基于估计点的双窗宽核密度估计算法

邓 飙[1,2]，于传强[2]，李天石[1]，苏文斌[1]，盘仰珂[3]
（1 西安交通大学机械工程学院 西安 710049；2 第二炮兵工程学院 西安 710025；
3 后勤指挥学院 北京 100037）

摘 要：针对核密度估计中窗宽确定困难的问题，提出了基于估计点的滑动双窗宽核密度估计算法。通过采用固定窗宽的密度估计函数代替假设的正态分布密度函数以及增加求解二次导数的窗宽的方法，对估计域中的每一估计点求取其最优窗宽值，实现了窗宽根据样本的分布情况，在不同的估计点自动调整窗宽的取值。文中给出了算法的具体推导以及实现步骤，给出了多组对比实验。结果表明双窗宽算法在估计结果的精度、平滑度等方面比现有固定窗宽的算法有明显提高。

关键词：核密度估计；双窗宽；估计点；算法
中图分类号：TP2 **文献标识码**：A **国家标准学科分类代码**：520.604

Dual-bandwidth kernel density estimation algorithm based on estimate points

Deng Biao[1,2], Yu Chuanqiang[2], Li Tianshi[1], Su Wenbin[1], Pan Yangke[3]
(*1 Xi'an Jiaotong University, Xi'an 710049, China; 2 Second Artillery Engineering Institute, Xi'an 710025, China; 3. Logistic Command Institute, Beijing 100037, China*)

Abstract: Aiming at the problem that the bandwidth of kernel density estimation is difficult to choose, this paper proposes a slide double bandwidth kernel density estimation algorithm based on estimate points. We adopt the kernel density estimation function obtained with fixed bandwidth to replace the hypothetic normal distribution density function and increase the bandwidth for solving the second derivative to obtain the optimal bandwidth at every point in the estimate interval. The method can automatically adjust the value of the bandwidth at different estimate points according to the sample distribution. In the paper, the detailed deduction process, realization steps are given. Groups of comparison experiments were carried out. Results show that the new algorithm improves the precision and gliding property of kernel density estimation obviously compared with fixed bandwidth algorithm.

Key words: kernel density estimation; dual-bandwidth; estimate point; algorithm

1 引 言

概率密度估计是一个应用范围十分广泛的问题，有了概率密度，几乎可以做所有与统计特性有关的计算和分析。密度估计通常有两种方法：参数估计和非参数估计。前者是密度函数结构已知，只有其中某些参数未知，此时的密度估计就是传统的参数估计问题；后者是密度函数未知，仅从即有样本出发得出密度函数的表达式[1]。

非参数估计始于直方图法，后来发展为最近邻法、Rosenblatt法以及核密度估计法（Parzen窗核密度估计法）。前几种方法都存在对区间某部分估计精度较差等问题，不适合整体估计，目前只是在一些简单和要求不高的场合有所应用；核密度估计法在理论上是比较完善的方法，它克服了前几种方法的缺点，能够在整个估计区间都获得较好的精度和平滑度，是当前非参数概率密度估计领域中研究和发展的主要方向[2]。

2 核密度估计及其窗宽选择

Parzen窗方法核密度估计函数可写成如下形式：

收稿日期：2010-08 Received Date: 2010-08

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - x_1}{h}\right) \quad (1)$$

式中: $\hat{f}(x)$ 表示估计的概率密度,n 表示样本数量,h 为窗宽,d 表示空间的维数,$K(x)$ 为核函数。

窗宽 h 对于核密度估计是一个至关重要的参数,在给定样本时,不同核函数对核密度估计精度的影响很小,而窗宽 h 对计算结果起着决定性的影响。如果 h 太小,那么结果就会不稳定; 反之,如果 h 太大,则会导致结果的分辨率太低。一般来说,h 会随着 n 的增大而减小。这样在有限样本个数的约束下,需要寻找合适的 h,希望达到稳定性与分辨率之间的折中。

