局所線形カーネル関数による ノンパラメトリック確率密度関数の推定

岐阜大学工学研究科 山 本 けい子

岐阜大学工学部 寒河江 雅 彦

要 旨 近年、ノンパラメトリックな密度関数の推定法の中でもカーネル法が大きく注目されている。カーネル法は1つのデータ点にある関数(カーネル関数)を当てはめ、これをすべてのデータ点について行い、配置された関数を重ね合わせるものである。その結果としてカーネル法は、なめらかな推定値が得られるという利点を持っている。但し、計算量に着目すると、データ量が増えるにつれて計算が非常に大変になる推定法であることがわかる。それはデータの各点に関数を配置することから、データ量が多くなるとそれに比例して関数を合成しなければならないためである。計算機の速度や精度の向上に伴い、改善されている問題ではあるが、大規模なデータに対しては依然として計算効率は重要な問題として位置付けられる。そこで、我々は、カーネル関数を計算の容易な線形関数を用いて効率的に構築することを考え、漸近的な MISE の面で精度の良い、台形カーネル関数を提案する。これによって、カーネル法においてカーネル関数の計算を直接的に減らすことが可能となる。また、既存のカーネル関数との比較により、効率的に推定値を得ることができるということを示す。

1. はじめに

ノンパラメトリックな確率密度関数の推定法であるカーネル法について効果的な計算手段を提案する。現在、ノンパラメトリックな推定の分野においてカーネル法は理論的にも研究が盛んであり、実用的にも広く利用されている方法である。カーネル法の理論的な性質は、例えば Silverman (1982), Scott (1992), Wand and Jones (1995) で述べられている。カーネル法は、基礎となる特定の関数(カーネル関数)をすべてのデータ点を中心に構成し、全区間でそれらの関数を重ね合わせることによってなめらかな推定結果を得るものである。しかし、この方法はカーネル関数の形に大きく依存する推定であるため、一般的な連続関数をカーネル関数に選択した場合には、計算効率が悪いという問題が生じる。また、カーネル推定値の正確な計算は線形以外の関数をカーネル関数に用いた時、膨大な量となる。そこで、Jones (1989)では実際に推定する場合、いくつかの格子点(grid point)でのみ推定値を厳密に計算し、その値を区分的定数あるいは区分的線形関数である直線で内挿するといった離散化を行っている。つまり、カーネル法で推定した後で内挿するといった事後処理を施しているのである。これらの漸近的な性質は Jones (1989)で詳述され

ている。

我々は、連続なカーネル関数での計算量の問題を回避し簡潔に表現できる局所線形カーネル関数として台形カーネルを提案する。2.1 節では、カーネル法の定義と理論的な性質について簡単に述べ、2.2 節では台形カーネルを定義し、その漸近的な性質を述べる。既存のカーネル関数との理論的な比較は 2.3 節で行う。3 節では数値例として推定誤差や計算時間の比較、いくつかの推定例におけるグラフィカルな結果を示す。計算面に焦点を限ったカーネル法とその周辺の話題について 4 節で議論する。

2. カーネル密度推定

2.1. カーネル法の定義と MISE 基準

未知の確率密度関数 f を推定することを考える。 f からの n 個のランダムサンプルを $X_1, X_2, \dots, X_n$ とし、 h をバンド幅とするカーネル関数を K_h とすると、カーネル密度推定は次式で定義される：

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) \quad (1)$$

但し、 $K_h(t) = h^{-1} K(t/h)$ である。ここでは一般の確率密度関数の条件として非負性、積分値は 1、対称性を満たすような 2 次のカーネル関数について議論する。つまり、

$$\begin{aligned} K(t) &> 0, \quad \int K(t) dt = 1, \\ \int t K(t) dt &= 0, \quad \int t^2 K(t) dt \equiv \mu_2(K) > 0 \end{aligned} \quad (2)$$

