ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 3,May 2025,pp. 711－729
DOI:10. 16337/j. 1004-9037. 2025. 03. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

针对模相近数据的启发式核密度估计器

何玉林[1,2], 陈纯佳[2], 黄哲学[1,2], 李俊杰[2], FOURNIER-VIGER Philippe[2]

(1. 人工智能与数字经济广东省实验室(深圳),深圳 518107;2. 深圳大学计算机与软件学院,深圳 518060)

摘　要: 区别于经典的基于Parzen窗口法的概率密度函数估计器构建策略,提出了基于近邻误差度量函数的启发式核密度估计器(Heuristic kernel density estimator, HKDE),用以提升对模相近数据概率密度函数拟合的准确性。首次从数据不确定性和模型不确定性的角度分析了传统核密度估计器解决模相近数据概率密度函数估计问题时的缺陷:利用概率密度值对于直方图箱宽参数的收敛性确定观测数据的启发式概率密度值,降低数据概率密度值计算的不确定性;基于启发式概率密度值构建用于确定核密度估计器最优带宽的目标函数,降低最优带宽优化过程中的不确定性。在18个模相近数据集上对新估计器HKDE的可行性、合理性和有效性进行了系统性的验证。实验结果表明,与7种具有代表性的概率密度函数估计器相比,HKDE能够获得更加优异的概率分布近似表现,具有比其他估计器更低的估计误差,能够确定出更接近真实值的概率密度函数估计值。

关键词: 核密度估计器;模相近观察值;不确定性;启发式概率密度值;直方图箱宽

中图分类号: TP391.9　　**文献标志码:** A

Heuristic Kernel Density Estimator for Modal-Proximity Data

HE Yulin[1,2], CHEN Chunjia[2], HUANG Zhexue[1,2], LI Junjie[2], FOURNIER-VIGER Philippe[2]

(1. Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen), Shenzhen 518107, China; 2. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China)

Abstract: Different from the classical probability density estimator construction strategies based on the Parzen window method, we propose a heuristic kernel density estimator (HKDE) based on nearest neighbor error measurement function, to improve the accuracy of fitting probability density function of modal-proximity data. From the perspective of data and model uncertainties, we analyze the defects of traditional kernel density estimators in solving the problem of probability density estimation of modal-proximity data. The heuristic probability density values that can reduce the uncertainty of observed data are obtained by referring to the convergence of probability density values with respect to the histogram box width. Based on the heuristic probability density value, we construct the sophisticated objective function to determine the optimal bandwidth for kernel density estimator by reducing the model uncertainty. Extensive experiments on 18 modal-proximity datasets are conducted to validate the feasibility, rationality and effectiveness of the designed HKDE. Results show that HKDE can obtain a better approximate

基金项目:广东省自然科学基金面上项目(2023A1515011667);深圳市基础研究面上项目(JCYJ20210324093609026);广东省基础与应用基础研究基金粤深联合基金重点项目(2023B1515120020);深圳市科技重大专项项目(KJZD20230923114809020)。
收稿日期:2024-05-29;**修订日期:**2024-10-09

performance of probability distribution than seven existing representative probability density function estimators. HKDE has lower estimation error and closer probability density function estimates to the real density values than other kernel density estimators.

Key words: kernel density estimator (KDE); modal-proximity data; uncertainty; heuristic probability density value; histogram box width

引　言

概率密度函数(Probability density function,PDF)估计是利用统计学知识和给定的观测值估计未知分布样本总体的PDF。现实生活中,PDF在进行数据分析和建模时发挥着重要作用,例如,高速列车信号频谱分析[1-2]、金融风险度量[3]以及物理现象模拟和预测[4-6]等。PDF估计技术在受到业界广泛关注的同时,也面临着高精度要求的挑战。目前,PDF估计根据其模型建立时是否依赖特定参数分为有参估计和无参估计。无参估计具备完全从数据出发、无需假设数据服从某种特定概率分布的特性,因此更为常见且通用。核密度估计(Kernel density estimator,KDE)是一种用于估计连续型随机变量PDF的无参估计统计方法,其基本思想是对每个数据点对应的核函数进行加权平均,最终得到数据整体对应的PDF。在KDE估计过程中,窗口宽度参数的选择是一个关键问题,其决定了估计PDF的平滑程度:若窗口宽度参数过大会导致估计的PDF过度平滑;若过小会导致估计的PDF波动性较大[7]。

在核函数已知时,选择的窗口宽度参数应使实际概率密度和估计概率密度之间的误差最小,需要通过构造可行的误差目标函数来选取最优窗口宽度。目前流行的误差度量标准是积分平方误差(Integrated squared error,ISE)和平均积分平方误差(Mean ISE, MISE)[8]。ISE计算每个数据点对应的估计PDF和真实PDF的差值平方,再在整个数据范围上积分,使得KDE估计的PDF尽可能接近真实的PDF;而MISE是基于ISE在不同样本上的平均作出对整体估计效果的评价,能够获得更为稳定和通用的PDF估计。

尽管经典的KDE在应用中取得了良好的表现,在一定程度上提高了PDF估计的准确性,然而通过深入分析发现,经典的KDE在处理模相近数据时的表现不尽如人意。与常见的数据类型不同,模相近数据所服从的概率分布通常是具有多个局部最大值的PDF,且这些局部最大值非常接近。近年来,随着大数据技术的快速普及,模相近数据在实际应用中出现的频率越来越高,对模相近数据高质量统计分析和信息挖掘的需求日益迫切。例如,按日统计的风力发电量数据就是模相近数据。经典的KDE在处理模相近数据时存在无法有效处理数据不确定性和模型不确定性的缺陷。数据不确定性,即模相近的数据样本点会出现重叠现象,在估计过程中难以区分数据来源于哪个模;模型不确定性,MISE误差度量和ISE误差度量都涉及到估计PDF和真实PDF,在最小化MISE和ISE时,展开式中都包含未知的估计PDF,这给最优窗口宽度的选取带来了较大的不确定性。由于上述缺陷的存在,当多个模的数据样本发生重叠时,使用经典KDE进行PDF估计会明显影响估计的准确性。如何突破不确定性对KDE性能的限制是目前研究模相近数据概率密度函数估计问题的关键。

为解决上述数据不确定性和模型不确定性所带来的KDE构建缺陷,本文提出了一种基于启发式概率密度值的核密度估计(Heuristic KDE, HKDE)。该方法利用数据样本概率密度值对于直方图箱宽参数的收敛性,计算确定观测数据的启发式概率密度值,用其替代误差度量目标函数中各个样本点的真实PDF值,降低数据不确定性;随后利用基于启发式概率密度值构建的目标函数来确定HKDE的最优窗口参数,以进一步降低模型不确定性;最后在18个模相近数据集上对HKDE的可行性、合理性和有效

性进行了系统性的验证。实验结果表明,与最小二乘交叉验证(Least squares cross-validation,LSCV)[9]、有偏交叉验证(Biased cross-validation,BCV)[10]、拇指原则(Rule of thumb,RoT)[11]、Scott's RoT[12]、Silverman's RoT[13]和最大似然交叉验证(Max-likelihood cross-validation,MLCV)[14]这6种具有代表性的经典概率密度函数估计器相比,并与常用于多模PDF估计的高斯混合模型(Gaussian mixture model,GMM)[13]相比,HKDE在多模情况下能够获得优异的概率分布近似表现,从而证实了HKDE是一种能够处理模相近数据PDF估计问题的高效概率密度估计器。

1 相关工作

1.1 KDE数学模型

KDE是一种用于估计随机变量PDF的非参方法,通过对每个数据点周围的邻域应用核函数并进行加权求和来估计概率密度[15],从而提供对数据分布的平滑估计。在KDE框架下,核函数的选择以及带宽的选取对PDF估计的准确性有重要影响。本部分所涉及到的数学符号及其含义已经总结在表1中。

表1 数学符号及其含义

Table 1 Mathematical symbols and their meanings

符号	含义
x	待估计的数据点
$\hat{f}(x)$	点x处的估计PDF
$f(x)$	点x处的真实PDF
N	样本点个数
h	窗口宽度
$K(\cdot)$	核函数
$\{X_1,X_2,\cdots,X_N\}$	样本点集合

一维情况下KDE的数学模型可以表示为

$$\hat{f}(x)=\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x-X_i}{h}\right) \tag{1}$$

在式(1)中,窗口宽度h与核函数$K(\cdot)$的选取决定了KDE的表现,而对于已选定$K(\cdot)$的KDE模型,其目标就是确定最优的h,使得$\hat{f}(x)$与真实值$f(x)$之间的误差达到最小。由于Gaussian核函数具有完美的数学特性[16],例如连续性和可导性等,在实际中获得了广泛的应用。其他类型的核函数,如均匀核、三角核、余弦核等亦有使用,但报道较少。因此,在本文选取了Gaussian核函数,$K(\cdot)$的表达式为

$$K(\mu)=\frac{1}{\sqrt{2\pi}}\left(-\frac{\mu^2}{2}\right) \tag{2}$$

式中μ为随机变量的取值,$\mu\in(-\infty,+\infty)$。

1.2 KDE方法优化

当核函数确定后,窗口宽度h的选取在很大程度上影响着KDE模型的表现[7]。最优窗口宽度参数要使得估计PDF与真实PDF的误差最小,如何度量这个误差成了关键。目前各种KDE方法中最流行的误差度量标准为ISE和MISE,其数学表达式分别为

$$\mathrm{ISE}(h)=\int_{-\infty}^{+\infty}\left[\hat{f}(x)-f(x)\right]^2\mathrm{d}x \tag{3}$$

$$\mathrm{MISE}(h)=E\left[\int_{-\infty}^{+\infty}\left[\hat{f}(x)-f(x)\right]^2\mathrm{d}x\right] \tag{4}$$