目前,有关窗宽选择的方法可分为以下几种: 交错鉴定方法、惩罚函数法、插入法(The plug-in method) 以及对比方法[1-4]。常用的是插入法,即把未知函数的估计插入到渐近公式里以选择最佳窗宽,Sain(1994) ,Jone(1996) 提出了改进的插入法,改进的插入法有较好的理论分析性质和实际应用效果,并且被认为要优于交错鉴定方法和惩罚函数法。插入法的研究中又可分为以下两个方向:

1) 固定窗宽算法(fixed bandwidth algorithm,FBA) ,这是目前最为常见和有效的方法[5-7],大多密度估计的问题都是基于此而展开。

2) 变窗宽算法(variable bandwidth algorithm,VBA)[8-9],窗宽的取值与样本的稀疏程度有关,而固定窗宽无法适应这些变化,于是出现了变窗宽的方法。从核密度估计的基本原理可以推断,变窗宽能够更好地反映估计区间不同点的光滑程度,降低拟合曲线在峰顶区域的偏差以及尾部区域的方差,提高拟合曲线的灵活性。

3 固定窗宽算法(FBA) 的基本思想

现有关于窗宽的优化算法,主要是基于积分均方误差 MISE(mean integral square error) 的固定窗宽优化[3-4]。MISE 定义如下:

$$MISE(\hat{f}(x)) = E\int[\hat{f}(x) - f(x)]^2 dx = \int[E\hat{f}(x) - f(x)]^2 dx + \int Var\hat{f}(x)\,dx \quad (2)$$

式(2) 分解后的第一项表示 $\hat{f}$ 的期望值与真实值之间偏差平方的积分,简称偏差,将 $E\hat{f}(x) - f(x)$ 记为 $bias(\hat{f}(x))$; 第二项表示估计值的方差积分,简称方差。对上式进行求解,得到如下结果:

$$bias(\hat{f}(x))^2 \approx \frac{1}{4}h^4 k_2 f''(x)^2 \quad (3)$$

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

其中 $k_2 = \int t^2 K(t)\,dt$,令

$$AMISE(\hat{f}(x)) = \frac{1}{4}h^4 k_2^2 \int f''(x)^2 dx + \frac{1}{nh}\int K(t)^2 dt \quad (5)$$

所以要求 $MISE(\hat{f}(x))$ 的最小值,近似地只需求解 $AMISE(\hat{f}(x))$ 的最小值即可,对式(7) 求导,并令一阶导数为零得到最优窗宽:

$$h = \left(\frac{\int K(t)^2 dt}{k_2^2 \int f''(x)^2 dx}\right)^{1/5} n^{-1/5} \quad (6)$$

最优固定窗宽的表达式依赖于 $f(x)$,但其未知,通常对其作服从 $N(u,\sigma^2)$ 的假设,经过计算,便可求得 FBA 算法的窗宽表达式:

$$h = 1.059\sigma n^{-1/5} \quad (7)$$

4 基于估计点的双窗宽算法(dual-bandwidth algorithm based on estimation point, DBABEP)

4.1 基本思想

从 FBA 算法的论述中,可以得到启发: 如果对定义域中的每一点都能够计算出其最优值,那么,自然整个定义域内的估计结果也是最优的。因此,利用前面得到的结果,设 $MSE = bias(\hat{f}(x))^2 + Var\hat{f}(x)$,对其关于 h 求导得到:

$$MSE' \approx h^3 f''(x)^2 k_2^2 - \frac{1}{nh^2}\int K(t)^2 dt \quad (8)$$

令其为零,求得窗宽 $h(x)_*$ 的表达式:

$$h(x)_* = \left(\frac{f(x)\int K(t)^2 dt}{k_2^2 f''(x)^2}\right)^{1/5} n^{-1/5} \quad (9)$$

从式(9) 可以看出,的值与待求密度估计点 x 有关,是基于估计点的优化,其值随着估计点的不同而不同,利用上式得到的结果求解 x 分布密度的时侯,窗宽的取值随着 x 的变化而变化,而 FBA 算法中,最优窗宽值与 x 的具体取值无关[10-11]。