である。主な 2 次のカーネル関数には、正規(Normal)、Biweight、Triweight、Epanechnikov、三角(Triangular)、一様(Uniform)などがあり、三角と一様カーネルは局所線形関数に属する。上記の関数を表 1 に示す。

表 1. カーネル関数(Scott (1992) による)

Kernel	Equation
Epanechnikov	$\frac{3}{4}(1-t^2)_+$
Biweight	$\frac{15}{16}(1-t^2)_+^2$
Triweight	$\frac{35}{32}(1-t^2)_+^3$
Normal	$N(0, 1)$
Triangular	$(1- t)_+$
Uniform	$U(-1, 1)$
Cosine arch	$\frac{\pi}{4} \cos \frac{\pi}{2} t$

注) 正規カーネルを除くすべてのカーネルのサポートは $[-1, 1]$ である。

また、この 2 次クラスの推定において $n \rightarrow \infty$ での漸近的な性質を計る基準として平均積分二乗誤差(MISE: Mean Integrated Squared Error)がある。MISE は Bias の二乗積分と Variance の積分の和で表されることから次式で定義されている；

$$\begin{aligned}\text{MISE}[\hat{f}(\cdot; h)] &= \int E[\hat{f}(t; h) - f(t)]^2 dt, \\ &= \text{ISB}[\hat{f}(\cdot; h)] + \text{IV}[\hat{f}(\cdot; h)].\end{aligned}$$

但し,

$$\begin{aligned}\text{ISB}[\hat{f}(\cdot; h)] &= \int \text{Bias}[\hat{f}(t; h)]^2 dt, \\ \text{IV}[\hat{f}(\cdot; h)] &= \int \text{Var}[\hat{f}(t; h)] dt\end{aligned}$$

である. 漸近的な MISE を AMISE とすると, 次式で与えられる:

$$\text{AMISE}[\hat{f}(\cdot; h)] = \frac{1}{4}h^4\mu_2(K)^2R(f'') + \frac{1}{nh}R(K). \quad (3)$$

但し, $R(g) = \int g(x)^2 dx$ である. AMISE を最小化するようなバンド幅を最適なバンド幅 h^* として, 最適な AMISE* は次のように $O(n^{-\frac{4}{5}})$ となる:

$$h^* = \left[\frac{R(K)}{\mu_2(K)^2 R(f'') n} \right]^{\frac{1}{5}}, \quad (4)$$

$$\text{AMISE}^*[\hat{f}(\cdot; h)] = \frac{5}{4}\{\mu_2(K)^2 R(K)^4 R(f'')\}^{\frac{1}{5}} n^{-\frac{4}{5}}. \quad (5)$$

2.2. 台形カーネル

前節で示したようにカーネル推定の効率は, カーネル関数の形状や性質に依存して決定される. カーネル関数として容易に構築できないような関数を用いれば, それだけ計算効率は悪くなる. そこで, 計算量を少なく, かつ効率の良い結果を得るために, カーネル関数を直線の組み合わせである局所線形関数を用いて構築することを考える. このクラスは, 一様カーネル, 三角カーネルが漸近的な性質も含め既に議論されている. このことから, これらのカーネル関数を応用して精度のよい推定を提案する. 一様カーネルは 1 本の直線, 三角カーネルは 2 本の直線で表現されることから, さらに拡張して 3 本の直線で構成する. つまり, 台形カーネルである.

次に台形カーネルの理論的な性質を挙げる. 2 次のカーネル関数の性質 (2) 式を満たし, MISE を最適にするように構築された台形(Trapezoid)カーネルを $K_{Tra}(t)$ と表し(これ以降, 他のカーネ

図 1. 台形カーネル

ル関数については添字に頭 3 英字を使って表すことにする), 次の式で定義する;

$$K_{Tra}(t) = \begin{cases} \frac{9}{8}(1+t) & t \in [-1, -\frac{1}{3}] \\ \frac{3}{4} & t \in [-\frac{1}{3}, \frac{1}{3}] \\ \frac{9}{8}(1-t) & t \in [\frac{1}{3}, 1] \end{cases}$$

評価基準となる最適な AMISE は (5) 式の定義を使って

$$AMISE^*[\hat{f}_{Tra}(\cdot; h)] = 0.437176R(f'')^{\frac{1}{5}}n^{-\frac{4}{5}}$$

となる.