交叉验证(Cross-validation,CV)方法中多数是基于ISE误差度量确定带宽,其发展最早可以追溯到1974年,Habbma等[17]发明了一个伪似然CV方法。1984年,Bowman[9]提出最小二乘交叉验证LSCV,也称为无偏交叉验证(Unbiased CV, UCV),旨在最小化ISE目标函数来确定最优窗口宽度参

数。在后续发展中为了提高稳定性，1991年，Chui[18]通过修改观测值的周期图来稳定残差平方和，提出更为稳定的带宽选择器。1992年，针对交叉验证Stute[19]提出了修正交叉验证(Modified CV, MCV)，其借助Hajek投影近似真实PDF来优化ISE目标函数。1998年，Hart等[20]因回归问题提出了更为稳定的单侧交叉验证(One-sided CV, OSCV)，通过优化构造新定义的单侧核目标函数得到最优窗口参数。

与交叉验证方法相比，Plug-in插件法倾向于最小化MISE目标函数，并且其波动性较小。1986年，Silverman[11]推广了著名的经验法则——拇指原则RoT，该方法基于数据服从正态分布的假设来近似优化MISE，并因其估计速度较快得到了广泛应用。然而，由于其固有的局限性，该方法可能会忽略数据特征或其他重要因素，因此后续引入了各种改进方法，例如Park[21]、Hall[22]等都是在MISE目标函数的基础上对其进行了改进。1987年，Scott等[10]提出了有偏交叉验证BCV，其基于渐近MISE设计了一个平滑的目标函数以优化窗口参数的选取。1989年，Taylor[23]提出一种基于重采样数据的Bootstrap方法，该方法可以根据不同情况选择最小化目标函数MISE或者ISE来得到最优带宽。1991年，Sheather等[24]基于平滑交叉验证(Smoothed CV, SCV)开发了一种更为可靠的插件法，其拥有在高阶核中仍保持收敛速度较快的优点。1994年，Kim等[25]认为基于精准的MISE能获得理论上的最佳带宽估计器，但在后续的实验中证实其效果一般。

通常，由多个单峰PDF所组成的多模PDF中存在多个局部最大值，每个单峰PDF对应于一组非独立且同分布的随机变量，这意味着在多模分布中，其PDF会更加多样化、复杂化。到目前为止，针对多模概率密度估计的研究方法较少，主要是利用统计矩和GMM[13]来实现。2018年，Rajan等[26]提出了一种相对较新的改进参数分布拟合技术，即矩约束最大熵法，其克服了假设分布是单峰的缺点。2019年，Zhang等[27]基于GMM针对多模态分布问题，提出了一种高精度概率不确定性传播方法。2022年，Li等[28]提出了一种基于分数阶矩的最大熵估计方法，并结合非线性变换和多峰识别方法来提高估计精度。

2 预备知识

为了更加清晰地展示本文提出的针对模相近数据所构建的基于启发式概率密度值的KDE的细节，本节将对模相近数据、数据不确定性和模型不确定性的相关概念进行简要介绍。

2.1 模相近数据

本文所研究的“模相近”数据通常在多模分布下出现。多模PDF具有多个局部最大值，当两个相邻模的均值相距较近或方差差距很大时，生成的仿真样本点会呈现相邻两个模的样本点重合度较高的特点。为直观展示“模相近”数据，使用基于高斯混合分布的多模概率密度函数$f(\boldsymbol{x})=\sum_{t=1}^{K}\boldsymbol{w}_t\times G(\boldsymbol{x};\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)$来生成仿真样本，其中$\boldsymbol{w}=[w_1,w_2,\cdots,w_K]$为权重，$\boldsymbol{\mu}=[\mu_1,\mu_2,\cdots,\mu_K]$为均值，$\boldsymbol{\Sigma}=[\Sigma_1,\Sigma_2,\cdots,\Sigma_K]$为协方差。假设现有下述两个分布的仿真样本点，其中一维分布包含200个样本点，二维分布包含500个样本点：(1) 1维2模概率密度函数$f_{12}(\boldsymbol{x})$：$\boldsymbol{\Sigma}=[[0.23][2.25]]$，$\boldsymbol{\mu}=[0\ 2]$，$\boldsymbol{w}=[0.57\ 0.43]$；(2) 2维2模概率密度函数$f_{22}(\boldsymbol{x})$：$\boldsymbol{\Sigma}=[[[7.63\ 0][0\ 7.63]][[0.56\ 0][0\ 0.56]]]$，$\boldsymbol{\mu}=[[3.07\ 3.83][3.70\ 4.46]]$，$\boldsymbol{w}=[0.74\ 0.26]$。

图1给出了1、2维模相近的数据样本点集合示意图。可以看见两个模的样本点均出现了交集，交集越多表示模距越近。

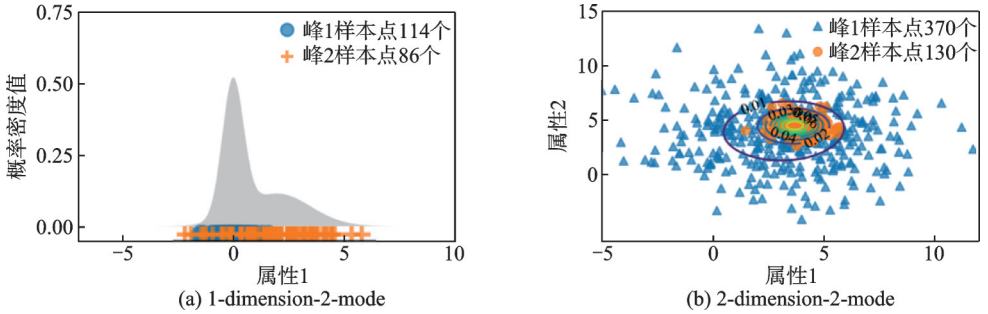

图1 模相近高斯分布仿真样本点示意图

Fig.1 Schematic of simulated samples from modal-proximity Gaussian distribution

2.2 数据不确定性

数据不确定性,即模相近的数据样本点会出现重叠现象,在估计过程中难以区分数据来源于哪个模,所以在多模PDF估计中鲜少利用KDE来实现。多模分布的数据不确定性具体表现如下:

(1)在多模分布中,不同模之间存在一个过渡区域。在这些区域内,样本点可能同时受到多个模的影响,无法明确归属某一个模。例如,在图1中,两个相邻模之间,过渡区域的数据点同时具有这两个模的特征。在多模分布中,过渡区域越多或过渡区域范围越大,其PDF复杂性和多样性更为明显,给KDE进行PDF估计带来了较大的困难。

(2)多模分布中的过渡区域通常表现为平滑过渡,而不是突变。这意味着从一个模到另一个模的PDF值是逐渐变化,这进一步增加了归属的模糊性。例如,GMM中的每个高斯构件在边界区域会相互重叠,使得这些区域的数据点具有多种特性。在图1的一维分布中可以看出真实PDF两个模的交界处变化平缓,在二维分布中两个模更是无法直观区分开。而高维数据通常具有更高的复杂性,各个维度之间可能存在复杂的相关性。某些维度上的相似性会掩盖样本点在其他维度上的差异,使得数据点具体属于哪个模变得更加不明确,更易出现数据不确定性。

2.3 模型不确定性

前文所提到的MISE误差度量和ISE误差度量都涉及到估计PDF和真实PDF。真实PDF值指的是目标分布的实际密度函数。在理论分析中,通常假设其是已知的,但在实际应用中往往是未知的,还是需要通过估计方法来近似真实PDF。因此,在KDE的估计过程中涉及到的估计PDF和真实PDF二者实际上均为近似值,这给最优窗口宽度的选取带来了较大的不确定性。

将ISE的计算公式(3)进行展开

$$\mathrm{ISE}(h)=\int_{-\infty}^{+\infty}\left[\hat{f}(x)-f(x)\right]^2\mathrm{d}x=\int_{-\infty}^{+\infty}\left[\hat{f}(x)\right]^2\mathrm{d}x-2\int_{-\infty}^{+\infty}\hat{f}(x)f(x)\mathrm{d}x+\int_{-\infty}^{+\infty}\left[f(x)\right]^2\mathrm{d}x \quad (5)$$

由式(5)可以发现展开式中的第3项与待估计的窗口参数h无关,使得式(5)前两项之差取得最小值的窗口参数必然能使式(5)取得最小值。现存的基于ISE的传统KDE方法都是优化前两项求得最优窗口参数,而直接忽略掉第3项。但第2项中仍然包含真实PDF的计算,为了获得优化的窗口参数,通常需要利用近似技术对其进行替换。

再将MISE的计算式(4)进行展开

$$\mathrm{MISE}(h)=E\left[\int_{-\infty}^{+\infty}\left[\hat{f}(x)-f(x)\right]^2\mathrm{d}x\right]=\int_{-\infty}^{+\infty}E\left[\hat{f}(x)-E\left[\hat{f}(x)\right]\right]^2\mathrm{d}x+\int_{-\infty}^{+\infty}E\left[E\left[\hat{f}(x)\right]-f(x)\right]^2\mathrm{d}x \quad (6)$$

可以发现式(6)的第2项同样与真实PDF相关,因此也需要对其进行近似以确定最优窗口宽度参数。

式(5)和式(6)反映的共性问题是确定未知窗口宽度参数的目标函数中包含了未知的真实PDF,这就导致了“利用未知PDF确定未知窗口宽度”现象的发生,真实PDF的近似增加了窗口宽度参数优化的不确定性。

3 模相近核密度估计器HKDE

本节将详细介绍针对模相近分布的PDF估计问题而设计的HKDE,其中包括启发式概率密度值的确定和目标函数的构建:利用收敛性确定数据样本点的启发式概率密度值,降低过渡区域数据样本点概率密度值计算的数据不确定性;利用新的目标函数去得到最优窗口参数,尽可能减少误差度量计算方法给窗口参数估计带来的模型不确定性。