由于密度函数 $f(x)$ 未知,所以 $h(x)_*$ 无法计算。比较式(9) 和式(6) 可以发现: 式(6) 是基于整个定义域的优化,h_* 的取值取决于 $\int f''(x)^2 dx$,其对未知分布准确性的要求不是很严格,所以,采用服从 $N(u,\sigma^2)$ 的假设分布就能够取得较好效果; 式(9) 是基于点的优化,$h(x)_*$ 在每一点的取值都受到 $f(x)$ 和 $f''(x)$ 的双重影响,其对未知分布准确性的要求更加严格,采用服从 $N(u,\sigma^2)$ 的假设无法取得满意效果[12]。

一个自然的想法是,采用 FBA 算法求得的密度估计函数代替真实的密度函数,而不是采用一个假设的密度函数来代替。因为对于一个未知分布,通过核密度估计获得的密度函数,多数情况下,应该比一个假设的分布更加接近实际[13-14]。

另外一点,也是更为重要的一点,$f''(x)$ 表示密度估计函数各点的凹凸程度,如果直接对采用 FBA 算法求得的 $\hat{f}(x)$ 二次求导来代替 $f''(x)$,会有较大偏差。因为,一般情况下,按照 FBA 算法求得 $\hat{f}(x)$ 密度图形光滑性不是很好,曲线局部起伏程度较大,虽然 $\hat{f}(x)$ 与 $f(x)$ 偏差可能不大,可用于代替 $f(x)$,但是对其二次求导来代替 $f''(x)$,失真严重。所以如果能够对 $\hat{f}(x)$ 进行适当的处理,使其更加平滑,有利于减小 $\hat{f}''(x)$ 与 $f''(x)$ 的偏差[15]。

我们知道,核密度估计图形的光滑程度主要取决于窗宽: 窗宽越大,密度图形的光滑程度越好。因此,为了减小 $\hat{f}''(x)$ 与 $f''(x)$ 之间的偏差,可以适当的加大固定窗宽值,从而取得较好效果。

4.2　算法推导

根据上述思想,我们采用 FBA 算法求得的 $\hat{f}(x)$ 代替 $f(x)$,采用加大窗宽值求得的 $\hat{f}''_H(x)$ 代替 $f''(x)$,对式(9)求解。由于求解过程中 $\hat{f}(x)$ 和 $\hat{f}''_H(x)$ 采用了不同的窗宽值,因此,称之为基于估计点的双窗宽算法,简称双窗宽算法(dual-bandwidth algorithm, DBA)。下面以高斯核函数 $K(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$(其他核函数也可通过相同方法解决)为例对 DBA 算法进行推导,有如下结果:

$$\hat{f}_H(x) = \frac{1}{nH}\sum_{i=1}^{n} K\left(\frac{x - x_i}{H}\right) \tag{10}$$

$$\hat{f}''_H(x) = \frac{1}{nH^3}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\left(t^2 e^{-t^2/2} - e^{-t^2/2}\right) \tag{11}$$

式中: H 为调整后的窗宽值, $H = ah$,通常 $a > 1$,称为窗宽调节系数, h 为 FBA 算法窗宽的优化值; $t = \frac{x - x_i}{H}$。

据式(9),得到 DBA 算法的窗宽值:

$$h(x)_* = c^{0.2} n^{-0.2} H^{0.8} \frac{\hat{f}(x)^{0.2}}{\left(\sum_{i=1}^{n} \frac{1}{nH} \frac{1}{(1 - e^{-1})\sqrt{2\pi}} t^2 e^{-t^2/2} - \hat{f}_H(x)\right)^{0.4}} \tag{12}$$

式中: $c = \frac{\int K(t)^2 \mathrm{d}t}{k_2}$。

则,基于 DBA 算法的核密度估计函数为:

$$\hat{f}(x)_* = \frac{1}{nh(x)_*}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(x)_*}\right) \tag{13}$$

因此,有如下的基于 DBA 算法的核密度估计算法:

1) 根据式(7),计算基于 FBA 算法的窗宽 h;

2) 根据式(1)求 $\hat{f}(x)$;

3) 计算 $H = ah, a > 1$;

4) 根据式(10),求 $\hat{f}_H(x)$;

5) 根据式(11),求 $\hat{f}''_H(x)$;

6) 根据式(12),在不同的估计点计算;