2.3. 効率性

(2) 式の条件を満たすような 2 次のカーネル密度推定は, (5) 式からわかるように最適な AMISE は $O(n^{-\frac{4}{5}})$ である. このクラスの推定値を比較する場合, AMISE のオーダーが $n^{-4/5}$ で共通なことから, n に依存しない定数項をできるだけ小さくすることが理論的な改良につながる. (5) 式における $R(f'')$ は, 未知であるが, ある有限の値をとる定数と考えることができる. よって, AMISE の効率は $R(f'')$ と n を除いたカーネル関数 K のみに依存する. このことから, カーネル関数は, 推定効率に大きな影響を与えることがわかる.

2.1 に挙げた 2 次のカーネル関数の中で, 理論的に最も効率がよいとされるものは $[-1, 1]$ の区間において Epanechnikov カーネル (Epanechnikov, 1969) と呼ばれ, 次式で定義される:

$$K_{Epa}(t) = \frac{3}{4}(1-t^2) \quad t \in [-1, 1]. \quad (6)$$

Epanechnikov カーネルを用いたカーネル推定値を厳密に計算するためには, Jones (1989) で指摘されているように連続なカーネル関数による計算量の増加が生じる.

一般にカーネル推定の漸近的な性質は (3) 式によって表されるが, ここではカーネル関数 K のもつ効果を調べるために Wand and Jones (1995) で述べられている尺度

$$C(K) = \{R(K)^4 \mu_2(K)^2\}^{\frac{1}{5}}$$

を使って他のカーネル関数との比較を行う.

台形カーネルの場合の $C(K_{Tra})$ は, $R(K_{Tra}) = \frac{5}{8}$, $\mu_2(K_{Tra}) = \frac{5}{27}$ より,

$$C(K_{Tra}) = 0.349741,$$

また, 最適なカーネル関数として知られている Epanechnikov カーネルを用いた場合の $C(K)$ を $C(K_{Epa})$ と定義すると

$$C(K_{Epa}) = 0.349086$$

である.

次に, 与えられた f に対して, 同じ最小の AMISE* を得るために必要なサンプルサイズの割合を調べるために, 表 1 に挙げたさまざまなカーネル関数 K と Epanechnikov カーネルとの比率

表 2. Epanechnikov カーネルおよび台形カーネルとの比率

Kernel	$\{C(K_{Epa})/C(K)\}^{5/4}$	$\{C(K_{Tra})/C(K)\}^{5/4}$
Epanechnikov	1.0000	1.0023
Cosine arch	0.9994	1.0018
Trapezoid	0.9977*	1.0000*
Biweight	0.9939	0.9962
Triweight	0.9867	0.9890
Triangular	0.9859	0.9882
Normal	0.9512	0.9534
Uniform	0.9295	0.9317

注) * は台形カーネルの効率を示す

$\{C(K_{Epa})/C(K)\}^{5/4}$ および台形カーネルとの比率 $\{C(K_{Tra})/C(K)\}^{5/4}$ を表 2 に示す。

表 2 における値は、例えば、1000 個の標本でカーネル推定の中で最もよく利用されている正規カーネルを用いて推定を行った場合、Epanechnikov カーネルでは約 951 個の標本で同じ AMISE* を得ることができることを示している。一方、我々が提案している台形カーネルでは、正規カーネル 1000 個の標本に対して約 953 個の標本で推定でき、Epanechnikov カーネルを使用した場合と大きな差はないことが分かる。

