区分的線形関数によるノンパラメトリック確率密度関数の推定

岐阜大学大学院工学研究科　山本けい子
岐阜大学工学部　寒河江雅彦

要旨　ノンパラメトリックな推定の最も単純な手法はヒストグラムであり、度数を用いたビン型推定法に属する。ヒストグラムを用いて推定された密度関数は滑らかでなく、理論的な性質も他の推定法と比べてよくない。そこで、ヒストグラムを改良する既存の推定モデルに関して統括的に示し、その理論的な性質を述べる。また、カーネル推定法についても、計算効率を考慮し、台形カーネル関数やビン化カーネル推定法について議論する。

1　ビン型推定法

ヒストグラムはある区間（ビン）に入る度数の高さを持った階段関数である。ヒストグラムの滑らかさと理論的な性能を改良する方法として、Frequency Polygon (Scott,1985), Edge Frequency Polygon (Jones 他,1998), Bias optimized Frequency Polygon (Minnotte,1996), Generalized Frequency Polygon(Sagae and Yamamoto,2000) が提案されている。これらの推定モデルは、節点の位置と高さを調整し、節点での 1 次の連続性を満たすような区分的線形関数で構築される。これらのビン型推定モデルは以下の表のように分類される;

表 1: 区分的線形関数を用いたビン型推定モデル

モデル	関数の次数	節点（節点での高さ）	性能
HIST	区分的 0 次	端点（度数）	$0.82548R(f')^{1/3}n^{-2/3}$
FP	区分的 1 次	中点（度数）	$0.52797R(f'')^{1/5}n^{-4/5}$
EFP	区分的 1 次	端点（隣接ビンの高さの平均）	$0.47232R(f'')^{1/5}n^{-4/5}$
BFP	区分的 1 次	中点（各ビンの面積 [推定前＝推定後]）	$0.50457R(f'')^{1/5}n^{-4/5}$
GFP4	区分的 1 次	中点, 端点（隣接ビンの線形結合）	$0.30175R(f'')^{1/5}n^{-4/5}$

HIST：ヒストグラム, FP：Frequency Polygon, EFP：Edge Frequency Polygon
BFP：Bias-optimized Frequency Polygon, GFP4：4-th Generalized Frequency Polygon

但し、$R(g) \equiv \int g(x)^2 dx$ であり、性能は $AMISE = AISB + AIV$(漸近平均積分二乗誤差）を表す。また、GFP4 の性能は 4 次の GFP の最適な重み $-\frac{1}{8}, \frac{5}{8}, \frac{5}{8}, -\frac{1}{8}$ を用いた時の値である。

次に区分的線形関数を用いたクラスの Bias および分散に関する定理を示す;

Theorem 1: *Bias* **の下限値** ：

k 次 GFP $\hat{f}_k(x)$ の漸近 2 乗バイアスの下限値は下記のとおり (但し、等号は $k \geq 3$ で達成可能である);

$$AISB \geq \frac{1}{720}\,\delta^4\,R(f'').$$

Theorem 2: 分散の下限値 ：

k が十分大きいと仮定すると、漸近的な IV の下限値;

$$AIV \geq \frac{1}{n\,\delta}\left(\frac{1}{k} + o(k^{-1})\right).$$

2　カーネル推定法

カーネル法における計算効率化 (Yamamoto and Sagae,2001)

カーネル推定法は、データの各点にカーネル関数と呼ばれるある関数を配置し、すべてのデータ点について配置したカーネル関数を重ね合わせて構築される。バンド幅 h を持つカーネル関数を K（但し、$K_h(x) = h^{-1}K(x/h)$）とすると、カーネル推定量は次式で定義される;

$$\hat{f}(x;h) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i) \qquad -\infty < x < \infty. \tag{1}$$

カーネル法の計算量は、1. 基底となるカーネル関数の形状や性質に依存するもの、2. データ数に比例して増加するものがある。それぞれに対して次のような計算効率化が考えられる;

1. カーネル関数に対する効率化：台形カーネル
   計算量を少なく、かつ効率の良い結果を得るために、カーネル関数に区分的線形な関数を用いた台形カーネルを構築する。台形 (Trapezoid) カーネルを $K_{Tra}(t)$ と表し、次式で定義する;

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

   最もよく用いられている正規カーネルと台形カーネルの効率を比較すると、Epanechnikov カーネルと同じ性能を得るために必要とされる標本数は Epanechnikov が 1000 個の時、台形カーネルは 1003 個、正規カーネルは 1049 個である。
2. データ数に対する効率化：ビン化カーネル推定法 (Silverman(1982), Scott and Sheather(1985))
   ビン化カーネル推定法は、各ビンの度数に比例した高さをもつカーネル関数をビンの数だけ配置して構築する推定法である。データをビンに集約することによってビンの数に対応したカーネル関数の重ね合わせで推定できることから、計算量の大幅な削減につながる。

参考文献

[1] Jones,M.C.(1989),"Discretized and Interpolated Density Estimates",*JASA*.

[2] Jones,M.C. ,Samiuddin,M.,Al-harbey,A.H. and Maatouk,T.A.H. (1998), "The edge frequency polygon",*Biometrika.*

[3] Minnotte,M.C.(1996),"The Bias-Optimized Frequency Polygon",*Computational Statistics.*

[4] Sagae,M. and Yamamoto,K.(2000),"On a generalized class of Frequency Polygon",*ISM cooperative report.*

[5] Scott,D.W.(1985),"Frequency Polygons",*Journal of the American Statistical Association.*

[6] Scott,D.W. and Sheather,S.J.(1985),"Kernel Density Estimation With Binned Data", *Communications in Statistics - Theory and Methods.*

[7] Silverman,B.W.(1982),"Kernel density estimation using the fast Fourier transform",*Appl. Statist.*

[8] Yamamoto,K. and Sagae,M.(2001), "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", 応用統計学,30,3.