岡山大学環境理工学部研究報告
第11巻，第1号，pp.27-30，2006年3月

区間上のデータに対するカーネル密度推定法

植木 優夫[1]　笛田 薫[2]

Kernel density estimation on the interval

Masao Ueki[1] and Kaoru Fueda[2]

(Received December 28, 2005)

abstract

In the field of data analysis, including environmental data, it is important to know the shape of underlying density function. In this case, we often use histogram which provides an information about the broad line of density's curve. However histogram can not be the best method when the true density function is continuous, as is often the cases. On the other hand, kernel density estimator is another popular one which gives a continuous function. In some practical cases, however, there is a case that some knowledges about the range of the data are previously given. For instance, data of percentage, such as mortality rate, only takes the values on $[0,1]$. This paper considers two different modifications in kernel density estimator for the data on known interval and compares them.

Key words: kernel density estimator, adaptive bandwidth, data on finite interval.

1　はじめに

得られたデータを解析するにあたって、初めに行うことの一つに，データの分布形を知ること，厳密に言うとデータが従っている密度関数の形を知ることがある．この大凡のデータの形に基づいて，次にどのような解析手法を使うか決めることができる．データの分布形を得るもっとも単純なツールとしてヒストグラムがもっとも広く知られており，統計の専門家でない人たちにも広く用いられている．ところがヒストグラムの推定する密度関数は連続でなく，またヒストグラムを書くためにデータを区切る端点によって形状が変化するため，十分な方法であるとは言いがたく，さらに2次元以上のデータに対し，ヒストグラムの適用は難しい．ヒストグラムと同じく密度関数を推定する別の方法として，カーネル関数を用いるカーネル密度推定法がある．これもまた広く用いられている方法であり，加えて推定された密度関数は連続関数となる．さらに次元が高くても適用できるというメリットがある (植木と笛田, 2003)．カーネル密度推定法で推定された密度関数は，データの分布形を知るというだけでなく，推定した密度関数自身を使い様々な応用が可能である．ただし，連続性を持たせたことにより，ヒストグラムでは生じなかった次のような問題が現われる．すなわちカーネル密度推定法で扱えるデータは実数上で定義された確率変数を考えるので，データが存在する区間が有限である場合に十分な結果を与えない．密度関数で考えると，密度関数の値が0でければならない区間上で，カーネル密度推定値が0でないということが起こりえる．実際，現実問題への応用の場面で，データの存在しうる範囲があらかじめ定まっているということはよくある．例えば死亡率などの割合のデータは $[0,1]$ 上の値しかとらず，また距離のデータなどは非負である．このようにデータの存在領域がわかっている場合にカーネル密度推定を適用するために，本論文では2つのカーネル密度推定法の改良案を示し，それらの性能

[1]岡山大学大学院環境学研究科博士後期課程
[2]岡山大学大学院環境学研究科人間生態学講座

比較を行う．

2 カーネル密度推定法

まず通常のカーネル密度推定量について述べる．データ $X_1, X_2, \ldots, X_n$ が未知の分布 q にしたがって観測されたとき，その分布の推定を

$$\hat{p}_h(x) = \frac{1}{n}\sum_{t=1}^{n} K_h(x - X_t)$$

によって行う．この $\hat{p}_h(x)$ をカーネル密度推定量とよぶ．ここで x は $\mathbb{R}$ 上の推定を行う点であり，そのとき $\hat{p}_h(x)$ は $q(x)$ の推定量である．また，K_h は

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$$

で，K はカーネル関数とよばれる非負対称単峰かつ定義域上での積分が 1 となる関数である．例えば標準正規密度関数やイパネクニコフカーネル

$$K(u) = 0.75(1 - u^2)I_{\{|u|\leq 1\}}$$

がよく用いられる．特にイパネクニコフカーネルは他のカーネル関数と比べてよい性質を持つことが知られており (Wand and Jones, 1994)，加えて計算コストも小さいので本論文でもこれをカーネル関数として採用する．

また h はバンド幅とよばれる密度推定の滑らかさを調整するパラメータで，$(0, \infty)$ 上に値をとり，$\hat{p}_h$ は $h = 0$ ならば推定した分布は経験分布に，$h = \infty$ ならば $\mathbb{R}$ 上の一様分布になる．つまり h によって経験分布と一様分布を滑らかに繋いでいる．この h は前もって適切に決める必要があって，例えば最小 2 乗クロスバリデーションで選択することができる (Loader, 1999)．以下の図 1 はカーネル密度推定量を標準正規乱数によるデータに対して当てはめたものであり，破線で推定曲線を示した．ここでイパネクニコフカーネルをバンド幅 $h = 0.5$ で用いており，下側の小さい 10 個の破線が各データ点の重みを示す．またヒストグラムもあわせて表示してある．