表 2 に示したカーネル関数の中では Cosine arch が Epanechnikov に最も近く、次いで台形カーネルがよい結果となった。カーネル推定は標本数だけカーネル関数を重ね合わせなければならぬことに着目すると、正規カーネルよりも台形カーネルを使用した方が計算量が少なく、かつ性能よく推定できるという結果を示している。

3. 数 値 例

本論文では、実用面を考慮して、効率的かつ容易に推定できるカーネル関数の構築を試みた。2 次のカーネル関数の中で理論的に最も効率がよい関数は Epanechnikov カーネルであるが、構築の容易さや視覚的な意味での欠点から、実際の推定において用いられることは稀である。また、さまざまな推定に用いられているソフトウェアにおいても基底関数として Epanechnikov カーネルがデフォルトで組み込まれていることは少ない。このような理由から、通常、組み込み関数として S-plus 上で定義されている正規カーネルと三角カーネルを用いて MISE と計算時間の比較を行う。

台形カーネルについては 4 節で後述するように基底となる三角カーネルを 2 つ組み合わせて構築できるということから、新たに関数を定義する必要がなく汎用性に優れている。ここでは、数値実験により台形カーネルの有効性を示す。

3.1. MISE

実際の推定における台形カーネルの推定誤差 MISE をヒストグラム、正規カーネルと共に求め、推定効率を調べる。

表 3 は、標準正規分布からの標本サイズを n とし、 n を 25, 50, 100, 1000 と変化させた時の MISE の値(実験回数 30 回の平均値)を表す。但し、それぞれの推定で使用したバンド幅は MISE

表 3. MISE 比較表

n	台形カーネル	正規カーネル	ヒストグラム
25	0.0116762	0.0124345	0.0407049
50	0.0085710	0.0088928	0.0250941
100	0.0044379	0.0047221	0.0174407
1000	0.0007431	0.0007974	0.0040027

注) 実験回数 30 回の平均値

を最小化する最適なバンド幅を h^* を用いた。

表 3 より、標本サイズが大きくなるにつれて MISE は小さくなり、どの標本サイズにおいても台形カーネルの MISE は正規カーネルやヒストグラムの MISE と比べて小さい値であることがわかる。この結果は、有限サンプルにおける台形カーネルの効率のよさを示すものであり、表 2 に示す漸近的な場合と同様の結果となった。

3.2. 計算時間

台形カーネルと正規カーネルを用いた時の計算時間を記載する。表 4 は標準正規分布からの n 個の標本を用いて、台形カーネルと正規カーネルに基づいたカーネル密度推定に要した計算時間 (500 回の平均値) を表す。ここでは、一般的によく用いられている統計ソフト S-plus に組み込まれているカーネル推定関数 ksmooth を使用し、台形カーネルについては三角の基底関数の組み合わせを用いて構築した結果である。(比較の際に使用した計算機の CPU は celeron450 MHz である。) また、Epanechnikov カーネルに関しては、S-plus の組み込み関数として定義されていないため、台形カーネルや正規カーネルと同等の定量的な比較ができないことから本論文では記載しない。

表 4. 推定に要した計算時間 [秒] (500 回平均)

n	台形カーネル	正規カーネル
10	0.0051	0.0051
100	0.0123	0.0156
500	0.1583	0.2257
1000	0.6204	0.9122
5000	15.056	21.395
10000	63.075	87.784

表 4 からわかるように、どの標本サイズにおいても、通常よく使用されている正規カーネルよりも台形カーネルの方が推定に要した時間は少なく効率的である。

3.3. 推定例

3.1 節と 3.2 節において台形カーネルを用いたカーネル密度推定の数値的な有効性を MISE および計算時間によって示した。本節では、正規、一様、三角、台形カーネルで同様の推定を行い、その推定値を図示することで視覚的になめらかさを比較する。