7) 基于 DBA 算法的核密度估计函数为:

$$\hat{f}(x)_* = \frac{1}{nh(x)_*}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h(x)_*}\right)。$$

5　算法验证

5.1　验证实例

用 3 个正态分布密度叠加产生 600 个数据样本,其中正态分布 1: 均值为 4,标准差为 1.5,样本数量 200; 正态分布 2: 均值为 0,标准差为 3,样本数量 200; 正态分布 3: 均值为 5,标准差为 2,样本数量 200。分布区域取 $[-8, 10]$,估计点个数取 900,固定窗宽为 $h = 0.7683$。

分别采用基于 FBA 算法和 DBA 算法对其进行密度估计,密度估计的图形、各估计点偏差绝对值的平均(对于 FBA 算法为: $\sum_{i=1}^{n} |f(x_i) - \hat{f}(x_i)|/n$, n 为估计点个数; 对于 DBA 算法为: $\sum_{i=1}^{n} |f(x_i) - \hat{f}(x_i)_*|/n$。以下分别用 $EA(\hat{f}(x))$ 和 $EA(\hat{f}(x)_*)$ 表示。),以及窗宽的变化分别如图 1 ~ 3 所示。

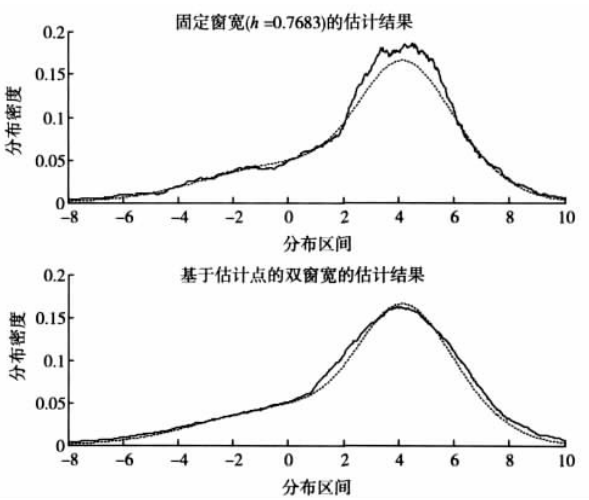

图 1　两种算法的密度估计结果

Fig. 1 Density estimation results for two algorithms

从图 1(虚线代表真实的密度值) 中可以看出,采用 DBA 算法的密度估计结果(下方的图) 比 FBA 算法的估计结果(上方的图) 要好,其图形与真实的密度分布更加接近,并且光滑度更好,DBA 算法中, $H = 9h$ 。

图 2 反映的是在各个估计点两种估计方法的偏差绝对值的大小,可以看出,DBA 算法密度估计(下方的图) 的精度高一些; 在本例中, $EA(\hat{f}(x)) = 0.0056$, $EA(\hat{f}(x)_*) = 0.0043$,精度提高 23.05%。

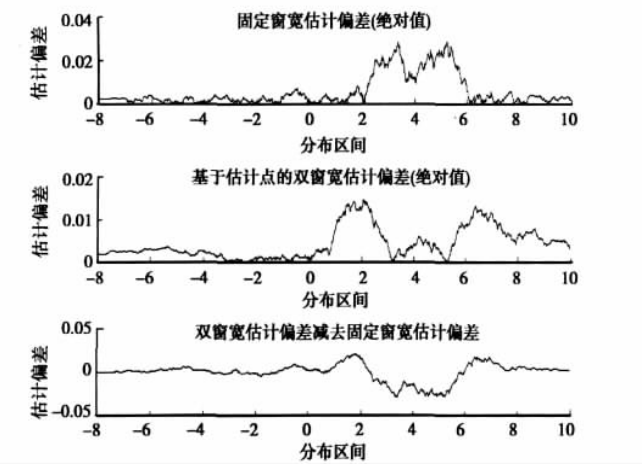

图 2 两种算法密度估计偏差绝对值的平均

Fig. 2 The means of absolute value of density estimation deviation for two algorithms