3.1 启发式概率密度值

未知密度函数的估计方法大多数都依赖于样本点落入观察值中心区域**R**中的概率,即$P=\int_{\mathrm{R}} p(x)\mathrm{d}x$。当区域**R**很小时样本点落在区域内的概率波动也小,概率计算公式近似为$P=p(x)V$,其中V表示区域**R**的空间体积大小。假设样本点均独立且服从同分布,则在D维情况下,点x处估计的启发式概率密度值可以表示为

$$f_{\mathrm{heu}}\left(X_i,\hat{h}\right)=\frac{n_i}{N\cdot\left(\hat{h}/2\right)^D} \tag{7}$$

式中:$\left(\hat{h}/2\right)^D$为边长为$\hat{h}$的区域**R**空间(或$\hat{h}$-邻域)体积大小;N表示样本集大小;n_i表示落在以样本点X_i为中心的$\hat{h}$-邻域内样本点的个数。显然,参数$\hat{h}$的选取直接影响估计值的准确性,过小过大都会导致估计的PDF误差偏大。

经典的非参估计方法包括K_N近邻估计法、Parzen窗[29]和直方图[30],其核心都是基于式(7)来进行概率密度估计。不同的是,K_N近邻估计是先确定n的大小,再对区域**R**的大小进行动态调整;Parzen窗和直方图是先确定区域**R**的大小,再得到落在区域内的样本点个数。在本文中,使用一种基于概率密度值之和对箱宽参数$\hat{h}$的收敛性方法来确定式(7)中的参数$\hat{h}$,从而计算出用于后续优化算法的样本点启发式概率密度值。具体的实现过程如算法1所示。

算法1 启发式概率密度值确定算法

输入:数据集DataSet$=\{X_1,X_2,\cdots,X_N\}$、窗口参数大小的上限max、阈值ξ和数据维度D

输出:各样本点对应的启发式概率密度值

(1) 找到合适的启发式窗口参数$\hat{h}$

for $h=0.01$ to max do

计算窗口参数h所对应的全部样本点启发式概率密度值之和P_{sum};

找出相邻两个P_{sum}的变化幅度小于阈值ξ所对应的窗口参数$\hat{h}$;

end for

返回所找到的合适窗口参数$\hat{h}$;

(2) 计算启发式概率密度值

计算D维下$\hat{h}$-邻域的空间体积大小$\left(\hat{h}/2\right)^D$;

for $X_i \in$ DataSet do

计算在点X_i的$\hat{h}$-邻域内的样本点个数;

计算X_i对应的启发式概率密度值$f_{\text{heu}}\left(X_i, \hat{h}\right)$;

end for

(3) 返回样本点以及其对应的启发式概率密度值。

对算法1中的关键步骤做如下解释说明。

(1) 为了寻找合适的R空间边长以确定启发式窗口参数数值$\hat{h}$,从h的最小值0.01开始,逐步增加到最大值max(考虑到计算时间与计算质量之间的平衡,本文选用max=1.00)。选取$\tau=0.01$作为遍历的步长,使得在遍历的过程中逐步逼近最佳参数的同时遍历时间不会过长。

(2) 因为所有数据样本点的概率密度值之和对箱宽参数具有收敛性(这一结论将在后续的可行性实验验证中得到证实),故在遍历时计算第j次循环所对应的箱宽参数$\hat{h}_j$下所有数据样本点的启发式概率密度值之和,即$P_{\text{sum_}j}=\sum_{i=1}^{N} f_{\text{heu}}\left(X_i, \hat{h}_j\right)$。当满足$|P_{\text{sum_}j+1}-P_{\text{sum_}j}|<\xi$时,可以确定合适的$\hat{h}=\hat{h}_j$,其中阈值$\xi$的大小可以根据维度的增加而适当扩大。

(3) 在选取合适的窗口参数数值$\hat{h}$后,计算数据样本点X的$\hat{h}$-邻域内样本点个数,具体的邻域是指与点X每个维度的距离都小于等于$\frac{\hat{h}}{2}$的空间,最后利用式(7)计算各样本点的启发式概率密度值$f_{\text{heu}}\left(X, \hat{h}\right)$。

对于数据样本点的概率密度值之和对箱宽参数的收敛性,此处给出如下简要分析说明。对于相邻的两个启发式窗口参数$\hat{h}_{j+1}$和$\hat{h}_j$,对应的所有数据样本点的启发式概率密度值和的差值为

$$|P_{\text{sum\_}j+1}-P_{\text{sum\_}j}|=\left|\sum_{i=1}^{N} f_{\text{heu}}\left(X_i, \hat{h}_{j+1}\right)-\sum_{i=1}^{N} f_{\text{heu}}\left(X_i, \hat{h}_j\right)\right|=\left|\sum_{i=1}^{N}\left[\frac{n_i^{(j+1)}}{N\cdot\left(\hat{h}_{j+1}/2\right)^D}-\frac{n_i^{(j)}}{N\cdot\left(\hat{h}_j/2\right)^D}\right]\right| \tag{8}$$

样本点X_i的h-邻域内样本点个数相同时,即$n_i^{(j+1)}=n_i^{(j)}=n_i$时,式(8)可被转化为

$$|P_{\text{sum\_}j+1}-P_{\text{sum\_}j}|=\left(\sum_{i=1}^{N}\frac{n_i}{N}\right)\cdot\left|\left[\frac{1}{\left(\hat{h}_{j+1}/2\right)^D}-\frac{1}{\left(\hat{h}_j/2\right)^D}\right]\right| \tag{9}$$

式中:$\hat{h}_j$和$\hat{h}_{j+1}$分别表示第j次和第$j+1$次循环所对应的箱宽参数;$P_{\text{sum_}j}$和$P_{\text{sum_}j+1}$分别表示第j次和第$j+1$次循环下的样本点启发式概率密度之和;$n_i^{(j+1)}$和$n_i^{(j)}$分别表示样本点X_i在启发式窗口参数$\hat{h}_{j+1}$和$\hat{h}_j$条件下h-邻域内样本点的个数;D表示数据维度。由于$\left|\hat{h}_{j+1}-\hat{h}_j\right|\to 0$,因此可得$|P_{\text{sum_}j+1}-P_{\text{sum_}j}|\to 0$,即存在一个这样的启发式窗口参数,能够使得数据集中所有样本点对应的概率密度函数值趋于稳定。

对算法1作简要的时间复杂度分析,其中第1步找到合适的启发式窗口值$\hat{h}$一共循环了$\frac{\hat{h}}{\tau}$次,每次循环内计算启发式概率密度值之和需要遍历一次数据集,每个样本点的概率密度值计算的时间复杂度

是$O(1)$,结合起来可得出算法1的时间复杂度是$O\left(\frac{N\hat{h}}{\tau}\right)$。

3.2 目标函数

在确定启发式概率密度值后,HKDE将优化目标函数找到该数据集对应的KDE的最优窗口宽度参数h_{best},即HKDE模型最终确定用于概率密度估计的带宽,其与确定启发式概率密度值时所找的$\hat{h}$不同点在于:(1)h_{best}是一个窗口参数数组,每个维度都对应一个最优窗口参数,而$\hat{h}$不论维度是多少都只有一个数值;(2)$\hat{h}$是一个中间参数,用来确定启发式概率密度值,辅助寻找模型最终用于估计概率密度的窗口参数h_{best}。

在本文中,用启发式概率密度值$f_{\text{heu}}(X,\hat{h})$代替原误差度量涉及到的真实概率密度值,对于$D$维数据,基于MSE的误差计算公式可以表示为

$$\text{MSE}=\frac{1}{N}\sum_{i=1}^{N}\left[\hat{f}(X_i)-f_{\text{heu}}(X_i,\hat{h})\right]^2 \tag{10}$$

式中$\hat{f}(X_i)$表示当前模型估计得到的概率密度值,其计算公式为式(1)的多维扩展,包含窗口参数$[h_1,h_2,\cdots,h_D]$。

利用2.1节中生成的200个一维分布$f_{12}(\boldsymbol{x})$、500个二维分布$f_{22}(\boldsymbol{x})$的随机真实样本点,计算得到使目标函数MSE最小化的最优窗口参数。图2为式(10)对应的MSE和该分布的真实PDF值与KDE估计值的全局MSE(Global MSE,GMSE)随窗口参数变化的曲线。从图中可以看出,一、二维下目标函数式(10)所找到的最优窗口参数与使得GMSE最小的窗口参数相比较都偏小。

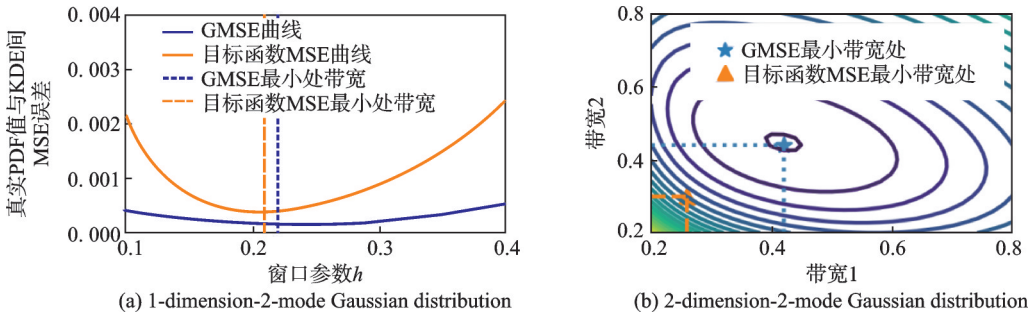

图2 目标函数MSE和GMSE随窗口参数变化示意图

Fig.2 Variations of MSE and GMSE with the change of window parameter

为了使HKDE得到的概率密度更接近真实值,可以取比使式(10)最小化的窗口参数偏大的值,即给目标函数添加一个补偿项。针对D维的模相近数据将误差目标函数调整为

$$\text{MSE}_{\text{HKDE}}=\text{MSE}+\sum_{d=1}^{D}\frac{\lambda_d}{h_d^2} \tag{11}$$