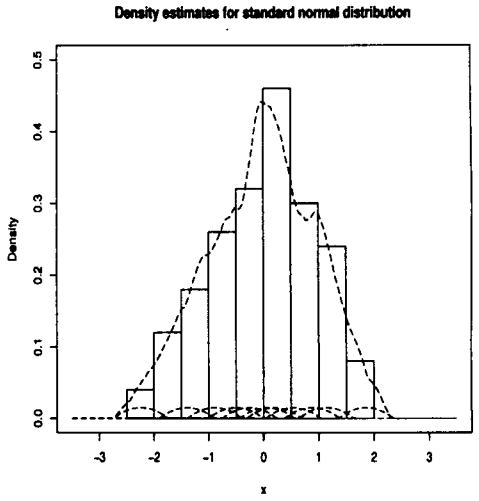

図 1: 標準正規乱数データに対するカーネル密度推定量

3 任意区間上のデータに対する改良

上で定義されたカーネル密度推定量は，有限区間上のデータに対して不十分な結果を導くことがある．すなわちあるバンド幅の値を超えたとき，データの定義域をはみ出すことが起こりうる．以下の図 2 はそのような推定の一例である．これはカーネル密度推定量を $[0, 1]$ 上の一様乱数データに対して当てはめたもので，図 2 と同様に下側の小さい 10 個の曲線が各データ点の重みを示す．バンド幅は $h = 0.1$ を用いた．点線は真の密度関数である．明らかに $[0, 1]$ の外で密度推定値が 0 になっていない．バンド幅を小さくとると 0 になりうるが，それでは $[0, 1]$ 区間上の推定曲線が滑らかでなくなってしまう．

もしデータがある区間でしか値をとらないということが事前に分かる場合，この知識を密度推定に反映させ推定量を改良できるはずである．以下ではそのような問題点を考慮した，カーネル密度推定量の 2 つの改良案を提示する．

いまデータは閉区間 $[a, b]$ 上で観測されると仮定する．要求される条件は

$$x \notin [a, b] \text{ に対して } \hat{p}_h(x) = 0$$

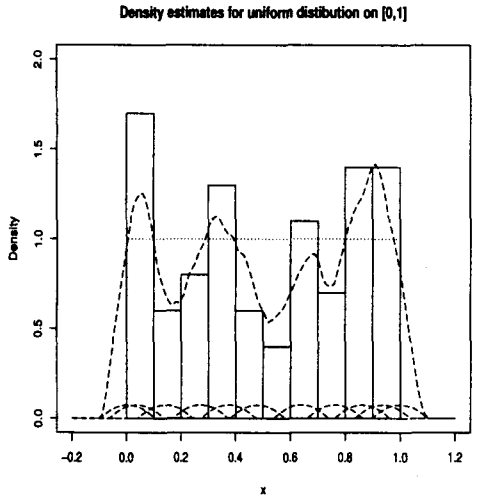

図 2: $[0,1]$ 上の一様乱数データに対するカーネル密度推定量

となることである.

方法 1.

各データ点におけるカーネル関数による重み $K_h(x-X_t)$ が $[a,b]$ をはみ出すと切り取りを行う方法. ただし切り取ればそれだけ積分値が 1 より小さくなるので補正を行う. すなわち

$$\hat{p}_{1,h}(x) = \frac{1}{n}\sum_{t=1}^{n} \frac{K_h(x-X_t)I_{\{x\in[a,b]\}}}{\int_a^b K_h(y-X_t)dy}.$$

方法 2.

区間 $[-1,1]$ で定義されたカーネル関数を考え, 各データ点におけるカーネル関数 $K_h(x-X_t)$ が $[a,b]$ をはみ出ないようにバンド幅を決める方法. 上で考えるカーネル関数として, イパネクニコフカーネルなどが挙げられる. すなわち

$$\hat{p}_{2,h}(x) = \frac{1}{n}\sum_{t=1}^{n} K_{h_t(h)}(x-X_t),$$

ここで $h_t(h) = \min\{h, X_t - a, b - X_t\}$. このように各データ点でのバンド幅をおけば, 必ず $|x-X_t| \le h_t(h)$ が成り立つ. 特に各点ごとにバンド幅を変える方法は適応的バンド幅選択法とよばれる. またそれぞれの推定点 x でバンド幅を変えることも考えられるが, その場合は密度推定値 $\hat{p}_h(x)$ の x に関する積分値が 1 とならず補正が必要となる. この方法として例えば最近傍バンド幅選択法が知られており, 良い性質をもつのであるが計算コストが大きい.

4　2 つの方法の比較

ここでは, 上で導入した 2 つの方法を例を用いて比較する.