图 3 表示的是两种算法方法的窗宽变化情况,图上方的曲线是 DBA 算法的窗宽,下方的直线为 DBA 算法的窗宽。从该图中可以看出: DBA 算法的窗宽值比 FBA 算法的窗宽值要大,因此,采用 DBA 算法的分布密度图形更加平滑,但是,其精度反而没有降低,这一点很重要。实际应用中,人们最希望得到的结果就是: 既有足够的精度,又要有很好的稳定性。DBA算法的密度估计结果比

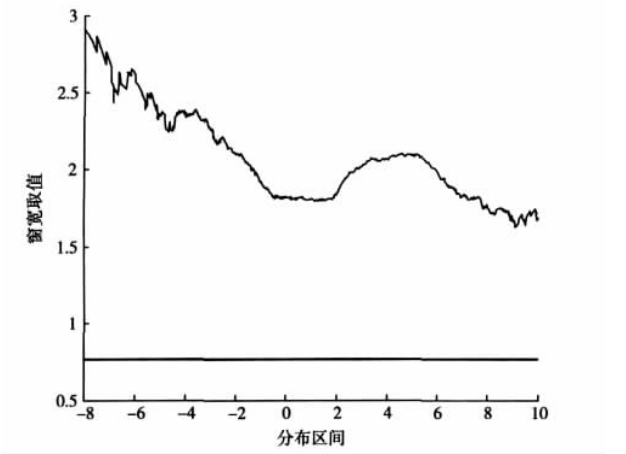

图 3 两种算法窗宽取值的变化情况

Fig. 3 The changing trend of bandwidth values for two algorithms

FBA 算法更加接近这个要求。另外,还可以看出,DBA 算法中,在分布密度大的区域,窗宽值降低,反之,窗宽加大。这说明 DBA 算法可以自动的调整窗宽的值,能够根据样本的分布情况,在不同的估计点而调整窗宽的取值,使其在该点最优。

图 4 为采用不同窗宽的密度估计函数的二阶导数绝对值的变化情况,虚线表示 $|\hat{f}''(x)|$,即,根据 FBA 算法求到的窗宽值 h,计算得到的二阶导数。实线表示 $|\hat{f}''_H(x)|$,即,根据加大窗宽值 H,计算的二阶导数。可以看出: $|\hat{f}''_H(x)|$ 和 $|\hat{f}''(x)|$ 的差异明显, $|\hat{f}''_H(x)|$ 的平滑性明显优于 $|\hat{f}''(x)|$,而通常情况下,真实密度函数的二阶导数应该是较光滑的。

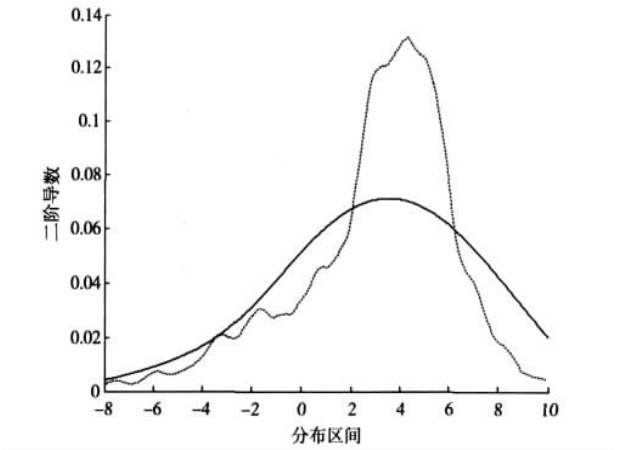

图 4 两种窗宽密度估计函数的二阶导数绝对值的变化情况

Fig. 4 The changing trend of the second derivative of density estimation functions for two different bandwidths

4.2 验证结果统计

采用不同分布的数据,选择适当的窗宽调节系数 a 值,做了多组对比验证,验证结果如表 1 所示,表中的混合分布采用 3 个或者 4 个标准分布混合而成,估计区间位于最小样本和最大样本之间。

通过实验,得到如下的结果:

1) 在保证精度提高的前提下,DBA 算法的估计图形的平滑度上总能优于 FBA 算法的估计结果,部分实验的精度虽然提高不明显,但是,其平滑度改善显著;