標準正規分布 f に従うデータに基づいた密度関数の推定を行う。但し、プロットする推定区間を $[-3, 3]$ とし、標本サイズは $n = 20$ および $n = 100$ 、バンド幅 h の値は、それぞれの推定に最小 AMISE を達成する漸近的に最適な h^* 、および一般的に使用されている Sturges の階級数による h_s 、統計ソフト S-plus で使用されている Doane の階級数 (Becker et al, 1988) から導いた h_d を用いて推定する。理論的に最適な h^* は (4) 式を用いて計算される。また、Sturges の階級数、Doane の階級数 k_s, k_d はそれぞれ次式で表され、推定区間を次式の値で分割したものがバンド幅となる。

$$k_s = \log_2(n) + 1$$

$$k_d = \log_2 \left[\frac{(n+1)(n+3)\hat{\mu}_3^2}{6(n-2)\hat{\mu}_2^3} + 1 \right] + 0.5 + \log_2(n) + 2$$

ここで、 k_d における $\hat{\mu}_2, \hat{\mu}_3$ はそれぞれ平均値まわりの標本 2 次モーメントと標本 3 次モーメントを表す。数値例の比較の際、Sturges の階級数に skewness (歪み) を考慮した Doane の階級数についても調べた。

それぞれのバンド幅を表 5 に示す。但し、 h_d は、本数値例に用いた標本より導出した値であり、標本サイズ $n = 20$ に対して $\hat{\mu}_2 = 1.003, \hat{\mu}_3 = 0.295$ 、 $n = 100$ に対して $\hat{\mu}_2 = 0.737, \hat{\mu}_3 = -0.046$ である。

表 5. バンド幅 h の値

h	$n = 20$	$n = 100$
h^*	1.339	0.971
h_s	1.127	0.785
h_d	0.857	0.667

注) h^* は台形カーネルのバンド幅を示す

これらの推定結果をもとに台形カーネルの性質を視覚的な観点から述べる。標本数 $n = 20, n = 100$ について、図 2.1-4 および図 3.1-4 はバンド幅として h^* を用いた推定値、図 4.1-4 および図 5.1-4 は h_s を用いた推定値、図 6.1-4 および図 7.1-4 は h_d を用いた推定値を図示したものである。但し、 $n = 20$ に関しては基底となる関数についても図示する。

すべての図において正規カーネルはなめらかであり、一様カーネルはなめらかでない階段状の推定値である。これは、連続関数と線形関数の違いに起因するものであるが、一様カーネルと同じ線形関数である三角や台形カーネルの場合は、正規カーネルに近いなめらかさを得ている。また、 $n = 100$ のようにサンプルサイズが大きくなると、 $n = 20$ と比べてさらになめらかさが増し、正規カーネルと視覚的に変わらない結果であることがわかる。このことは、理論的に効率のよい台形カーネルを用いる動機を与えるものである。

台形カーネルはその効率のよさのほかに関数自体が尖りが少なくなめらかであるため、視覚的によい推定値を得ることができると考えられる。一方、三角カーネルはデータのばらつきとバンド幅 h の値によってなめらかさに差が出ると思われる。バンド幅 h はスムージングパラメータとも言われ、大きすぎるとなめらかさが増しデータの特徴を損ない、小さくしすぎると個々のデータを強調し全体の傾向を把握できなくなる。そのため、今回使用した例は標準正規分布からの乱

それぞれに最適な h^* を用いた推定結果 $n = 20$

図 2.1 正規カーネルを用いた推定

図 2.2 一様カーネルを用いた推定

図 2.3 三角カーネルを用いた推定

図 2.4 台形カーネルを用いた推定

数であったが、バンド幅 h_d を用いた場合は、 h^* や h_s に比べて h_d の値が小さくなるため(表 5)、推定された値の形状は平滑化不足のように見える。

