



バンド幅をクロス・バリデーションで選んだ場合のカーネル密度推定による信頼区間に関するシミュレーション分析

難波, 明生

(Citation)

国民経済雑誌, 197(5):73-86

(Issue Date)

2008-05

(Resource Type)

departmental bulletin paper

(Version)

Version of Record

(JaLCD01)

<https://doi.org/10.24546/00056229>

(URL)

<https://hdl.handle.net/20.500.14094/00056229>



バンド幅をクロス・バリデーションで選んだ 場合のカーネル密度推定による 信頼区間に関するシミュレーション分析

難 波 明 生

カーネル密度推定では、データに基づいてバンド幅を選択する方法として、多くの場合クロス・バリデーションが用いられている。クロス・バリデーションによって選ばれたバンド幅は、平均自乗誤差の積分値を最小化する値に確率収束する事が知られている。しかしながら、バンド幅にこの値を用いた場合、カーネル密度推定量のバイアスと信頼限界は同じオーダーを持ち、信頼限界は漸近的にも推定量のバイアスの影響を受ける事になる。本稿では、クロス・バリデーションによりバンド幅を選択した場合、カーネル密度推定に基づく信頼区間がどのような小標本特性を持つのかをシミュレーションにより分析する。

キーワード ノンパラメトリック法, カーネル密度推定,
クロス・バリデーション, ブートストラップ法

1 は じ め に

カーネル密度推定は、ノンパラメトリックに確率密度関数を推定する方法として最も良く知られている方法である。カーネル密度推定を用いる事により、確率密度関数の点推定量のみならず、信頼区間を求めることができる。

カーネル密度推定を用いる際に非常に重要になるのが、バンド幅 (Bandwidth) の選択である。多くの場合、バンド幅は MISE (Mean Integrated Squared Error) を最小にするように選ぶ事が望ましいと考えられているが、MISE は推定の対象である未知の確率密度関数に依存した値となる。したがって、推定の対象となっている確率密度関数が分からなければ、最適なバンド幅の値を求める事はできない。そこで、通常は MISE の近似値を最小にする方法として、クロス・バリデーション (Cross Validation) によってバンド幅が選択される。クロス・バリデーションによって選択されたバンド幅は MISE を最小化する値に確率収束する事が Hall (1983), Stone (1984), Härdle, Hall and Marron (1988) 等によって示されている。

MISE を最小化するようにバンド幅を選択した場合、得られた推定量自体は一致性を持つが、信頼区間の上限と下限、つまり信頼限界は推定量のバイアスと同じオーダーを持つ事が知られている。この結果、信頼限界は漸近的にも推定量のバイアスから影響を受ける事になる。したがって、クロス・バリデーションによって得られたバンド幅を用いた場合にも、標本が十分に大きかったとしても信頼区間はバイアスを持つと考えられる。さらに、カーネル密度推定により確率密度関数の信頼区間を求める際には、通常は推定量の漸近正規性が用いられるため、得られた信頼区間が小標本でも正確であるとは言いがたい。推定量の漸近的な性質しか分かっていない場合にしばしば有効であるのが、Efron (1979) により提案されたブートストラップ法である。Hall (1991, 1992a, 1992b) は、カーネル密度推定においても、バイアスを除去すれば、ブートストラップ法が応用可能であることを示している。

しかしながら、以上の分析はほとんど標本が十分に大きいことを想定しており、小標本特性を分析したものはわずかしかない。また、小標本特性を分析した研究においても、バンド幅はクロス・バリデーション以外の方法で決定されたものがほとんどである。そこで本稿では、クロス・バリデーションを用いた場合の信頼区間の小標本特性をシミュレーションにより分析する。

2 カーネル密度推定

$X_1, X_2, \dots, X_n$ は確率密度関数 $f(x)$ をもつ確率分布から独立に得られた標本であるとする。 $f(x)$ の関数型が、例えば平均 μ , 分散 σ^2 の正規分布の確率密度関数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x-\mu^2}{2\sigma^2}\right] \quad (1)$$

であることが既知であれば、最尤法などの推定法を用いてパラメータ μ, σ^2 を推定することにより、 $f(x)$ の推定量を得ることができる。このような推定法は、パラメトリック推定と呼ばれる。パラメトリックな推定法を行う場合は、上の例のように $f(x)$ の関数型を前もって定式化しなければならない。しかし、多くの場合 $f(x)$ の関数型は未知であり、定式化の誤りが生じる可能性がある。

これに対し、ノンパラメトリック推定では、 $f(x)$ の関数型を前もって定めることなく推定を行うので、定式化の誤りが生じる可能性は無い。確率密度関数をノンパラメトリックに推定する最も簡単な方法は以下のようなものである。

$F(x)$ を $f(x)$ に対応する分布関数とすると、

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_{-\infty}^x f(x) dx \end{aligned} \quad (2)$$

であるから

$$\begin{aligned}
 f(x) &= \frac{d}{dx} F(x) \\
 &\approx \frac{F(x+h/2) - F(x-h/2)}{h} \\
 &= \frac{P(x-h/2 \leq X \leq x+h/2)}{h} \\
 &= \frac{P\left(\frac{1}{2} \leq \frac{X-x}{h} \leq \frac{1}{2}\right)}{h}
 \end{aligned} \tag{3}$$