2) 选择适当的 a 值,DBA 算法偏差的绝对值的平均值总能小于 FBA 算法的平均值(可以通过观察的方法对 a 进行取值: 在保证 DBA 算法的估计图形与 FBA 算法的估计图形大致相当情况下,尽量增大参数 a 的值,通常能够取得满意效果) ;

3) DBA 算法求得的窗宽,其最小值通常都大于 DBA 算法的窗宽值,然而其估计精度却没有降低。

表 1　两种算法的实验统计结果

Table 1　The experiment statistic results for two algorithms

分布	样本情况 数目	估计点数	方差	估计结果偏差 FBA 算法 h	$EA(\hat{f}(x))$	DBA 算法 a	$EA(\hat{f}(x)_*)$	%
混合正态分布	60	300	3.013 5	1.171 3	0.011 5	4	0.006 7	41.4
			2.193 0	0.852 4	0.018 5	2	0.016 8	9.5
			3.036 1	1.180 1	0.013 1	6	0.010 0	24.1
	300	900	3.115 5	0.877 7	0.007 0	6	0.003 9	44.0
			3.435 4	0.967 8	0.006 5	4	0.005 9	9.3
			2.888 8	0.813 8	0.010 6	2	0.009 4	11.8
	600	900	3.132 9	0.768 3	0.005 6	8	0.003 9	30.0
			3.408 0	0.835 8	0.005 5	5	0.004 8	12.7
			2.435 0	0.597 1	0.009 1	2	0.006 8	24.5
	1 200	1 400	2.896 9	0.618 5	0.006 6	5	0.006 1	6.9
			3.522 9	0.752 1	0.007 3	5	0.006 9	6.4
			2.321 5	0.495 6	0.006 8	8	0.005 6	17.9
混合指数分布	300	1 500	8.0305	2.2622	0.0047	3.5	0.0046	2.9
			10.7835	3.0378	0.0062	2	0.0056	8.4
	1 200	1 500	8.5298	1.8210	0.0066	2	0.0052	20.4
			10.4016	2.2206	0.0052	2	0.0037	27.3
混合 χ^2 分布	100	1 500	5.0995	1.7895	0.0147	3	0.0144	1.9
			4.6482	1.6312	0.0068	4	0.0045	33.6
	1200	1000	4.6558	0.9940	0.0090	5	0.0090	0.45
			5.0695	1.0823	0.0028	8	0.0025	11.7
混合瑞利分布	200	1 500	7.1393	2.1811	0.0038	4	0.0037	3.9
			4.8076	1.4687	0.0181	7	0.0118	34.9
	600	1500	7.2694	1.7827	0.0042	7	0.0039	8.0
			5.2347	1.2837	0.0192	10	0.0121	37.1
混合韦伯分布	60	1 200	0.4959	0.1927	0.1549	2	0.1520	1.9
			0.3390	0.1318	0.1140	1.5	0.1098	3.8
	1 200	1 200	0.2670	0.0570	0.0344	5	0.0310	9.8
			0.4252	0.0908	0.0821	2	0.0629	23.4

4) DBA 算法偏差的绝对值平均值,与估计点个数关系不大。

5) DBA 算法的效果受窗宽调节系数 a 的影响明显。

6　结　　论

DBA 算法采用基于估计点的变窗宽思想,利用 FBA 算法求得的估计函数来代替假设的密度分布函数,实现自适应变窗宽;通过加大估计函数的二阶导数的窗宽值,来减少其与真实密度函数的变差,从而改善估计效果(精度和平滑度) 。

根据目前的研究成果,还需要在以下两方面继续深入研究和探讨:

1) DBA 算法的估计效果受窗宽调节系数 a 的影响很大,关于参数 a 选取的理论或经验公式需要深入研究。

2) DBA 算法的计算量比 FBA 算法的计算量要大,特别是多维密度估计时的计算量更大,因此,对于实时性要求高的应用场合,需要研究减少 DBA 算法计算量的问题。

参考文献

[1] JONES M C, MARRON J S, SHEATHER S J. A brief survey of bandwidth selection for density estimation [J]. Journal of the American Statistical Association, 1996,91 (433) : 401-407.

