核密度估计背景建模方法 [J]. 计算机工程与应用, 2012, 48(5): 192–193.

(编辑: 关 毅)