となる。(3)の分子を推定値で置き換えることにより、 $f(x)$ は

$$\begin{aligned}
 \hat{f}_n(x) &= \frac{\left(\left[-\frac{1}{2}, \frac{1}{2}\right] \text{に入っている } \frac{x_i - x}{h} \text{ の個数} \right) / n}{h} \\
 &= \frac{1}{nh} \sum_{i=1}^n I\left(-\frac{1}{2} \leq \frac{X_i - x}{h} \leq \frac{1}{2}\right)
 \end{aligned} \tag{4}$$

のように推定される。ただし、 $I(A)$ は事象 A が起こったときに1、それ以外では0の値を取る関数であり、Indicator Function と呼ばれる。また、 $h(>0)$ はバンド幅 (Bandwidth) と呼ばれる、分析者が定めるパラメータである。この推定量は Naive Estimator, または、Local Histogram Estimator 等と呼ばれている。

(4)で与えられる推定量は、Local Histogram Estimator という名前も示しているように、滑らかな関数ではない。そこで、Rosenblatt (1956) は(4)を拡張し、次のような推定量を考案した。

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \tag{5}$$

h は(4)の場合同様にバンド幅である。また、 $K(\cdot)$ はカーネルと呼ばれる関数であり、 $\hat{f}(x)$ が一致性を持つために、通常以下の仮定が置かれる。ただし、表記を簡単にするために、今後は積分範囲が明示されていない場合には、積分範囲は積分する変数の定義域全体であるとする。

仮定1 カーネル $K(\cdot)$ は以下の性質を満たすものとする。

- (i) $\int K(v) dv = 1$
- (ii) $K(v) = K(-v)$
- (iii) $\int v^2 K(v) dv = \kappa_2 > 0$

ここで、仮定 1 の内容を簡単に見ておこう。(i) が満たされることにより、

$$\begin{aligned}
 \int \hat{f}(x) dx &= \int \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) dx \\
 &= \frac{1}{nh} \sum_{i=1}^n \int K\left(\frac{X_i - x}{h}\right) dx \\
 &= \frac{1}{nh} \sum_{i=1}^n \int K(\phi_i) h d\phi_i \\
 &= \frac{1}{n} \sum_{i=1}^n \int K(\phi_i) d\phi_i = 1
 \end{aligned} \tag{6}$$

となる。ただし、 $\phi_i = (X_i - x)/h$ である。このことから、 $K(\cdot) \geq 0$ であれば、 $\hat{f}(x) \geq 0$ かつ $\int \hat{f}(x) dx = 1$ となるので、カーネル密度推定量 $\hat{f}(x)$ は確率密度関数であるといえる。また、(ii) は $K(v)$ が $v=0$ について左右対称であることを意味する。(i)、(ii) を満たす $K(\cdot)$ としては $v=0$ について左右対称である確率密度関数を用いれば良い。 $K(\cdot)$ としてこのような確率密度関数を用いた場合、(iii) は確率密度関数 $K(\cdot)$ を持つ分布の分散が有限な正の値を持てば良いことを意味している。また、この時、(ii) により確率密度関数 $K(\cdot)$ を持つ分布の平均は 0 である。ただし、仮定 1 では $K(\cdot) \geq 0$ という仮定は置かれていないので、必ずしも確率密度関数をカーネルとして用いなければならない訳ではない。

$X_1, X_2, \dots, X_n$ が 3 回微分可能な確率密度関数 $f(x)$ を持つ確率分布から独立に得られた標本であるとする、 $\hat{f}(x)$ のバイアスは次のように計算される。

$$\begin{aligned}
 \text{Bias}[\hat{f}(x)] &= E[\hat{f}(x)] - f(x) \\
 &= E\left[\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)\right] - f(x) \\
 &= h^{-1} E\left[K\left(\frac{X_1 - x}{h}\right)\right] - f(x) \\
 &= h^{-1} \int f(x_1) K\left(\frac{x_1 - x}{h}\right) dx_1 - f(x) \\
 &= h^{-1} \int f(x + hv) K(v) h dv - f(x) \\
 &= \int \left\{ f(x) + f^{(1)}(x) hv + \frac{1}{2} f^{(2)}(x) h^2 v^2 + O(h^3) \right\} k(v) dv - f(x) \\
 &= \frac{h^2}{2} f^{(2)}(x) \kappa_2 + O(h^3)
 \end{aligned} \tag{7}$$

となる。2 行目から 3 行目の変形では X_i が独立に同一の分布に従う事、4 行目から 5 行目の変形では $(x_1 - x)/h = v$ という変数変換、5 行目から 6 行目への変形では $f(x + hv)$ のテ

ーラ一展開を用いている。

同様の計算により、 $\hat{f}(x)$ の分散は

$$\begin{aligned}\text{Var}[\hat{f}(x)] &= E[\hat{f}(x) - E[\hat{f}(x)]] \\ &= \frac{1}{nh} \{ \kappa f(x) + O(h) \}\end{aligned}\quad (8)$$