在确定调节系数$\lambda_1,\lambda_2,\cdots,\lambda_D$后,由于$\text{MSE}_{\text{HKDE}}$对应的最优$h_1,h_2,\cdots,h_D$的解析式无法确定,因此本文采用了粒子群优化算法(Particle swarm optimization, PSO)[31]确定最优窗口参数h_{best}。以一维的概率密度函数估计为例,结合式(10,11)可得

$$\text{MSE}_{\text{HKDE}}=\frac{1}{N}\sum_{i=1}^{N}\left[\hat{f}(X_i)-f_{\text{heu}}(X_i,\hat{h})\right]^2+\frac{\lambda}{h^2} \tag{12}$$

考虑到PSO算法的收敛特性,假设某前后两次PSO算法对应的最优窗口参数分别为h_{t+1}和h_t,满

足$|h_{t+1}-h_t|\to 0$。对应h_{t+1}和h_t的估计误差之间的差值可以表示为

$$\begin{aligned}&|\mathrm{MSE}_{\mathrm{HKDE}}(h_{t+1})-\mathrm{MSE}_{\mathrm{HKDE}}(h_t)|=\\&\left|\left[\frac{1}{N}\sum_{n=1}^{N}\left[\hat{f}(X_n,h_{t+1})-f_{\mathrm{heu}}(X_n,\hat{h})\right]^2+\frac{\lambda}{h_{t+1}^2}\right]-\left[\frac{1}{N}\sum_{n=1}^{N}\left[\hat{f}(X_n,h_t)-f_{\mathrm{heu}}(X_n,\hat{h})\right]^2+\frac{\lambda}{h_t^2}\right]\right|\leqslant\\&\left|\frac{1}{N}\sum_{n=1}^{N}\left[\hat{f}(X_n,h_{t+1})-f_{\mathrm{heu}}(X_n,\hat{h})\right]^2-\frac{1}{N}\sum_{n=1}^{N}\left[\hat{f}(X_n,h_t)-f_{\mathrm{heu}}(X_n,\hat{h})\right]^2\right|+\left|\frac{\lambda}{h_{t+1}^2}-\frac{\lambda}{h_t^2}\right|=\\&\left|\frac{1}{N}\sum_{n=1}^{N}\left[\left[\hat{f}(X_n,h_{t+1})-\hat{f}(X_n,h_t)\right]\left[\hat{f}(X_n,h_{t+1})+\hat{f}(X_n,h_t)-2f_{\mathrm{heu}}(X_n,\hat{h})\right]\right]\right|+\left|\frac{\lambda}{h_{t+1}^2}-\frac{\lambda}{h_t^2}\right|\end{aligned}\tag{13}$$

由于$|h_{t+1}-h_t|\to 0$,可得$\left|\hat{f}(X_n,h_{t+1})-\hat{f}(X_n,h_t)\right|\to 0$与$\left|\frac{\lambda}{h_{t+1}^2}-\frac{\lambda}{h_t^2}\right|\to 0$,故$|\mathrm{MSE}_{\mathrm{HKDE}}(h_{t+1})-\mathrm{MSE}_{\mathrm{HKDE}}(h_t)|\to 0$,这表明用于确定最优带宽的目标函数$\mathrm{MSE}_{\mathrm{HKDE}}$具备收敛性。

具体的算法流程如算法2所示。

算法2 优化目标函数找最优参数算法

输入:数据集DataSet$=\{X_1,X_2,\cdots,X_N\}$、数据维度D、算法1确定的启发式概率密度值$f_{\mathrm{heu}}(X,\hat{h})$

输出:最优窗口参数数组h_{best}

利用PSO算法求解最小化式(10)的窗口参数数组h_{MSE};

再利用数组h_{MSE}计算得到各数据样本点估计概率密度值与启发式概率密度值$f_{\mathrm{heu}}(X,\hat{h})$之间的MSE误差;

for $d=1$ to D do

计算使当前风险结构项$\frac{\lambda_d}{h_d^2}$与MSE误差保持在同一个量级的调节系数λ_d;

end for

目标函数加入确定了系数$\lambda_1,\lambda_2,\cdots,\lambda_D$的风险结构项,即式(11);

利用PSO算法优化新构建的目标函数,得到最终用于概率密度估计的窗口参数数组h_{best}。

对算法2中的关键步骤做如下解释说明。

(1) 在计算MSE的公式中,使用算法1计算所得的启发式概率密度值来替代原式中的真实概率密度值。在算法2中两次应用了PSO优化算法:第1次是寻找使未添加风险结构项的式(10)最小化的数组h_{MSE},以便确定每个维度下的调节系数λ;第2次是为了寻找使添加风险结构项后的最终目标函数式(11)最小化的h_{best},即HKDE最终用于概率密度估计的窗口参数数组。

(2) 算法1中确定的数据样本点启发式概率密度值记为$f_{\mathrm{heu}}(X,\hat{h})$,在算法2中确定参数数组$h_{\mathrm{MSE}}$后得到数据样本点对应Gaussian核的概率密度估计值记为$f_{\mathrm{MSE}}(X)$。为确保后续优化过程的有效性,令每个维度的风险结构项与MSE误差保持在同一量级,有

$$\frac{\lambda_d}{h_{\mathrm{MSE\_d}}{}^2}=\frac{1}{N}\sum_{i=1}^{N}\left[f_{\mathrm{MSE}}(X_i)-f_{\mathrm{heu}}(X_i,\hat{h})\right]^2\tag{14}$$

通过式(14)计算出第d维对应的系数$\lambda_d(d\in[1,D])$,从而确定本文所构建的目标函数,其中$h_{\mathrm{MSE_}d}$表示在参数数组h_{MSE}中对应的第d维窗口参数数值。

(3) PSO优化算法通过Python启发式算法库scikit-opt中的工具包sko.PSO实现。当目标函数的复杂度增加时,可以通过调整PSO算法的粒子数和迭代次数来进一步优化结果。

算法2主要利用PSO算法来进行优化,PSO算法的时间复杂度主要和迭代次数I、粒子个数P以及问题的维度D有关。对于一个D维的数据集,使用PSO进行优化的时间复杂度为$O(DPI)$。除了PSO外,还可以选择其他优化算法来最小化目标函数,如遗传算法、蚁群算法等,算法2的时间复杂度也会随着优化算法的改变而变化。

在经过算法1和算法2后,利用窗口参数数组h_{best}即可得到在Gaussian核函数下的最终估计PDF。

4 实验设置与结果

本节将针对HKDE模型的可行性、合理性和有效性进行验证。下面给出了9个1维和9个2维的多模高斯分布PDF信息,本文根据这些分布生成真实随机样本点数据集,用于后续估计PDF的实验中。本节所有实验均在配置为Intel(R) Core(TM) i5-7400 3.00 GHz CPU、16 GB内存、Windows 10专业版操作系统和Spyder 5.4.3编程环境的台式电脑上实现。符合D维K模高斯混合分布的仿真数据集生成依据简述如下。(1)均值$\boldsymbol{\mu}$。生成1个数值范围在[0,3)的$1\times D$维随机初始均值数组,后续$K-1$个均值都在前一个均值的基础上增加1个在范围[0,4)的随机偏移值,均保留小数点后两位;(2)方差$\boldsymbol{\Sigma}$。生成1个数值范围在[0,10)的$D\times D$维随机矩阵作为第1个模的方差,后续$K-1$个方差数值范围在[0,3),均保留小数点后两位;(3)权重$\boldsymbol{w}$。用Dirichlet分布随机生成K个权重,保证权重之和为1,均保留小数点后两位。

随机生成1维9种多模、2维9种多模,共18个符合高斯混合分布的仿真数据集用于后续的实验验证,实验所用数据集均能从公开的百度网盘下载获得:https://pan.baidu.com/s/1-lGL-ZXJHn0uJC9nNCG6UAg(提取码MY29)。

4.1 HKDE可行性验证

在本节实验中,为验证算法1中的收敛性(基于概率密度值对箱宽参数的收敛,来确定用于估算启发式概率密度值的窗口参数$\hat{h}$)和算法2中的收敛性(基于PSO算法的MSE和窗口参数对于迭代次数的收敛),分别生成服从$f_{14}(\boldsymbol{x})$分布的400个真实随机样本点,服从$f_{28}(\boldsymbol{x})$分布的500个真实随机样本点,其中$f_{DK}(\boldsymbol{x})$的下标分别表示数据的维度D和模数K。

(1) $f_{14}(\boldsymbol{x})(D=1,K=4)$:$\boldsymbol{w}=[0.30\ 0.60\ 0.02\ 0.08]$,$\boldsymbol{\mu}=[1\ 3\ 6\ 9]$,$\boldsymbol{\Sigma}=[[1.52][0.34][2.67][2.17]]$。

(2) $f_{28}(\boldsymbol{x})(D=2,K=8)$:$\boldsymbol{w}=[0.39\ 0.03\ 0.18\ 0.12\ 0.06\ 0.12\ 0.02\ 0.08]$;$\boldsymbol{\mu}=[[0.59\ 1.79][0.80\ 2.00][1.39\ 2.59][1.43\ 2.63][2.20\ 3.40][2.82\ 4.02][3.15\ 4.35][3.27\ 4.47]]$; $\boldsymbol{\Sigma}=[[[9.19\ 0][0\ 9.19]][[0.22\ 0][0\ 0.22]][[2.94\ 0][0\ 2.94]][[1.10\ 0][0\ 1.10]][[2.68\ 0][0\ 2.68]][[0.19\ 0][0\ 0.19]][[0.48\ 0][0\ 0.48]][[2.08\ 0][0\ 2.08]]]$。