それぞれに最適な h^* を用いた推定結果 $n = 100$

図 3.1 正規カーネルを用いた推定

図 3.2 一様カーネルを用いた推定

図 3.3 三角カーネルを用いた推定

図 3.4 台形カーネルを用いた推定

4. 議 論

ノンパラメトリックな確率密度関数の推定で最もよく利用されているカーネル推定の計算面に焦点を絞った議論を行ってきた。計算量を減らすための工夫としては次の3つが挙げられる。

- 1) カーネル関数として線形関数を組み合わせて用いること。

Sturges' rule による h_s を用いた推定結果 $n = 20$

図 4.1 正規カーネルを用いた推定

図 4.2 一様カーネルを用いた推定

図 4.3 三角カーネルを用いた推定

図 4.4 台形カーネルを用いた推定

我々の提案する台形カーネルもこれに属する。

- 2) ビン化カーネル推定〈データをある区間(ビン)に区切り, 各ビンに入る度数を用いたカーネル推定〉を用いること。

これはビンの数のみカーネル関数を配置すればよく, データ数に比べて, かなり少ない数のカーネル関数を用いた推定値を得ることができる。

Sturges' rule による h_s を用いた推定結果 $n = 100$

図 5.1 正規カーネルを用いた推定

図 5.2 一様カーネルを用いた推定

図 5.3 三角カーネルを用いた推定

図 5.4 台形カーネルを用いた推定

3) カーネル関数のサポートを有限区間でとること.

たとえば, 正規カーネル関数を用いた場合, サポートは無限であるため各データ点においてすべてのデータに配置したカーネル関数の影響を受ける. しかし, 有限サポートであれば, 隣接するいくつかのデータのカーネル関数を重ね合わせればよい.

Doane の方法による h_d を用いた推定結果 $n = 20$

図 6.1 正規カーネルを用いた推定

図 6.2 一様カーネルを用いた推定

図 6.3 三角カーネルを用いた推定

図 6.4 台形カーネルを用いた推定

4.1. 台形カーネル関数の構築

2.2 節で定義された台形カーネルは、図 8 に示すように 2 つの三角カーネルの組合せとみなすことができる。これは、1 つのデータに対してシフトした 2 つの単純な三角形の基底関数を重ね合わせることによって構築することができ、計算が非常に容易であるという特徴をもつ。例えば、

Doane の方法による h_d を用いた推定結果 $n = 100$

図 7.1 正規カーネルを用いた推定

図 7.2 一様カーネルを用いた推定

図 7.3 三角カーネルを用いた推定

図 7.4 台形カーネルを用いた推定

3.2 節の例で用いたように、S-plus では三角カーネルが基底関数として組み込まれている。この三角カーネルを組合わせて台形カーネルを構築すれば、新たに台形カーネルを定義する必要がなく効率良く生成できる。

図 8. 台形カーネルの構築

4.2. 多角形カーネル関数への拡張

台形カーネルは一樣あるいは三角といった区分的線形関数の拡張であることから、台形をさらに多角形へと区分的線形関数のまま拡張させることが可能である。この多角形を Epanechnikov カーネルを近似するような関数に構築すれば、精度は Epanechnikov カーネルにより近くなる。

また、多角形カーネルにおいても、4.1 節で述べたように単純な三角形の基底関数を複数組み合わせることで得ることができる。しかし、多角形の次数が増えるにつれて基底関数の数も増え、計算量は増加する。例えば図 9 のように基底関数が 3 つの場合、表 2 に対応させると、効率は 0.99889 であるが計算量は台形カーネルの 1.5 倍である。このことを考慮しても、実用上は表 2 に示すように台形カーネルで十分な効率を得ることができるのでこれ以上の議論は避ける。