である事を示す事ができる。ただし、 $\kappa = \int K^2(v) dv$ である。(7)と(8)を用いれば、 $\hat{f}(x)$ の平均自乗誤差 (Mean Squared Error; MSE) は

$$\begin{aligned}\text{MSE}[\hat{f}(x)] &= E[(\hat{f}(x) - f(x))^2] \\ &= \text{Var}[\hat{f}(x)] + (\text{Bias}[\hat{f}(x)])^2 \\ &= \frac{h^4}{4} [\kappa_2 f^{(2)}(x)]^2 + \frac{\kappa f(x)}{nh} + o(h^4 + (nh)^{-1}) = O(h^4 + (nh)^{-1})\end{aligned}\quad (9)$$

となる。したがって、 $c > 0, \beta > 1$ を定数とし $h = cn^{-1/\beta}$ のようにバンド幅を選択すれば、 $n \rightarrow \infty$ の時 $h \rightarrow 0, nh \rightarrow \infty$ となり $\hat{f}(x)$ は $f(x)$ に平均自乗収束する。したがって、バンド幅が以上の条件を満たせば、 $\hat{f}(x)$ は $f(x)$ の一致推定量であると言える。さらに、次の定理が成立する事が知られている。

定理 1 $X_1, X_2, \dots, X_n$ が確率密度関数 $f(x)$ を持つ確率分布から独立に得られた標本であるとする。また、 $n \rightarrow \infty$ の時、 $h \rightarrow 0, nh \rightarrow \infty$ かつ $nh^7 \rightarrow 0$ であるとする。この時、

$$\sqrt{nh} \left(\hat{f}(x) - f(x) - \frac{\kappa_2 h^2}{2} f^{(2)}(x) \right) \xrightarrow{d} N(0, \kappa f(x)) \quad (10)$$

が成立する。(証明は Li and Racine (2007) 等を参照。)

3 バンド幅の選択

前節で紹介したカーネル密度推定を用いるためには、分析者はカーネル $K(\cdot)$ およびバンド幅 h を選択しなければならない。一般にカーネルの関数型が結果に及ぼす影響はさほど大きくないが、バンド幅が及ぼす影響は非常に大きいと言われている。したがって、ここではバンド幅の選択方法について説明する。

3.1 Mean Integrated Squared Error

バンド幅の選択基準の一つは、(9)で与えられる MSE を x について積分した Mean Integrated Squared Error (MISE)

$$\text{MISE}[\hat{f}(x)] = \int \text{MSE}[\hat{f}(x)] dx$$

$$\begin{aligned}
&= \int \left\{ \frac{h^4}{4} [\kappa_2 f^{(2)}(x)]^2 + \frac{\kappa f(x)}{nh} + o(h^4 + (nh)^{-1}) \right\} dx \\
&= \frac{1}{4} h^4 \kappa_2^2 \int [f^{(2)}(x)]^2 dx + \frac{\kappa}{nh} + o(h^4 + (nh)^{-1}) \quad (11)
\end{aligned}$$

を最小にすることである。そのようなバンド幅は(11)を h に関して微分して 0 とおく事で

$$h_{\text{opt}} = c_0 n^{-1/5} \quad (12)$$

である事が容易に分かる。ただし

$$c_0 = \kappa^{1/5} \kappa_2^{-2/5} \left\{ \int [f^{(2)}(x)]^2 dx \right\}^{-1/5} \quad (13)$$

は正の定数である。したがって、 h_{opt} は $f^{(2)}(x)$ に依存する事が分かるが、 $f^{(2)}(x)$ は通常未知であるため、実際には h_{opt} をバンド幅として利用する事はできない。このバンド幅を利用する方法の一つは、 $f(x)$ に何らかの特定の分布（例えば正規分布）を仮定して(13)の値を計算し、必要であれば $f(x)$ に含まれるパラメータを推定値で置き換えて利用する事である。例えば、カーネル $K(\cdot)$ として標準正規分布の確率密度関数を用いた場合、 $f(x)$ が平均 μ , 分散 σ^2 の正規分布の確率密度関数であると仮定すれば、 $h \approx 1.06 \sigma n^{-1/5}$ となる。したがって、分析者は h として $1.06 \hat{\sigma} n^{-1/5}$ を用いる事になる。ただし、 $\hat{\sigma}$ は $X_1, X_2, \dots, X_n$ から得られる標本標準偏差である。

3.2 クロス・バリデーション

データに基づいてバンド幅を選択する基準としてよく用いられるのがクロスバリデーション (Cross Validation) と呼ばれる方法である。以下では、クロスバリデーションについて簡単に説明する。

まず、 $\hat{f}(x)$ と $f(x)$ の距離の 2 乗、つまり

$$\int [\hat{f}(x) - f(x)]^2 dx = \int \hat{f}(x)^2 dx - 2 \int \hat{f}(x) f(x) dx + \int f(x)^2 dx \quad (14)$$

を考える。バンド幅 h はこの値を最小にするように定めるのが望ましい。第 3 項は h に依存しないので、 h は最初の 2 項、つまり

$$\int \hat{f}(x)^2 dx - 2 \int \hat{f}(x) f(x) dx \quad (15)$$

を最小にするように定めればよい。第 2 項については $\int \hat{f}(x) f(x) dx = E_X[\hat{f}(X)]$ と考える事ができるが、この値は未知の確率密度関数 $f(x)$ に依存しているので実際に求める事はできない。ただし、 $E_X[\cdot]$ は X に関する期待値である。そこで、 $E_X[\hat{f}(X)]$ は推定値 $\frac{1}{n} \sum_{i=1}^n \hat{f}_i(X_i)$ で置き換える。ただし、

$$\hat{f}_{-1}(X_i) = \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K\left(\frac{X_i - X_j}{h}\right) \quad (16)$$