图3给出了基于$f_{14}(\boldsymbol{x})$和$f_{28}(\boldsymbol{x})$两个分布的随机样本点启发式概率密度值之和对于窗口参数的收敛情况,即算法1的收敛性验证。在一维数据$f_{14}(\boldsymbol{x})$对应的图3(a)和二维数据$f_{28}(\boldsymbol{x})$对应的图3(b)中均可以观察到,随着窗口参数的增大,样本点的概率密度值之和呈现出收敛的趋势。对于同一分布生成的多组不同随机样本点数据集,可以观察到计算出的概率密度值之和的误差波动情况:在一维情况下波动较小,在二维情况下波动可以忽略不计,说明同一分布的收敛趋势基本相同。这些观察结果充分表明,本文所设计的算法1具有良好的收敛性。通过启发式概率密度值的收敛点,能够确定启发式窗口参数,为后续的概率密度函数估计提供可靠的依据。

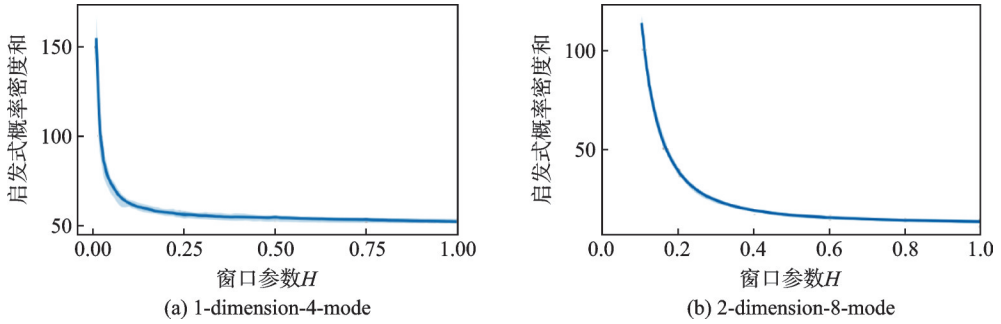

图3 算法1的收敛性示意图

Fig.3 Convergence of Algorithm 1

图4和图5给出了算法2中PSO算法涉及到收敛性的验证。图4清晰展示了式(11)的MSE误差随着迭代次数的增加而收敛的情况,可以发现一维数据在迭代20次左右、二维数据在迭代60次左右能达到收敛,表明PSO算法在优化过程中能够有效地减少MSE,并找到合适的解。图5则进一步验证了利用PSO算法最小化MSE误差寻找最优的窗口参数的收敛性,其中,图5(a)是基于$f_{14}(\boldsymbol{x})$的最终最优带宽收敛情况,图5(b)和图5(c)分别是基于$f_{28}(\boldsymbol{x})$的属性1和属性2的最终最优带宽收敛情况。从图中可以看出,无论窗口参数的初始化大小如何,都能在MSE误差达到收敛的迭代次数时稳定。实验结果表明,算法2能够确定最优窗口参数。通过上述的两个收敛性验证,证实了HKDE的可行性。

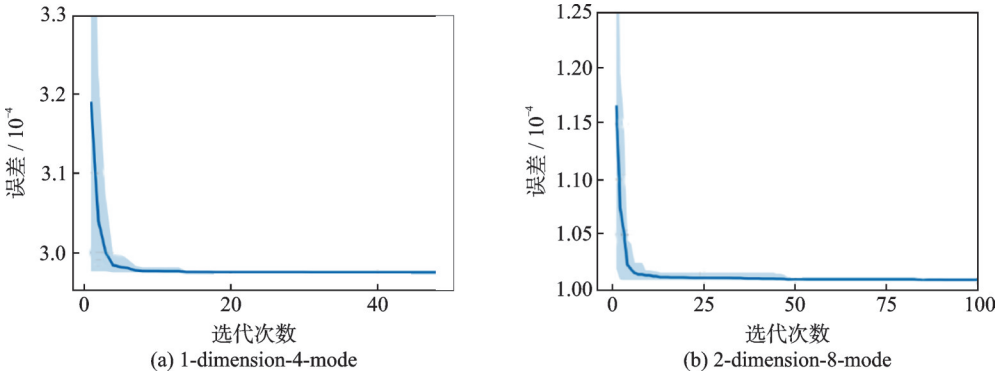

图4 MSE误差基于PSO算法迭代次数收敛示意图

Fig.4 Convergence of MSE error based on iterations of PSO algorithm

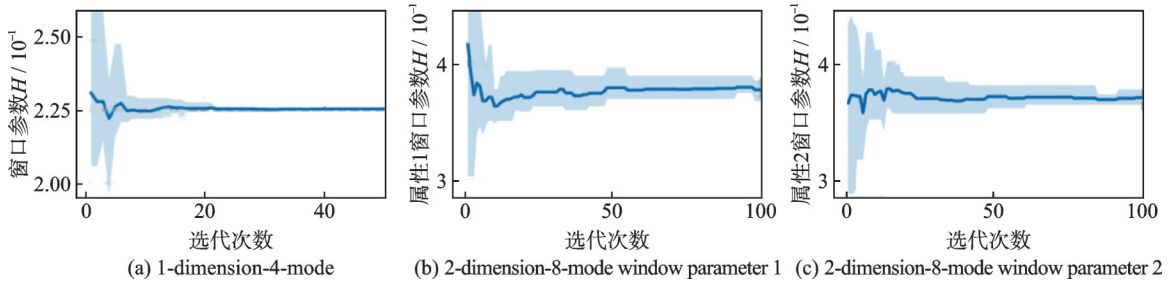

图5 窗口参数基于PSO算法迭代次数收敛示意图

Fig.5 Convergence of window parameter based on iterations of PSO algorithm

4.2 HKDE合理性验证

本节进行了实验证实HKDE的合理性，即HKDE能够降低KDE构建过程中的数据不确定性和模型不确定性，并能有效地估计随机样本点的PDF。

从部分的角度，实验基于图1给出的1维2模$f_{12}(\boldsymbol{x})$和2维2模$f_{22}(\boldsymbol{x})$，分别在数据样本点重叠较多的部分随机挑选10个样本点，利用算法1计算其启发式概率密度值，利用算法2计算其最终概率密度值，并与经典的LSCV[9]和RoT[11]进行比较，验证算法1解决数据不确定和算法2解决模型不确定的合理性。

(1) $f_{12}(\boldsymbol{x})(D=1,K=2)$

$$\begin{aligned}\text{Data}=[&-0.307\,322\,4,\ -0.202\,148\,1,\ -0.161\,549\,8,\ -0.094\,931\,5,\ -0.023\,058\,1,\\ &0.212\,640\,64,\ 0.228\,642\,47,\ 0.668\,534\,77,\ 0.745\,888\,69,\ 0.928\,233\,8\,]\end{aligned}$$

(2) $f_{22}(\boldsymbol{x})(D=2,K=2)$

$$\text{Data}=\begin{bmatrix}[3.949\,721\,5,5.136\,114\,9],[3.546\,337\,5,4.606\,866\,4],[4.888\,827\,7,4.406\,063\,8],\\ [3.294\,832\,9,5.137\,717\,5],[3.522\,410\,6,4.650\,521\,8],[4.000\,750\,7,4.202\,794\,6],\\ [3.276\,703\,7,4.607\,002\,7],[3.789\,589\,9,4.536\,032\,2],[3.751\,853\,6,5.072\,792\,6],\\ [3.862\,095\,3,4.265\,851\,4]\end{bmatrix}$$

在一、二维情况下，随机挑选10个重叠部分的样本点，通过与经典的LSCV[9]和RoT[11]进行数值上的比较。根据表2和表3可以发现，利用算法1计算得到的样本点启发式PDF值、利用算法2计算得到的HKDE最终估计PDF值，均比经典交叉验证方法LSCV和插件法RoT所得估计值与真实值之间的误差小。

从整体的角度，实验基于1维7模$f_{17}(\boldsymbol{x})$和2维9模$f_{29}(\boldsymbol{x})$分布分别生成的600个和500个真实随机样本点，利用HKDE对样本点分布的PDF进行估计。

(1) $f_{17}(\boldsymbol{x})(D=1,K=7)$

$\boldsymbol{w}=[\,0.06\ 0.21\ 0.08\ 0.46\ 0.01\ 0.06\ 0.12\,]$;$\boldsymbol{\mu}=[2\ 4\ 6\ 8\ 11\ 12\ 13]$;

$\boldsymbol{\Sigma}=[[1.79][1.53][2.74][0.84][0.39][2.14][0.12]]$。

表2 各方法基于1维2模高斯分布的PDF对比

Table 2 PDF comparison based on 1-dimension-2-mode Gaussian synthetic samples

数据点	真实PDF值	最终估计PDF值	启发式PDF值	LSCV	RoT
x_1	0.421 180	0.419 116	0.425 925	0.411 359	0.326 930
x_2	0.472 779	0.464 264	0.490 740	0.452 582	0.345 074
x_3	0.488 494	0.476 294	0.472 222	0.463 823	0.350 515
x_4	0.508 081	0.488 497	0.481 481	0.480 016	0.361 835
x_5	0.519 664	0.491 305	0.490 740	0.019 797	0.019 798
x_6	0.485 996	0.445 287	0.444 444	0.439 883	0.354 599
x_7	0.480 165	0.440 068	0.435 185	0.434 942	0.352 967
x_8	0.256 580	0.260 302	0.240 740	0.261 808	0.270 853
x_9	0.222 098	0.230 984	0.212 962	0.232 129	0.252 904
x_{10}	0.161 452	0.165 120	0.157 407	0.167 850	0.212 204