図 9. 多角形カーネルの構築

4.3. ビン化カーネル密度推定への応用

大規模データに対しカーネル推定を適応する場合、データ量に比例して計算量が増大する。それを軽減するためには、Silverman (1986), Scott and Sheather (1985), Hall and Wand (1996) などで提案されているビン化カーネル法が効果的である。この方法はデータをいくつかのビンに区切り (binning), 各ビンにその度数に比例した高さをもつカーネル関数を配置する。つまり、ビンの数だけのカーネル関数を重ね合わせればよいのである。このビン化カーネル法を用いることでデータ量に依存して増加する計算量を大幅に減らすことが可能である。台形カーネルを含めてここで議論してきたカーネル法をビン化カーネル推定に応用することは容易である。Binning したデータに台形カーネル関数を用いれば、さらなる計算量の減少を導くことができる。

4.4. グラフィックにおける台形カーネルの誤差

通常、正規カーネルなどのカーネル関数を用いた密度推定で得られる推定値を図示する場合、推定区間内のいくつかの点(grid point)での推定値を順次つないで表示される。つなぐ方法は直線であったり、2次関数であったりする。このようにしてプロットされた推定結果は、grid pointでの推定値は正確な値であるが、それ以外の点における推定値は実際の推定値と一致しない。しかし、我々が提案した台形カーネルの場合は、基底となる台形カーネルの節点での推定値を順次直線でつなぐことで grid point 以外の点においても正確な推定値を得ることができる。

以上の議論から、カーネル密度推定において台形カーネル関数を使用することは、理論的にも計算効率の面でも優れていることがわかる。また、視覚的にも正規カーネル推定値とほとんど変わらないなめらかな推定値を与えている。

参 考 文 献

- Becker, R.A. and Chambers, J.M. and Wilks, R.A. (1988): 渋谷政昭・柴田里程 訳, S 言語-データ解析とグラフィックスのためのプログラミング環境-I, II. 共立出版.
- Epanechnikov, V.A. (1969): Nonparametric estimation of a multivariate probability density. *Theory of Probability and Its Applications* **14**, 153-158.
- Hall, P. and Wand, M.P. (1996): On the accuracy of binned kernel density estimators. *Journal of Multivariate Analysis* **47**, 103-122.
- Jones, M.C. (1989): Discretized and interpolated density estimates. *Journal of the American Statistical Association* **84**, 733-741.
- Scott, D.W. and Sheather, S.J. (1985): Kernel density estimation with binned data. *Communications in Statistics - Theory and Methods* **14**, 1353-1359.
- Scott, D.W. (1992): *Multivariate Density Estimation*. Wiley, New York.
- Silverman, B.W. (1982): Kernel density estimation using the fast Fourier transform. *Appl. Statist.* **31**, 93-99.
- Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
- Wand, M.P. and Jones, M.C. (1995): *Kernel Smoothing*. Chapman and Hall, London.

(2001 年 8 月 20 日受付 12 月 14 日最終修正 12 月 19 日採択)

著者連絡先: 〒 501-1112 岐阜市柳戸 1-1
岐阜大学工学部応用情報工学科 寒河江 雅彦

Nonparametric Density Estimate by Piecewise Linear Kernel Function

Keiko Yamamoto¹ and Masahiko Sagae^{1,*}

¹ Faculty of Engineering, Gifu University,
1-1 Yanagido, Gifu-shi, Gifu 501-1193, Japan

Abstract

Kernel density estimation is one of the well-known technique in nonparametric density estimations. Kernel method is constructed by placing a kernel function disposed at each data point and it can obtain smooth estimations. We note that this method has cost of calculation increasing in proportion with the number of data. Its computational burden is relieved by the development of the computer, however, if we choose a continuous probability density function, then the calculation increases more. Therefore we suggest the trapezoid kernel function generated by the piecewise linear function, because we can reduce the cost of calculating the estimate by using trapezoid kernel. We show that it is easy to compute and has the better efficiency compared with other kernel functions. We discuss about the property of trapezoid kernel function focused on the calculational efficiency and show the numerical example, also.

Key words: binned kernel density estimator, MISE, trapezoid kernel function.

* Corresponding author.

(Masahiko Sagae)

Received August 20, 2001; Received in final form December 14, 2001; Accepted December 19, 2001.