は $f(x)$ の leave-one-out estimator と呼ばれる。第1項については

$$\begin{aligned} \int f(x)^2 dx &= \frac{1}{n^2 h^2} \sum_{i=1}^n \sum_{j=1}^n \int K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right) dx \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \int K\left(\frac{X_i - X_j}{h} - t\right) K(t) dt \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n (K * K)\left(\frac{X_i - X_j}{h}\right) \end{aligned} \quad (17)$$

と計算される。第1行目から2行目への変形では $t = (x - X_j)/h$ と変数変換を行っている。

また、 $(K * K)(u) = \int K(u - t) K(t) dt$ はカーネル $K(\cdot)$ の重畳積分 (Convolution) である。

以上の結果をまとめて

$$\frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n (K * K)\left(\frac{X_i - X_j}{h}\right) - \frac{2}{n(n-1)h} \sum_{j=1, j \neq i}^n K\left(\frac{X_i - X_j}{h}\right) \quad (18)$$

を最小化するように h を求める方法をクロス・バリデーションと呼ぶ。実際の h はコンピュータを用いて数値的に求める事になる。Hall (1983), Stone (1984) はクロス・バリデーションにより選択された h は $n \rightarrow \infty$ の時、(12)で与えられる h_{opt} に確率収束する事を示している。

4 信頼区間と問題点

本節では $f(x)$ の信頼区間を求める方法とその問題点について説明する。

4.1 正規分布による近似

$f(x)$ の信頼区間を簡単に求めるには定理1の結果を用いれば良い。 z_α を標準正規分布の $\alpha\%$ 点とすると、定理1により近似的に

$$P\left(\left|\frac{\sqrt{nh}\left(\hat{f}(x) - f(x) - \frac{\kappa_2 h^2}{2} f^{(2)}(x)\right)}{\sqrt{\kappa \hat{f}(x)}}\right| < z_{\alpha/2}\right) = \alpha \quad (19)$$

が成立する。ここで問題となるのが(19)の括弧内の左辺第3項、つまり推定のバイアス項である。 $h \propto n^{-1/\beta}$ かつ $1 < \beta < 5$ であれば $n \rightarrow \infty$ の時 $\sqrt{nh} \cdot h^2 \rightarrow 0$ となるのでバイアス項の存在は無視できる。しかし、前節でみたように、クロス・バリデーションによって選ばれた h は $h_{\text{opt}} = c_0 n^{-1/5}$ に確率収束する。したがって、クロスバリデーションによりバンド幅を選択した場合、漸近的にもバイアス項の存在は無視できないものとなる。したがって、正規分

布による近似に基づいて信頼区間を求める場合には、単純にバイアス項を無視してしまうか、 $f^{(2)}(x)$ を推定値で置き換える等の方法が用いられる事になる。

4.2 ブートストラップ法

前述の正規分布による近似法は、漸近的にのみ有効な方法であり、仮にバイアス項が無いと仮定しても、小標本ではあまり正確でない可能性がある。統計量の漸近分布しか分かっていない場合でも、小標本特性を改善できる可能性がある方法としてしばしば用いられるのが Efron (1979) により提案されたブートストラップ法である。

ブートストラップ法は手元にある標本からリサンプリングを行う事により、統計量の分布を近似する方法である。以下では、その手順を説明する。

1. 手元にある標本 $X_1, X_2, \dots, X_n$ から重複を許して大きさ n の標本を無作為抽出する。抽出した標本を $X_1^*, X_2^*, \dots, X_n^*$ とする。
2. ステップ 1 で得られた標本 $X_1^*, X_2^*, \dots, X_n^*$ を用いて、分布を近似したい統計量の値を計算する。ただし、未知パラメータは $X_1, X_2, \dots, X_n$ を用いて計算される推定値で置き換える。

ここでは、分布を近似したい統計量は

$$\frac{\sqrt{nh}(\hat{f}(x) - f(x))}{\sqrt{\kappa f(x)}} \quad (20)$$

であるから、

$$\frac{\sqrt{nh}(\hat{f}^*(x) - \hat{f}(x))}{\sqrt{\kappa \hat{f}^*(x)}} \quad (21)$$

を計算する。ただし、 $\hat{f}(x)$ および $\hat{f}^*(x)$ はそれぞれ $X_1, X_2, \dots, X_n$ および $X_1^*, X_2^*, \dots, X_n^*$ を用いて計算したカーネル密度推定値である。

3. ステップ 1-2 を B 回繰り返し、 B 個の (21) の推定値を得る。得られた B 個の推定値の経験分布を用いて統計量の分布を近似する。

上記のブートストラップ法では、 $X = \{X_1, X_2, \dots, X_n\}$ を所与として、 $X_1, X_2, \dots, X_n$ の経験分布、つまり $P(X = X_i) = \frac{1}{n}$ である分布から得られた $X_1^*, X_2^*, \dots, X_n^*$ を用いて統計量の分布を近似している事になる。しかし、(21) の分子を考えると、

$$\begin{aligned} E[\hat{f}^*(x) | X] &= \frac{1}{nh} \sum_{i=1}^n E\left[K\left(\frac{x - X_i^*}{h}\right) | X\right] \\ &= \frac{1}{h} E\left[K\left(\frac{x - X_i}{h}\right) | X\right] \end{aligned}$$

$$= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) = \hat{f}(x) \quad (22)$$