表3 各方法基于2维2模高斯分布的PDF对比

Table 3 PDF comparison based on 2-dimension-2-mode Gaussian synthetic samples

数据点	真实PDF值	最终估计PDF值	启发式PDF值	LSCV	RoT
x_1	0.059 880	0.051 744	0.051 020	0.050 635	0.034 518
x_2	0.085 662	0.064 589	0.063 775	0.062 599	0.039 264
x_3	0.033 067	0.028 495	0.038 265	0.027 466	0.028 002
x_4	0.056 320	0.050 201	0.051 020	0.048 888	0.033 627
x_5	0.084 217	0.063 995	0.063 775	0.062 010	0.038 986
x_6	0.078 604	0.059 397	0.082 908	0.057 756	0.038 838
x_7	0.076 596	0.058 454	0.076 530	0.056 439	0.037 537
x_8	0.087 492	0.064 745	0.089 285	0.062 867	0.039 684
x_9	0.066 530	0.057 027	0.076 530	0.055 595	0.035 874
x_{10}	0.084 334	0.063 196	0.063 775	0.061 465	0.039 645

(2) $f_{29}(\boldsymbol{x})(D=2,K=9)$

$\boldsymbol{w}$=[0.10 0.01 0.01 0.12 0.16 0.05 0.40 0.01 0.14];$\boldsymbol{\mu}$=[[1.60 0.69][4.63 3.72][7.95 7.04][10.30 9.39][12.81 11.90][16.71 15.80][18.72 17.81][20.89 19.98][23.41 22.50]];$\boldsymbol{\Sigma}$=[[[8.77 0][0 8.77]][[2.33 0][0 2.33]][[2.16 0][0 2.16]][[2.00 0][0 2.00]][[2.26 0][0 2.26]][[2.52 0][0 2.52]][[2.66 0][0 2.66]][[0.77 0][0 0.77]][[1.70 0][0 1.70]]]。

通过与LSCV[9]、BCV[10]、MLCV[14]、RoT[11]、Scott's RoT[12]和Silverman's RoT[13]这6种估计方法进行比较,即对比不同方法的估计PDF与真实PDF之间的一致性来验证HKDE的合理性,其中BCV使用R语言的kedd、ks包实现,其余在Python中使用Statsmodels库实现;PDF的一致性通过KL(Kullback-Leibler)散度来进行数值上的衡量。

为了直观展示不同方法间的性能差异,对HKDE以及上述6种方法进行了PDF等高线图的可视化对比,同时还计算了KL散度作为数值上的评估指标。图6和7分别是$f_{17}(\boldsymbol{x})$和$f_{29}(\boldsymbol{x})$分布的估计PDF对比示意图。从图中可以明显看出,RoT、Scott's RoT和Silverman's RoT这3种方法估计的PDF与真实PDF之间存在明显差距。这主要是因为它们都是基于正态分布假设来处理数据样本点,因此在面对多模分布时,相邻模对应的样本点无法被正确区分。这导致当模相近时,多模分布往往被估计得过于平滑,从而失去了真实分布的特征。相比之下,BVC、LSCV和MLCV方法表现出了一定的改进。然而,与这些方法相比,HKDE方法得到的PDF具有更小的KL散度值,即HKDE估计的PDF更接近于真实PDF,这表明HKDE在估计模相近数据样本点的PDF时具有更高的准确性和合理性。

4.3 HKDE有效性验证

本节在4.1节中提到的18个分布上,对HKDE和LSCV[9]、BCV[10]、MLCV[14]、RoT[11]、Scott's RoT[12]和Silverman's RoT[13]这6种方法以及高斯混合模型GMM[13]方法进行了数值比较,采用均方误差MSE作为衡量标准来评估HKDE的有效性。MSE表示估计PDF与真实PDF之间误差平方的平均值。通过比较不同方法的MSE值,可以判断估计PDF与真实PDF之间的接近程度,较小的MSE值意味着估计PDF更加接近真实PDF。

对于每个多模分布,生成一组服从该分布的随机样本点作为训练集,并生成一组均匀分布在训练数据集范围内的等距数据点作为测试集。在训练集上得到的训练均方误差(Training MSE, TMSE)反

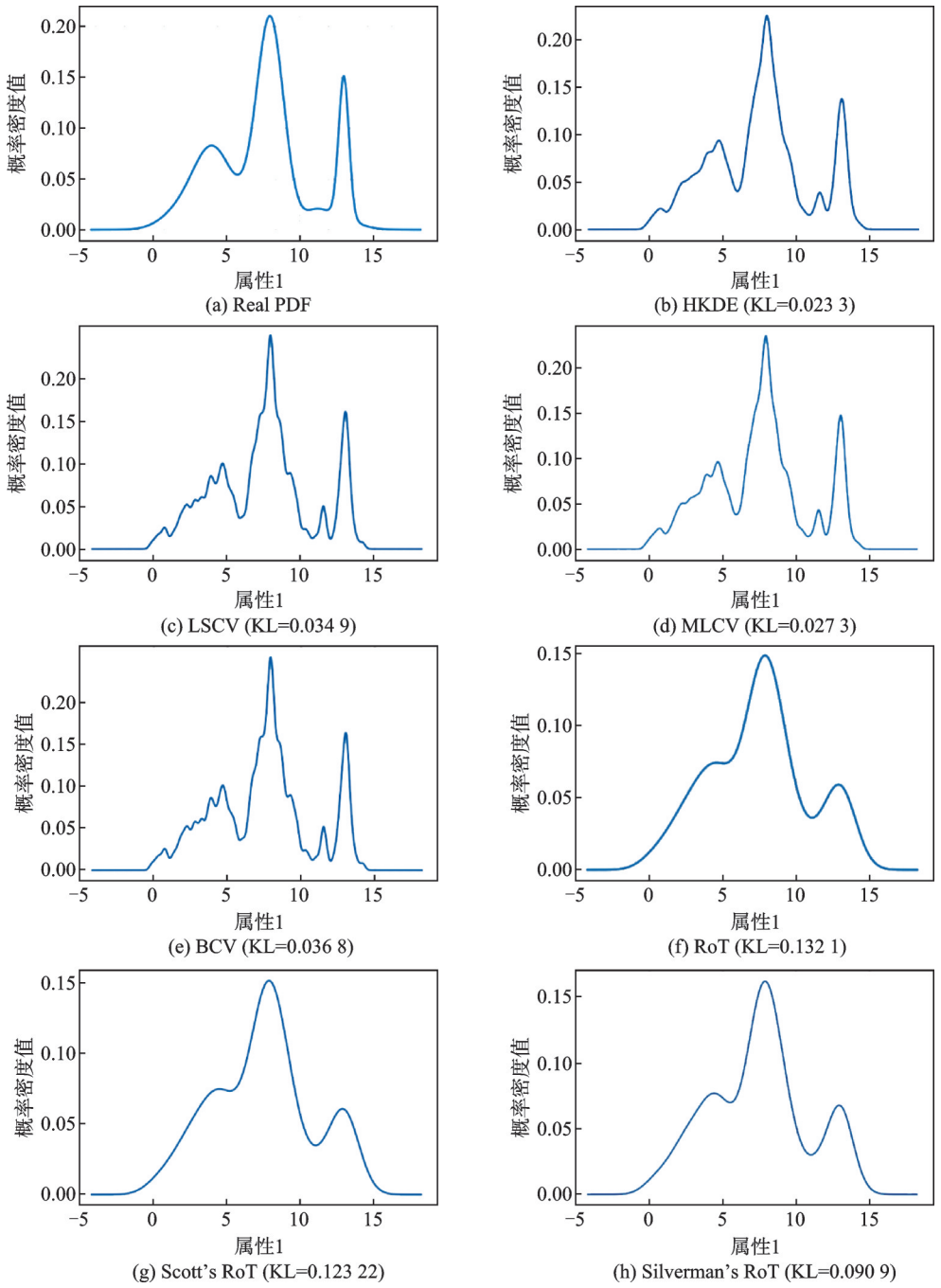

图6 7种KDE方法在基于高斯分布1维7模的PDF对比示意图

Fig.6 PDF estimations corresponding to seven different KDE methods on 1-dimension-7-mode Gaussian synthetic samples

映了模型在原始样本点上的拟合精度,即模型能否精确地逼近数据分布。而在测试集上得到的全局均方误差GMSE则衡量了模型在观测样本空间上的全局误差。通过TMSE和GMSE这两个指标可以客观全面地评估各种方法在多模分布下的表现。