[2] SAIN S R, BAGGERLY K A, SCOTT D W. Cross-validation of multivariate densities [J]. Journal of the American Statistical Association, 1994,89(427) : 807-817.

[3] BOLANCE C, GUILLEN M, NIELSEN J P. Kernel density estimation of actuarial loss functions [J]. Mathematics and Economics, 2003,32: 19-36

[4] FUKUNAGA K, HOSTETLER L D. The estimation of the gradient of a density function with applications in pattern recognition [J]. IEEE Trans. Information Theory, 1975, 21: 32-40.

[5] HALL P, KANG K. Bandwidth choice for nonparametric classification [J]. The Annals of Statistics, 2005, 33 (1) : 284-306.

[6] SCOTT D W, SAIN S R. Multidimensional density estimation [J]. Handbook of Statistics, 2005: 229-261.

[7] KARUNAMUNI R J, ALBERTS T. On boundary correction in kernel density estimation [J]. Statistical Methodology, 2005,2(3) : 191-212.

[8] BROWME M. A geometric approach to non-parametric density estimation [J]. Pattern Recognition, 2007, 40 (1) : 134-140.

[9] CAO R, JANSSEN P, VERAVERBEKE N. Relative density estimation and local bandwidth selection for censored data [J]. Computational Statistics & Data Analysis, 2001,36(4) : 497-510.

[10] 黄良沛,王广斌,赵先琼. 基于分形维和局部切空间均值重构的非线性降噪方法 [J]. 电子测量与仪器学报, 2010,24(8) : 699-705.
HUANG L P, WANG G B, ZHAO X Q. Nonlinear noise reduction method based on fractal dimension and the local tangent space mean reconstruction [J]. Journal of Electronic Measurement and Instrument, 2010, 24 (8) :

699-705.

[11] 王代华,周锋,吴朝明. 基于阈值的调焦方向判断方法[J]. 仪器仪表学报, 2010,31(8):1813-1818.
WANG D H, ZHOU F, WU CH M. Direction judgment method for autofocus based on threshold [J]. Chinese Journal of Scientific Instrument, 2010, 31 (8): 1813-1818.

[12] JEFFREYS H. Theory of Probability (3nd. ed) [M]. London: Oxford Univ. Press, 1997.

[13] JONES M C. Variable kernel density estimates and variable kernel density estimates [J]. Australian J. Statist, 1990,32:361-371.

[14] 黄美发,景晖,匡兵,等. 基于拟特罗方法的测量不确定度评定[J]. 仪器仪表学报, 2009,30(1):120-125.
HUANG M F, JING H, KUANG B, et al. Measurement uncertainty evaluation based on quasi Monte-Carlo method [J]. Chinese Journal of Scientific Instrument, 2009,30(1):120-125.

[15] 赵春晖,王炜薇,崔颖. 基于 FPGA 的镜像阈值层叠滤波器实现方法[J]. 电子测量与仪器学报, 2009,23(11):42-47.
ZHAO CH H, WANG W W, CUI Y. Implementation of stack filters with mirrored image threshold cascade based on FPGA [J]. Journal of Electronic Measurement and Instrument, 2009,23(11):42-47.

作者简介

邓飙,1996 年于第二炮兵工程学院获得硕士学位,西安交通大学博士研究生,现为第二炮兵工程学院副教授,主要研究方向系统仿真与故障诊断。
E-mail: djm202@163.com

Deng Biao received M. Sc from Second Artillery Engineering Institute in 1996, now he is an associate professor in Second Artillery Engineering Institute. His research interests are system simulation and fault diagnosis.

于传强,分别 2000 年、2003 年、2007 年于第二炮兵工程学院获得学士、硕士、博士学位,现为第二炮兵工程学院讲师,主要研究方向为故障诊断,统计模式识别。
E-mail: fishychq@163.com

Yu Chuanqiang received B. Sc, M. Sc and Ph. D. all from Second Artillery Engineering Institute in 2000, 2003 and 2007, respectively; and he is a lecturer in the same university. His research interests are fault diagnosis and statistical pattern recognition.