であるから、(21)の平均は0であるという事になり、平均0ではない(20)の分布を近似できない。(21)の経験分布は、実際には

$$\frac{\sqrt{nh}(\hat{f}(x) - E[\hat{f}(x)])}{\sqrt{\kappa \hat{f}(x)}} \quad (23)$$

の分布を近似している事になる¹⁾。したがって、ブートストラップ法を用いた場合も、(19)のバイアス項を単純に無視するか、推定値で置き換える等の方法をとらなければならない事になる。

5 シミュレーション分析

ここまで見てきたように、カーネル密度推定により $f(x)$ 信頼区間を求める時、クロス・バリデーションによりバンド幅を選んだ場合には、大標本においてさえもバイアスの存在が無視できなくなる。さらに、小標本においては、正規分布による近似自体もどの程度正確かは分からない。したがって本節では、クロス・バリデーションによりバンド幅を選択したとき、信頼区間がどのような小標本特性を持つのかをシミュレーションにより分析する。分析対象とする信頼区間は以下の通りである。

方法1. バイアス項を無視して、単純に正規近似から信頼区間を求める。つまり、信頼係数 $100 \times \alpha$ % 信頼区間は

$$(\hat{f}(x) - z_{\alpha/2}(nh)^{-1/2}\sqrt{\kappa \hat{f}(x)}, \hat{f}(x) + (nh)^{-1/2}z_{\alpha/2}\sqrt{\kappa \hat{f}(x)}) \quad (24)$$

となる。

方法2. バイアス項を理論的に求め、バイアスを修正して正規近似により信頼区間を求める。この時、信頼係数 $100 \times \alpha$ % の信頼区間は

$$\left(\hat{f}(x) - \frac{\kappa_2 h^2}{2} f^{(2)}(x) - z_{\alpha/2}(nh)^{-1/2}\sqrt{\kappa \hat{f}(x)}, \right. \\ \left. \hat{f}(x) - \frac{\kappa_2 h^2}{2} f^{(2)}(x) + (nh)^{-1/2}z_{\alpha/2}\sqrt{\kappa \hat{f}(x)} \right) \quad (25)$$

となる。この方法は、 $f^{(2)}(x)$ が未知であるため、現実の問題には実行可能ではない点に注意が必要である。

方法3. バイアス項は理論的に求め、正規近似の代わりにブートストラップ法を用いて信頼区間を求める。この時、信頼係数 $100 \times \alpha$ % 信頼区間は

$$\left(\hat{f}(x) - \frac{\kappa_2 h^2}{2} f^{(2)}(x) - t_{(1-\alpha/2) \times B}^* (nh)^{-1/2} \sqrt{\kappa \hat{f}(x)}, \right. \\ \left. \hat{f}(x) - \frac{\kappa_2 h^2}{2} f^{(2)}(x) + (nh)^{-1/2} t_{(\alpha/2) \times B}^* \sqrt{\kappa \hat{f}(x)} \right) \quad (26)$$

となる。ただし、 t^* は、ブートストラップ法によって得られた B 個の (21) の値を昇順に並べた時の l 番目の値である。方法 2 と同様に、この方法も現実の問題には実行可能ではない。

方法 4. バイアス項を推定値で置き換え、正規近似により信頼区間を求める。つまり、信頼係数 $100 \times \alpha \%$ の信頼区間は

$$\left(\hat{f}(x) - \frac{\kappa_2 h^2}{2} f^{(2)}(x) - z_{\alpha/2} (nh)^{-1/2} \sqrt{\kappa \hat{f}(x)}, \right. \\ \left. \hat{f}(x) - \frac{\kappa_2 h^2}{2} f^{(2)}(x) + (nh)^{-1/2} z_{\alpha/2} \sqrt{\kappa \hat{f}(x)} \right) \quad (27)$$

となる。ただし、 $f^{(2)}(x)$ の推定値 $\hat{f}^{(2)}(x)$ は Bhattacharya (1967), Schuste (1969) 等が提案しているように、 $\hat{f}(x)$ を微分する事により得られる

$$\hat{f}^{(2)}(x) = \frac{1}{nh^3} \sum_{i=1}^n K^{(2)} \left(\frac{X_i - x}{h} \right) \quad (28)$$

を用いる。

方法 5. バイアス項を推定値で置き換え、ブートストラップ法により信頼区間を求める。つまり、信頼係数 $100 \times \alpha \%$ の信頼区間は

$$\left(\hat{f}(x) - \frac{\kappa_2 h^2}{2} \hat{f}^{(2)}(x) - t_{(1-\alpha/2) \times B}^* (nh)^{-1/2} \sqrt{\kappa \hat{f}(x)}, \right. \\ \left. \hat{f}(x) - \frac{\kappa_2 h^2}{2} \hat{f}^{(2)}(x) + (nh)^{-1/2} t_{(\alpha/2) \times B}^* \sqrt{\kappa \hat{f}(x)} \right) \quad (29)$$