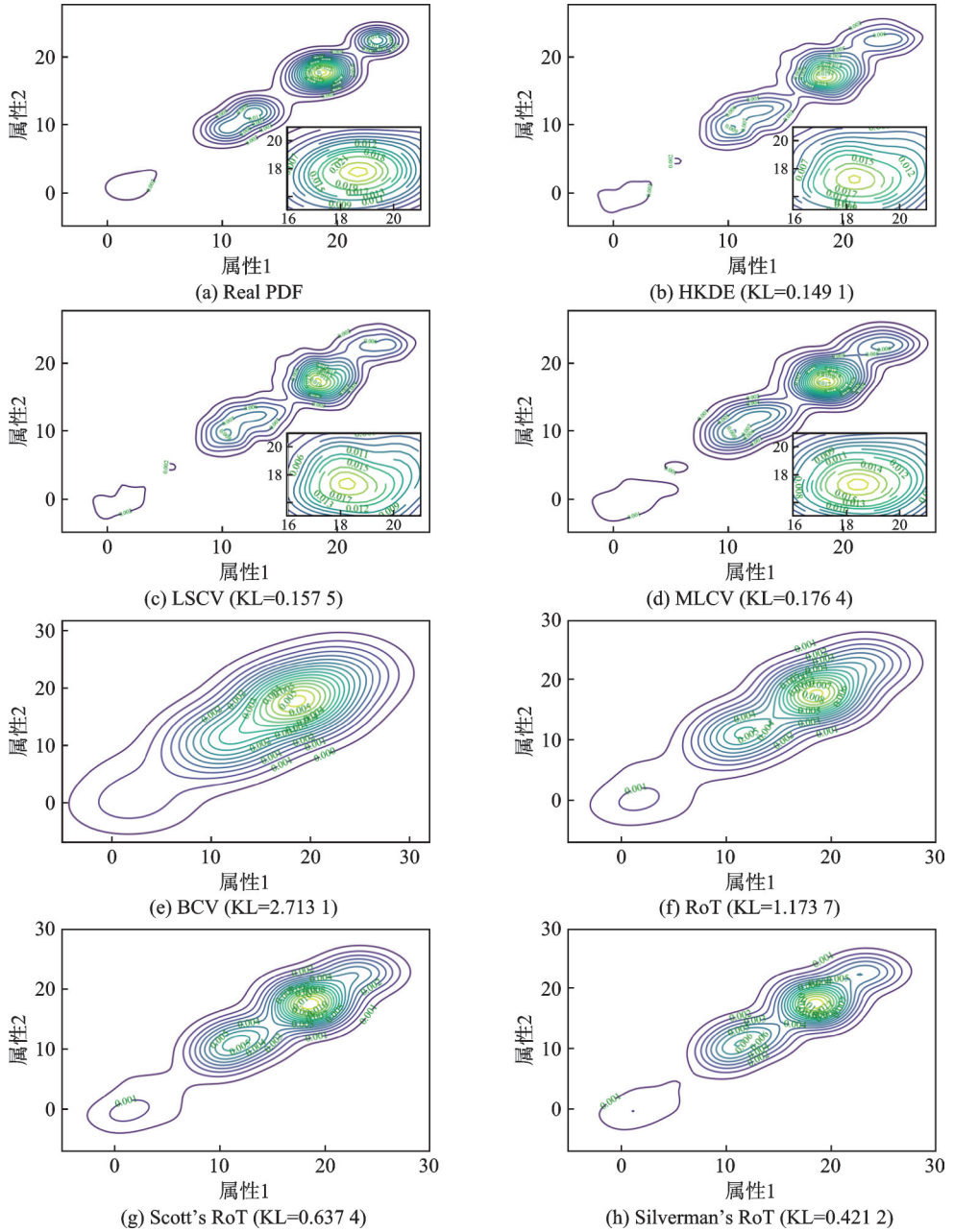

图7 7种KDE方法在基于高斯分布2维9模的PDF对比示意图

Fig.7 PDF estimations corresponding to seven different KDE methods on 2-dimension-9-mode Gaussian synthetic samples

在基于独立训练的18个高斯分布上,分别在训练数据集上使用TMSE和测试数据集上使用GMSE来测量平均估计误差,结果如表4~7所示。可以看出,与现有的6种代表性PDF核密度估计器相比,HKDE在训练数据集和测试数据集上的误差都较低;与GMM相比,在模数较少的情况下性能略显差异。其主要原因在于模相近数据集的GMM构件数在确定时较小,对于模数较少的数据集会更加接近真实的模数,当模数较大时所确定的构件数与真实的模数出现较大误差,导致GMM模型的估计准确度降低。

表4 各方法基于1维的9种多模高斯分布的TMSE对比

Table 4 TMSE comparison based on 1-dimension-multi-mode Gaussian synthetic samples

数据集	LSCV	MLCV	BCV	RoT	Scott's RoT	Silverman's RoT	GMM	HKDE
$f_{12}(\boldsymbol{x})$	0.000 571	0.001 392	0.000 623	0.007 238	0.005 346	0.003 315	**0.000 263**	0.000 401
$f_{13}(\boldsymbol{x})$	0.003 992	0.014 027	0.004 059	0.023 497	0.005 716	0.004 458	**0.002 388**	0.003 588
$f_{14}(\boldsymbol{x})$	0.001 231	0.001 884	0.001 051	0.006 826	0.001 954	0.001 457	**0.000 234**	0.000 905
$f_{15}(\boldsymbol{x})$	0.000 478	0.000 578	0.000 575	0.001 272	0.001 272	0.000 963	0.001 087	**0.000 462**
$f_{16}(\boldsymbol{x})$	0.001 540	0.004 290	0.026 931	0.019 797	0.019 798	0.017 434	**0.000 552**	0.000 833
$f_{17}(\boldsymbol{x})$	0.000 288	0.000 178	0.003 798	0.001 545	0.001 445	0.001 067	0.000 331	**0.000 137**
$f_{18}(\boldsymbol{x})$	0.000 429	0.000 502	0.004 979	0.002 915	0.002 915	0.002 519	0.001 387	**0.000 426**
$f_{19}(\boldsymbol{x})$	0.000 324	0.000 122	0.000 938	0.001 566	0.001 566	0.001 373	0.000 929	**0.000 115**
$f_{1,10}(\boldsymbol{x})$	0.000 373	0.000 367	0.000 421	0.000 380	0.000 380	0.000 329	0.000 466	**0.000 268**

表5 各方法基于2维的9种多模高斯分布的TMSE对比

Table 5 TMSE comparison based on 2-dimension-multi-mode Gaussian synthetic samples

数据集	LSCV	MLCV	BCV	RoT	Scott's RoT	Silverman's RoT	GMM	HKDE
$f_{22}(\boldsymbol{x})$	0.000 058	0.000 177	0.000 203	0.000 202	0.000 088	0.000 062	**0.000 011**	0.000 050
$f_{23}(\boldsymbol{x})$	0.000 084	0.000 441	0.000 187	0.000 379	0.000 223	0.000 141	**0.000 049**	0.000 058
$f_{24}(\boldsymbol{x})$	0.001 231	0.007 342	0.005 800	0.008 135	0.006 690	0.005 510	0.005 697	**0.001 008**
$f_{25}(\boldsymbol{x})$	0.003 121	0.019 997	0.009 527	0.021 064	0.006 428	0.004 642	**0.000 607**	0.002 709
$f_{26}(\boldsymbol{x})$	0.000 245	0.000 850	0.000 202	0.000 459	0.000 311	0.000 238	0.000 334	**0.000 186**
$f_{27}(\boldsymbol{x})$	0.028 973	0.006 766	0.006 624	0.007 333	0.006 274	0.005 568	0.004 892	**0.002 062**
$f_{28}(\boldsymbol{x})$	0.000 111	0.000 453	0.000 388	0.000 402	0.000 276	0.000 214	0.000 178	**0.000 096**
$f_{29}(\boldsymbol{x})$	4.64E−06	6.93E−06	5.24E−05	3.47E−05	2.35E−05	1.71E−05	6.92E−06	**4.61E−06**
$f_{2,10}(\boldsymbol{x})$	0.000 023	0.000 035	0.000 278	0.000 220	0.000 154	0.000 125	0.000 027	**0.000 020**

表6 各方法基于1维的9种多模高斯分布的GMSE对比

Table 6 GMSE comparison based on 1-dimension-multi-mode Gaussian synthetic samples

数据集	LSCV	MLCV	BCV	RoT	Scott's RoT	Silverman's RoT	GMM	HKDE
$f_{12}(\boldsymbol{x})$	0.000 148	0.000 264	0.000 154	0.001 216	0.000 905	0.000 572	**0.000 053**	0.000 133
$f_{13}(\boldsymbol{x})$	0.000 645	0.001 677	0.000 651	0.002 646	0.000 811	0.000 688	**0.000 429**	0.000 610
$f_{14}(\boldsymbol{x})$	0.000 200	0.000 294	0.000 176	0.001 026	0.000 304	0.000 232	**0.000 045**	0.000 159
$f_{15}(\boldsymbol{x})$	0.000 128	0.000 150	0.000 150	0.000 322	0.000 322	0.000 245	0.000 225	**0.000 125**
$f_{16}(\boldsymbol{x})$	0.000 246	0.000 246	0.003 555	0.002 479	0.002 679	0.002 365	0.000 252	**0.000 188**
$f_{17}(\boldsymbol{x})$	0.000 105	0.000 076	0.001 193	0.000 496	0.000 465	0.000 345	0.000 153	**0.000 064**
$f_{18}(\boldsymbol{x})$	0.000 125	0.000 136	0.001 566	0.000 874	0.000 874	0.000 737	0.000 364	**0.000 125**
$f_{19}(\boldsymbol{x})$	0.000 124	0.000 047	0.000 323	0.000 516	0.000 516	0.000 455	0.000 368	**0.000 046**
$f_{1,10}(\boldsymbol{x})$	0.000 140	0.000 138	0.000 157	0.000 153	0.001 426	0.000 126	0.000 199	**0.000 112**

表7 各方法基于2维的9种多模高斯分布的GMSE对比

Table 7 GMSE comparison based on 2-dimension-multi-mode Gaussian synthetic samples