4.1　一様分布

死亡率など $[0,1]$ 上の密度関数の例として $[0,1]$ 上の一様分布を考える. これより乱数を用いて発生させた人工データに対し, 2 つの方法のそれぞれの推定結果を次の図 3 で示す. ただし, カーネル関数としてイパネクニコフカーネルを用いた. 密度関数の定義域 $[a,b]$ として $a=0, b=1$ をそれぞれ事前に入力し, バンド幅は $h=0.1$ を用いた. 実線は方法 1, 破線は方法 2, 点線は真の密度関数をそれぞれあらわす.

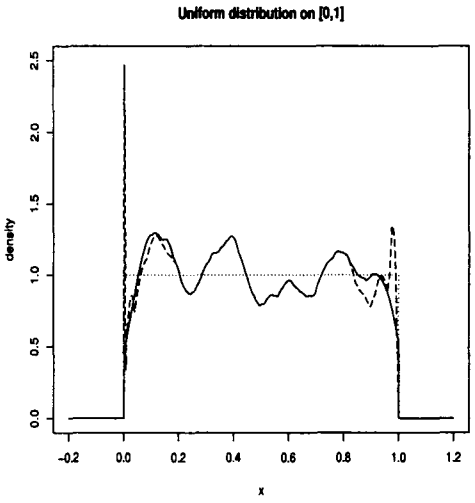

図 3: $[0,1]$ 上の一様分布に対する密度推定

図 3 において, 方法 1(実線) と方法 2(破線) のプロットは定義域の中心の辺りで一致してい

る．ところが定義域の端に近い点つまり $x=0$ または $x=1$ の付近で，方法2は明らかに不正に大きい密度関数の値を推定している．この理由として方法2は，定義域の端に近いデータは，そうでないデータに比べて小さいバンド幅が選ばれるせいである．特にデータが境界に非常に近いとバンド幅は0に近くなり，1点で鋭く上昇するカーネル関数となってしまう．それに対して，方法1はバンド幅を大きくとれば $[0,1]$ 上の一様分布に近づく．これは方法1の定義より，どんなデータに対しても言えることで，今の場合は明らかに良い推定を与える．

4.2 カイ2乗分布

距離などの非負データの密度関数の例として自由度10のカイ2乗分布を考える．この場合定義域は $[0,\infty)$ である．これより発生させた人工データに対し，それぞれ2つの方法の推定結果を次の図4で示す．ただし，カーネル関数としてイパネクニコフカーネルを用いた．密度関数の定義域 $[a,\infty)$ として $a=0$ を事前に入力し，バンド幅は $h=5$ を用いた．実線は方法1，破線は方法2，点線は真の密度関数をそれぞれあらわす．

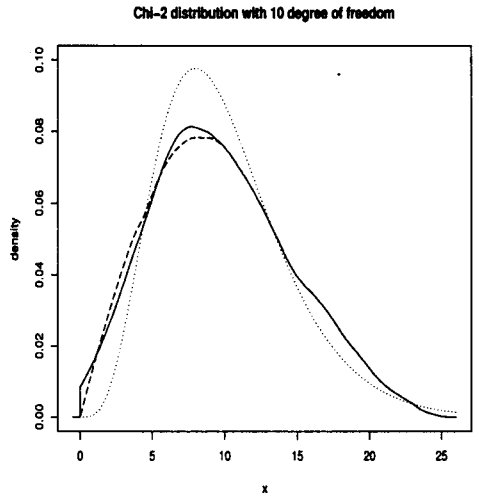

図4: 自由度10のカイ2乗分布に対する密度推定

図4の結果をみると，方法1と方法2は定義域の端，つまり点 $x=0$ の近く以外でほぼ同じような結果を与えていることがわかる．ところが方法1は境界 $x=0$ において不連続な曲線を推定している．これは打ち切りを行ったためであり，今の例ではあまり望ましい結果であるとは言えない．一方で方法2は連続になっており，定義域の外へ向かうとき滑らかに0へ収束している．

5 まとめと今後の展望

本論文は，あらかじめデータの存在しうる区間がわかっている場合のカーネル密度推定量の2つの改良案を示しそれらを比較した．4節の結果をみると2つの改良法はどちらも従来の方法の問題点は解決しているものの，方法1はデータの存在する区間の端点で密度関数が連続である場合にも不連続性を生じ，方法2は区間の端点で密度関数が正である場合に端点で鋭く上昇するという欠点がある．両者の良い性質のみを引き継ぐ新しい方法の開発が次回の目標である．このようなデータの可視化に関する需要は今後さらに高まることが予想され，実際の環境データ解析へ応用することが重要な課題であろう．またデータの次元が2次元以上となる場合も考える必要がある．

参考文献

[1] Loader, C. R. (1999). *Local regression and likelihood.* Springer, New York.

[2] 植木 優夫, 笛田 薫. (2003). カーネル密度推定におけるカーネル関数の比較. 日本計算機統計学会, 第17回大会論文集, 147-150.

[3] Wand, M. P. and Jones, M. C. (1994) *Kernel Smoothing.* Monographs on Statistics and Applied Probability 66, Chapman and Hall, London.