となる。

シミュレーションは以下の設定で行った。まず、 $f(x)$ としては標準正規分布 (I), 自由度 3 のカイ 2 乗分布 (II), 自由度 3 の t 分布 (III) を考えた。信頼区間を求める点 x は分布の平均 μ および、平均 μ から $\pm 1.0, \pm 0.5$ 離れた点を用いた。カーネル $K(\cdot)$ には標準正規分布の確率密度関数を用い、バンド幅 h は前節で説明したクロス・バリデーションにより求めた。信頼係数に関しては $\alpha = 0.1, 0.05, 0.01$ とし、90%, 95%, 99% の信頼区間を求めた。標本の大きさは $n = 20, 50, 100, 200, 500$, ブートストラップ法における繰り返しの回数 $B = 2000$ とし、10000 回の繰り返し実験を行い、上記の信頼区間が真の $f(x)$ を含んでいる割合を計算した。特徴的な結果である $n = 50, n = 200$ の場合の結果が表 1, 表 2 である。

表 1 $n = 50$ の結果

分布	$x - \mu$	方法 1			方法 2			方法 3			方法 4			方法 5		
		99%	95%	90%	99%	95%	90%	99%	95%	90%	99%	95%	90%	99%	95%	90%
I	-1.0	0.9873	0.9625	0.9333	0.9873	0.9625	0.9333	0.9857	0.9384	0.8838	0.9580	0.8954	0.8305	0.9103	0.8060	0.7198
	-0.5	0.9791	0.9139	0.8436	0.9922	0.9777	0.9583	0.9806	0.9336	0.8799	0.9612	0.8923	0.8308	0.8865	0.7530	0.6605
	0.0	0.9540	0.8582	0.7733	0.9938	0.9827	0.9643	0.9818	0.9289	0.8754	0.9597	0.8893	0.8206	0.8517	0.7127	0.6207
	0.5	0.9738	0.9137	0.8437	0.9888	0.9741	0.9550	0.9818	0.9291	0.8744	0.9549	0.8870	0.8283	0.8848	0.7566	0.6692
	1.0	0.9873	0.9627	0.9312	0.9873	0.9627	0.9312	0.9859	0.9386	0.8823	0.9554	0.8894	0.8268	0.9076	0.8011	0.7155
II	-1.0	0.9829	0.9492	0.9059	0.9855	0.9586	0.9216	0.9788	0.9185	0.8503	0.9536	0.8833	0.8167	0.8946	0.7771	0.6894
	-0.5	0.9855	0.9623	0.9324	0.9854	0.9621	0.9322	0.9842	0.9382	0.8824	0.9538	0.8860	0.8194	0.8995	0.7955	0.7115
	0.0	0.9840	0.9574	0.9236	0.9833	0.9557	0.9218	0.9836	0.9363	0.8819	0.9478	0.8730	0.8023	0.8997	0.7934	0.7124
	0.5	0.9799	0.9551	0.9220	0.9790	0.9538	0.9214	0.9841	0.9407	0.8930	0.9443	0.8724	0.8026	0.9113	0.8093	0.7289
	1.0	0.9776	0.9528	0.9175	0.9763	0.9505	0.9175	0.9833	0.9435	0.8937	0.9416	0.8717	0.8053	0.9207	0.8269	0.7472
III	-1.0	0.9875	0.9651	0.9370	0.9735	0.9264	0.8779	0.9631	0.8821	0.8059	0.9553	0.8902	0.8241	0.8945	0.7867	0.7130
	-0.5	0.9548	0.8749	0.8026	0.9920	0.9763	0.9571	0.9831	0.9413	0.8896	0.9506	0.8761	0.8069	0.8772	0.7541	0.6630
	0.0	0.8726	0.7566	0.6694	0.9880	0.9670	0.9357	0.9074	0.7659	0.6423	0.9259	0.8356	0.7589	0.8135	0.6800	0.5883
	0.5	0.9545	0.8757	0.8014	0.9927	0.9776	0.9581	0.9824	0.9370	0.8872	0.9524	0.8790	0.8058	0.8774	0.7466	0.6536
	1.0	0.9877	0.9658	0.9375	0.9759	0.9287	0.8737	0.9650	0.8848	0.8058	0.9558	0.8932	0.8215	0.8961	0.7885	0.7105

