円周上のカーネル密度推定量とそのバンド幅選択法の漸近的性質

鶴田　　靖人　　金沢大学大学院人間社会環境研究科
寒河江　雅彦　　金沢大学経済学経営学系

概 要

周期性を持つデータは円周上に台を持つ確率分布を用いてモデリングできる．方向カーネル密度推定量は、このような円周上に定義された密度関数を推定するためのノンパラメトリックな手法である．本稿では方向統計学におけるカーネル密度推定量のいくつかの研究成果を報告する．1 つ目は，巻き込みコーシーカーネル密度推定量の漸近的性質を導いたことである．2 つ目は，高次オーダーカーネル密度推定量と具体的な高次オーダーカーネル密度推定量の構成法を提案し，それらの漸近的性質を求めたことである．3 つ目は，カーネル密度推定量の平滑化パラメータの選択法の理論的性質を議論するために，代表的な選択法である least square cross validation 法と direct plug-in rule 法の収束レートを含むいくつかの漸近的な性質を導出したことである．

1　はじめに

周期性を持つデータは、円周上の角度として値を取る確率変数として扱うことができるので方向データと呼ばれる．このような方向データを分析することを目的とした統計学が方向統計学である．方向統計学では周期性を持つ確率変数 $\Theta \sim f(\theta)$ を扱う．ただし，$\theta \in [-\pi, \pi)$，$f(\theta) = f(\theta + 2\pi)$ とする．標本は $\Theta_1, \ldots, \Theta_n \overset{\text{i.i.d.}}{\sim} f(\theta)$ とする．このとき，$f(\theta)$ の方向カーネル密度推定量は以下の式で定義される：

$$\hat{f}_\kappa(\theta) := \frac{1}{n} \sum_{i=1}^{n} K_\kappa(\theta - \Theta_i),$$

ただし，$K_\kappa(\theta)$ は対称なカーネル関数であり，$\kappa > 0$ は集中度パラメータ (バンド幅の逆数に対応する平滑化パラメータ)．

方向データの例として Stephens (1969) の turtles データとそのカーネル密度推定量を図 1 に示した．図 1 のカーネル密度推定量は turtles データが持つ 2 峰性の性質をうまく表している．方向データは，有限区間の台を持つが，図 1 から分かるように周期性の性質から境界を持たない．ゆえに，方向カーネル密度推定量は実数空間上のカーネル密度推定量で発生する境界バイアスの問題を避けることができる．

本稿は、方向データのモデリングとしてよく用いられる巻き込みコーシー (WC) 分布をカーネル関数とした WC カーネル密度推定量の漸近的性質を与える．数値実験の結果から特定の条件の下では WC カーネル密度推定量は良い性質を持つことが分かっている．

本稿では新しいモーメントを定義し，p 次オーダーカーネル関数のクラスを提案する．このカーネル関数のクラスは，Hall et al.(1987) が与えた 2 次オーダーカーネル関数のクラスを p 次オーダーカーネル関数に拡張したものとなっている．p 次オーダーカーネル密度推定量は平均二乗誤差 (MISE) の収束レートが $O(n^{-2p/(p+1)})$ で漸近正規性を持つ．このカーネル関数は，Jones and Foster (1993) の加法型構成法や Terrell and Scott (1980) の乗法型構成法を用いて，2 次オーダーカーネルから 4 次以上の高次オーダーカーネル密度推定量を構成できる．

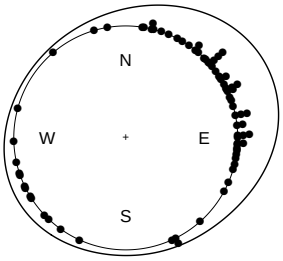

図 1: Stephens (1969) turtles データ (n=76) をプロットしたもの．turtles データは，子亀がふ化したときどの方向を進むかを表している．実線は turtles データのカーネル密度推定量である．

方向統計学での平滑化パラメータ (集中度パラメータ) κ の選択法として， least square cross validation (CV) 法（Hall et al. , 1987) や Direct Plug-in (PI) 法 (Mardio et al. , 2011) などがある．κ の推定についての研究は，数値実験による比較実験が主であり，選択法の理論的な性質はほとんど議論されていない．我々は CV 法と PI 法の漸近的な性質を導出した．CV 法による κ の推定量 $\hat{\kappa}$ の漸近的な収束レートは $O(n^{-1/10})$ であるが，PI 法による $\hat{\kappa}$ の漸近的な収束レートは $O(n^{-5/14})$ となる．この結果は PI 法の方が CV 法よりも優れた選択法であることを示している．