数据集	LSCV	MLCV	BCV	RoT	Scott's RoT	Silverman's RoT	GMM	HKDE
$f_{22}(\boldsymbol{x})$	2.78E−06	6.13E−06	6.97E−06	6.92E−06	3.43E−06	2.78E−06	**6.18E−07**	2.59E−06
$f_{23}(\boldsymbol{x})$	3.95E−06	12.1E−06	5.50E−06	12.1E−06	6.33E−06	4.48E−06	**2.10E−06**	3.30E−06
$f_{24}(\boldsymbol{x})$	2.12E−05	9.03E−05	6.98E−05	9.03E−05	8.07E−05	6.63E−05	7.56E−05	**1.94E−05**
$f_{25}(\boldsymbol{x})$	2.93E−05	13.9E−05	6.84E−05	14.4E−05	4.74E−05	3.69E−05	**6.10E−06**	2.80E−05
$f_{26}(\boldsymbol{x})$	6.16E−06	17.7E−06	5.63E−06	9.57E−06	6.93E−06	5.89E−06	8.89E−06	**5.60E−06**
$f_{27}(\boldsymbol{x})$	13.4E−05	5.84E−05	5.72E−05	6.32E−05	5.42E−05	4.84E−05	4.69E−05	**2.18E−05**
$f_{28}(\boldsymbol{x})$	2.69E−06	9.09E−06	7.62E−06	7.92E−06	5.31E−06	4.19E−06	4.71E−06	**2.52E−06**
$f_{29}(\boldsymbol{x})$	0.35E−06	0.43E−06	2.71E−06	1.81E−06	1.24E−06	0.93E−06	0.75E−06	**0.34E−06**
$f_{2,10}(\boldsymbol{x})$	0.66E−06	0.92E−06	6.51E−06	5.14E−06	3.62E−06	2.94E−06	0.73E−06	**0.62E−06**

结合上述的数值对比结果和图像对比结果,可以得出结论:与其他传统的KDE估计器相比,HKDE具有更好的性能表现,是一种能够处理模相近数据PDF估计问题的有效方法。

5 结束语

本文设计了一种针对模相近数据的启发式核密度估计器HKDE,用于提升对模相近数据的概率密度函数拟合的准确性。传统核密度估计器在处理模相近数据时,常常面临数据不确定性和模型不确定性的挑战,这些问题影响了KDE的估计性能。HKDE采用了创新的方法来克服不确定性带来的KDE估计缺陷。在降低数据不确定方面,HKDE通过利用观测数据概率密度值对于直方图箱宽参数的收敛性来确定启发式概率密度函数值,从而替代误差度量目标函数中的真实PDF值;在降低模型不确定性上,利用新构建的、添加了补偿项的目标函数来确定HKDE的最优窗口参数。通过一系列真实可信的实验证实了HKDE的可行性、合理性和有效性,表明HKDE是一种估计模相近数据概率密度的有效方法。未来的工作计划从3个方面进行深入研究:(1)针对大规模数据集,考虑将分布式随机样本划分技术融入HKDE中,以提高其在处理大规模数据时的效率和性能;(2)通过具体的应用场景,进一步深入验证HKDE的性能,探索其在不同领域中的潜在应用价值;(3)针对更高维度和更复杂的真实数据,将对HKDE目标函数进行改进和拓展,以提升其适应性和泛化能力。

参考文献:

[1] WANG P, DENG H, WANG Y M, et al. Kernel density estimation based Gaussian and non-Gaussian random vibration data induction for high-speed train equipment[J]. IEEE Access, 2020, 8: 90914-90923.

[2] 胡李军, 薛海, 周宇. 基于核密度估计的重载组合列车纵向载荷谱外推研究[J]. 兰州交通大学学报, 2022, 41(2): 94-100.
HU Lijun, XUE Hai, ZHOU Yu. Extrapolation research on longitudinal load spectrum of heavy haul combined train based on kernel density estimation[J]. Journal of Lanzhou Jiaotong University, 2022, 41(2): 94-100.

[3] WATADA J. A kernel density estimation-maximum likelihood approach to risk analysis of portfolio[C]//Proceedings of 2013 IEEE 8th International Symposium on Intelligent Signal Processing. [S.l.]: IEEE, 2013: 37-42.

[4] WAHIDUZZAMAN M, YEASMIN A. A kernel density estimation approach of north indian ocean tropical cyclone formation and the association with convective available potential energy and equivalent potential temperature[J]. Meteorology and Atmospheric Physics, 2020, 132(5): 603-612.

[5] 萧凌波. 基于核密度估计的清代中国自然灾害时空分布特征 [J]. 灾害学, 2019, 34(4): 92-99.
XIAO Lingbo. Spatio-temporal distribution of natural disasters in China during 1644-1911 based on kernel density estimation[J]. Journal of Catastrophology, 2019, 34(4): 92-99.

[6] 施剑玮,奚蔚.限带白噪声随机过程的雨流幅值概率密度函数模型[J].南京航空航天大学学报,2020,52(4): 659-665.
SHI Jianwei, XI Wei. Probability density function model of rain flow amplitude for random process of band-limited white noise[J]. Journal of Nanjing University of Aeronautics &Astronautics, 2020, 52(4): 659-665.

[7] HOROVÁ I, KOLACEK J, ZELINKA J. Kernel smoothing in MATLAB: Theory and practice of kernel smoothing[M]. Singapore: World Scientific, 2012.

[8] HEIDENREICH N-B, SCHINDLER A, SPERLICH S. Bandwidth selection for kernel density estimation: A review of fully automatic selectors[J]. ASTA Advances in Statistical Analysis, 2013, 97: 403-433.

[9] BOWMAN A W. An alternative method of cross-validation for the smoothing of density estimates[J]. Biometrika, 1984, 71(2): 353-360.

[10] SCOTT D W, TERRELL G R. Biased and unbiased cross-validation in density estimation[J]. Journal of the American Statistical Association, 1987, 82(400): 1131-1146.

[11] SILVERMAN B W. Kernel density estimation technique for statistics and data analysis[J]. Monographs on Statistics and Applied Probability, 1986, 26:34-74.

[12] SCOTT D W. Multivariate density estimation: Theory, practice, and visualization[M]. Hoboken, NJ: John Wiley & Sons, 2015.

[13] SILVERMAN B W. Density estimation for statistics and data analysis[M]. London: Routledge, 2018.

[14] HORNE J S, GARTON E O. Likelihood cross-validation versus least squares cross-validation for choosing the smoothing parameter in kernel home-range analysis [J]. The Journal of Wildlife Management, 2006, 70(3): 641-648.

[15] WĘGLARCZYK S. Kernel density estimation and its application[C]//Proceedings of ITM Web of Conferences. Lesvlis, France: EDP Sciences, 2018, 23: 00037.

[16] HEER J. Fast & accurate Gaussian kernel density estimation[C]//Proceedings of 2021 IEEE Visualization Conference (VIS). [S.l.]: IEEE, 2021: 11-15.

[17] HABBEMA J, HERMANS J, VAN DEN BROEK K. A stepwise discriminant analysis program using density estimation [C]//Proceeding in Computational Statistics. Vienna: Rudolf Liebling, 1974: 101-110.

[18] CHIU S T. Some stabilized bandwidth selectors for nonparametric regression[J]. The Annals of Statistics, 1991, 19(3): 1528-1546.

[19] STUTE W. Modified cross-validation in density estimation[J]. Journal of Statistical Planning and Inference, 1992, 30(3): 293-305.

[20] HART J D, YI S. One-sided cross-validation [J]. Journal of the American Statistical Association, 1998, 93(442): 620-631.

[21] PARK B U, MARRON J S. Comparison of data-driven bandwidth selectors[J]. Journal of the American Statistical Association, 1990, 85(409): 66-72.

[22] HALL P, SHEATHER S J, JONES M, et al. On optimal data-based bandwidth selection in kernel density estimation[J]. Biometrika, 1991, 78(2): 263-269.

[23] TAYLOR C C. Bootstrap choice of the smoothing parameter in kernel density estimation[J]. Biometrika, 1989, 76(4): 705-712.

[24] SHEATHER S J, JONES M C. A reliable data-based bandwidth selection method for kernel density estimation[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1991, 53(3): 683-690.

[25] KIM W, PARK B, MARRON J S. Asymptotically best bandwidth selectors in kernel density estimation[J]. Statistics & Probability Letters, 1994, 19(2): 119-127.

[26] RAJAN A, KUANG Y C, OOI M P-L , et al. Moment-constrained maximum entropy method for expanded uncertainty evaluation[J]. IEEE Access, 2018, 6: 4072-4082.

[27] ZHANG Z, JIANG C, HAN X, et al. A high-precision probabilistic uncertainty propagation method for problems involving multimodal distributions[J]. Mechanical Systems and Signal Processing, 2019, 126:21-41.

[28] LI G, WANG Y, ZENG Y, et al. A new maximum entropy method for estimation of multimodal probability density function[J]. Applied Mathematical Modelling, 2022, 102: 137-152.

[29] PARZEN E. On estimation of a probability density function and mode[J]. The Annals of Mathematical Statistics, 1962, 33(3): 1065-1076.

[30] WEGMAN E J. Nonparametric probability density estimation: A summary of available methods[J]. Technometrics, 1972, 14(3): 533-546.

[31] JAIN M, SAIHJPAL V, SINGH N, et al. An overview of variants and advancements of PSO algorithm[J]. Applied Sciences, 2022, 12(17): 8392.

作者简介:

何玉林(1982-),**通信作者**,男,博士,研究员,研究方向:数据挖掘、机器学习、大数据系统计算技术,E-mail:yulinhe@gml.ac.cn。

陈纯佳(2000-),女,硕士研究生,研究方向:数据挖掘、机器学习,E-mail:1571650576@qq.com。

黄哲学(1959-),男,博士,特聘教授,研究方向:大数据系统计算技术,E-mail:zx.huang@szu.edu.cn。

李俊杰(1980-),男,博士,副教授,研究方向:大数据系统计算技术、数据挖掘、机器学习,E-mail:jj.li@szu.edu.cn。

FOURNIER-VIGER Philippe(1980-),男,博士,特聘教授,研究方向:数据挖掘、人工智能、知识表示和推理、认知模型建构等,E-mail:philfv@szu.edu.cn。

(编辑:刘彦东)