表 2 $n = 200$ の結果

分布	$x - \mu$	方法 1			方法 2			方法 3			方法 4			方法 5		
		99%	95%	90%	99%	95%	90%	99%	95%	90%	99%	95%	90%	99%	95%	90%
I	-1.0	0.9934	0.9700	0.9334	0.9934	0.9700	0.9334	0.9881	0.9440	0.8902	0.9623	0.8887	0.8220	0.9286	0.8268	0.7474
	-0.5	0.9873	0.9357	0.8707	0.9959	0.9755	0.9468	0.9854	0.9371	0.8800	0.9598	0.8901	0.8235	0.9097	0.7935	0.7050
	0.0	0.9742	0.8858	0.8030	0.9962	0.9799	0.9571	0.9867	0.9389	0.8881	0.9642	0.8943	0.8278	0.9000	0.7744	0.6858
	0.5	0.9870	0.9366	0.8702	0.9946	0.9764	0.9509	0.9851	0.9406	0.8841	0.9640	0.8940	0.8272	0.9121	0.7923	0.7106
	1.0	0.9928	0.9650	0.9293	0.9928	0.9650	0.9293	0.9863	0.9376	0.8861	0.9591	0.8854	0.8166	0.9223	0.8201	0.7404
II	-1.0	0.9914	0.9674	0.9320	0.9919	0.9676	0.9343	0.9885	0.9436	0.8916	0.9597	0.8816	0.8128	0.9251	0.8189	0.7381
	-0.5	0.9922	0.9631	0.9246	0.9922	0.9630	0.9248	0.9878	0.9422	0.8899	0.9555	0.8805	0.8088	0.9290	0.8300	0.7482
	0.0	0.9905	0.9620	0.9255	0.9899	0.9611	0.9249	0.9884	0.9462	0.8941	0.9558	0.8792	0.8056	0.9381	0.8380	0.7556
	0.5	0.9906	0.9594	0.9171	0.9898	0.9586	0.9166	0.9902	0.9454	0.8908	0.9517	0.8744	0.8031	0.9392	0.8421	0.7635
	1.0	0.9877	0.9544	0.9138	0.9872	0.9541	0.9125	0.9880	0.9466	0.8969	0.9508	0.8743	0.8012	0.9440	0.8517	0.7745
III	-1.0	0.9930	0.9714	0.9383	0.9912	0.9646	0.9259	0.9886	0.9419	0.8869	0.9598	0.8887	0.8188	0.9220	0.8187	0.7386
	-0.5	0.9800	0.9201	0.8546	0.9946	0.9742	0.9417	0.9857	0.9354	0.8815	0.9581	0.8858	0.8173	0.9092	0.7972	0.7113
	0.0	0.9200	0.7982	0.7052	0.9945	0.9776	0.9563	0.9842	0.9402	0.8898	0.9487	0.8674	0.7937	0.8847	0.7563	0.6686
	0.5	0.9817	0.9213	0.8596	0.9961	0.9752	0.9468	0.9874	0.9392	0.8811	0.9600	0.8858	0.8158	0.9094	0.7934	0.7096
	1.0	0.9934	0.9696	0.9282	0.9917	0.9602	0.9220	0.9878	0.9362	0.8777	0.9586	0.8806	0.8079	0.9141	0.8096	0.7327

$f(x)$ が標準正規分布の場合、 $x-\mu \pm 1.0$ において $f^{(2)}(x)=0$ となるので、方法1と方法2の結果は同じになる。

まず、 $n=50$ の時を見ていこう。分布Iで $x-\mu = \pm 1.0$ の時、方法1による信頼区間の性質はあまり悪くないようである。これは、この点ではバイアス項の影響は非常に小さく、さらに正規近似が比較的正確に行われているからであると思われる。しかし、バイアスの影響が大きくなる $x-\mu = 0$ の点においては、方法1による信頼区間は、信頼係数よりかなり低い割合でしか真の値を含んでいない。逆に、方法2による信頼区間は、信頼係数より高い割合の真値を含んでおり、方法1よりは若干正確なようである。また、方法3による信頼係数は方法2のものよりも改善されているようである。この結果は Hall (1991, 1992a, 1992b) によって示されたブートストラップ法の有効性とも整合的である。

また分布IIの場合は方法1, 2, 3の結果はほぼ同等である。分布IIIの場合は、分布Iの場合と同様にバイアスの影響が大きくなっているようであり、方法2を用いることにより信頼区間の性質はかなり改善されている。しかし、 $x-\mu = 0$ の場合に方法3を用いた結果は方法2のものよりも悪く、ブートストラップ法による近似があまりうまく行われていないようである。

次に、方法4, 5を用いた場合の結果を見ると、どの場合も方法1よりも正確でないようである。これは、方法2, 3の場合と比較すると、バイアスの推定があまり正確に行われていないことが原因のようである。方法2, 3が実行可能でないことを考えると、単純な方法1の結果がかなり良いといえる。

$n=200$ の場合をみると、方法1, 2, 3の結果は、 $n=50$ の時と比べて全体に改善されているようである。しかし、分布III $x-\mu = 0$ の場合、方法1による信頼区間は依然としてあまり正確でない。このケースで $n=20$ の時には、方法3による信頼区間もあまり正確ではなかった。ところが、 $n=200$ の場合にはかなり正確になっている。全体的に見ても、方法3を用いた信頼区間はかなり正確であるといえる。このことから、標本の大きさがある程度大きければ、ブートストラップ法による分布の近似はかなり良いといえる。しかし、方法4, 5による信頼区間は、 $n=200$ の場合でもあまり正確ではない。

以上の結果から、バイアス項を除去して考えれば、ブートストラップ法による分布の近似はかなり良好であると言える。また、方法3, 4の結果が正確でないのは、バイアス項の推定があまり正確ではないためではないかと思われる。方法2, 3が実行可能でないことを考えると、バンド幅をクロスバリデーションで選んだ場合、方法1のように単純にバイアス項を無視して信頼区間を求めるのが現実的であるようである。

6 お わ り に

本稿では、バンド幅をクロス・バリデーションによって選択した場合のカーネル密度推定に基づく信頼区間をいくつか考え、その小標本におけるパフォーマンスをシミュレーションにより分析した。得られた結果から、バイアス項を無視して正規近似により信頼区間を求める方法は、現実的にあまり悪くない事が分かった。また、標本の大きさがある程度大きい場合には、バイアスを除去した統計量の分布はブートストラップ法によりかなり正確に近似できる事が分かった。しかし、バイアスを推定値で置き換える事によって得られる信頼区間の精度はかなり低い。

この結果より、カーネル密度推定に基づく信頼区間をより正確なものにするには、

1. バイアス項より正確な²⁾推定値を求める
2. バイアス項の存在を意識することなく行えるようなブートストラップ法を用いる³⁾

等の方法を用いる必要があると言えるだろう。

注

- 1) (23)で与えられる統計量は漸近的に標準正規分布に従い、平均や分散といった未知のパラメータに依存しない分布になる。このように、未知パラメータを含まない分布を Pivot と呼ぶ。ブートストラップ法を用いる場合、一般に、漸近的に Pivot になる統計量の分布を近似する方が正確な近似が行える (Beran (1988), Hall (1992a) を参照)。実際、Hall (1991, 1992a, 1992b) は (21) を用いて (23) の分布を近似する事の有効性を示している。
- 2) 本稿においては、バイアス項に含まれる $f^{(2)}(x)$ の推定値 $\hat{f}^{(2)}(x)$ を求める際に、 $\hat{f}(x)$ で用いられたものと同じバンド幅を用いた。しかし、特定の条件の下で、 $f(x)$ の 2 回の導関数を推定する場合に MISE を最小化するバンド幅は $h \propto n^{-1/9}$ であることが示される。また、Härdle, Marron and Wand (1990) は導関数の推定に対するクロスバリデーションを提案し、その有効性を示している。
- 3) Hall (1992b) はバンド幅 h を通常より小さくとることにより、バイアス項が 0 に近付くように設定してブートストラップ法を行うことを提案している。彼のシミュレーションによれば、この方法はバイアス項を推定値で置き換えるよりも有効な方法であると考えられるが、 h をどの程度小さく取れば良いかという問題点がある。さらに、Hall (1992a) では、ブートストラップ法におけるリサンプリングを行う際に、単純に $X_1, X_2, \dots, X_n$ から無作為抽出するのではなく、 $\hat{f}(x)$ を確率密度関数として持つ分布から標本抽出する事が提案されている。また、Hall and Owen (1993) は経験尤度を用いて $f(x)$ の信頼区間を求める方法を提案している。Chen (1996) のシミュレーション分析によれば、経験尤度を用いた信頼区間はかなり正確である。したがって、経験尤度を用いた信頼区間に Brown and Newey (2002) のようなブートストラップ法を応用すれば、信頼区間はさらに改善される可能性がある。しかしながら、経験尤度に基づく信頼区間を用いる場合も、バイアス項の影響を小さくするために h を h_{opt} よりも小さく取る必要があり、どの程

度小さく取ればよいかは定かではないという問題点がある。

参 考 文 献

- Beran, R., 1988, "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements," *Journal of the American Statistical Association*, 83, 687-697.
- Bhattacharya, P., 1967, "Estimation of a Probability Density Function and Its Derivatives," *Sankhyā, Series B, Indian Journal of Statistics*, 29, 373-382.
- Brown, B., and W. Newey, 2002, "GMM, Efficient Bootstrapping, and Improved Inference," *Journal of Business and Economic Statistics*, 20, 507-517.
- Chen, S. X., 1996, "Empirical Likelihood Confidence Intervals for Nonparametric Density Estimation," *Biometrika*, 83, 329-341.
- Efron, B., 1979, "Bootstrap Methods: Another Look At the Jackknife," *Annals of Statistics*, 7, 1-26.
- Hall, P., 1983, "Large Sample Optimality of Least Squares Cross-Validation in Density Estimation," *The Annals of Statistics*, 11, 1156-1174.
- Hall, P., 1991, "Edgeworth Expansions for Nonparametric Density Estimators, with Applications," *Statistics*, 22, 215-232.
- Hall, P., 1992a, *The Bootstrap and Edgeworth Expansion*, Springer-Verlag Inc.
- Hall, P., 1992b, "Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density," *The Annals of Statistics*, 20, 675-694.
- Hall, P., and A. B. Owen, 1993, "Empirical Likelihood Confidence Bands in Density Estimation," *Journal of Computational and Graphical Statistics*, 2, 273-289.
- Härdle, W., P. Hall, and J. S. Marron, 1988, "How Far Are Automatically Chosen Regression Smoothing Parameters from Their Optimum?," *Journal of the American Statistical Association*, 83, 86-95.
- Härdle, W., J. S. Marron, and M. P. Wand, 1990, "Bandwidth Choice for Density Derivatives," *Journal of the Royal Statistical Society, Series B, Methodological*, 52, 223-232.
- Li, Q., and J. S. Racine, 2007, *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.
- Rosenblatt, M., 1956, "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, 27, 832-837.
- Schuster, E. F., 1969, "Estimation of a Probability Density Function and Its Derivatives," *The Annals of Mathematical Statistics*, 40, 1187-1195.
- Stone, C. J., 1984, "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *The Annals of Statistics*, 12, 1285-1297.