2　先行研究

Marzio et al. (2011) は，sin 型モーメント $\eta_j(K_\kappa) := \int \sin(\theta)^j K_\kappa d\theta$ を用いて sin 型 p 次オーダーカーネル関数を提案した．sin 型 p 次オーダーカーネル関数 K_κ を，

$$\eta_0(K_\kappa) = 1, \quad \eta_j(K_\kappa) = 0,\ 0 < j < p, \quad \eta_p(K_\kappa) \neq 0,$$

満たすことであると定義する．sin 型 p 次オーダーカーネル関数は、多くの対称な円周上の密度関数を含む一般的なクラスであるという特徴を持つ．sin 型 p 次オーダーカーネル密度推定量の MISE は以下の式となる．

$$\text{AMISE} = \frac{\eta_p^2(K_\kappa) R\left(f^{(p)}\right)}{(p!)^2} + \frac{1 + 2\sum_{j=1}^{\infty} \gamma_j^2(\kappa)}{2\pi n}, \tag{2.1}$$

ただし，$R(g) := \int \{g(\theta)\}^2 d\theta$，フーリエ係数 $\gamma_j(K_\kappa) := \mathrm{E}_K[\cos(j\theta)]$．式 (2.1) の第 1 項と第 2 項はそれぞれバイアスの 2 乗と分散に対応している．Di Marzio et al. (2011) は，(2.1) を用いて，代表的な sin 型 2 次オーダーカーネルであるフォン・ミーゼス (VM) カーネル密度推定量の収束レートが $O(n^{-4/5})$ となることを示した．これは，台が実数直線上で与えられる 2 次オーダーカーネル密度推定量の収束レートに一致する．

3　巻き込みコーシーカーネル密度推定量

WC カーネルを以下のように定義する:

$$K_\rho(\theta) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta)}, \quad 0<\rho<1,$$

ただし，ρ は集中度を調節するパラメータとする．WC カーネルの特性関数は以下のように定義できる：

$$\phi_p = \rho^{|p|}.$$

WC カーネルの $\gamma_j(\rho)$は$\gamma_j(\rho) = \rho^j$ となるので，WC カーネルの AMISE は次式で与えられる：

$$\mathrm{AMISE}_{\mathrm{WC}}[\hat{f}_\rho(\cdot)] = \frac{\{1-\rho^2\}^2 R(f'')}{16} + \frac{1}{n\pi(1-\rho^2)}. \tag{3.1}$$

(3.1) について $1-\rho^2 = h$ とおくと，次式が得られる：

$$\mathrm{AMISE}_{\mathrm{WC}}[\hat{f}_h(\cdot)] = \frac{h^2 R(f'')}{16} + \frac{1}{n\pi h}. \tag{3.2}$$

κ^* と同様にして (3.2) を最小にする最適な h^* は次式のようになる：

$$h^* = \left(\frac{8}{\pi R(f'')n}\right)^{1/3}, \quad n > 8(\pi R(f''))^{-1}. \tag{3.3}$$

(3.2)，(3.3) より AMISE の収束レートは $O(n^{-2/3})$ となる．

制約条件 $(0<h<1)$ を設けなければ，n が十分大きくないとき h は 1 を超えうる．実用上は最適な集中度パラメータは (3.2) を最小にする ρ^* とするのが望ましい．ρ^* を以下のように与える：

$$\rho^* = \arg\min_{0<\rho<1}\left\{\mathrm{AMISE}_{\mathrm{WC}}[\hat{f}_\rho(\cdot)]\right\}. \tag{3.4}$$

数値実験の結果から，真の分布が特定の条件を満たす場合に WC カーネルは，VM カーネルよりも MISE の値が小さくなることが分かっている．この数値実験の詳細は当日発表する．

WC カーネルの最適な収束レート $O(n^{-2/3})$ は同じ VM カーネルの収束レート $O(n^{-4/5})$ とは異なる． このことは，Sin 型モーメントは MISE の収束レートに対応したカーネル・モーメントとは言えないことを示唆している．また，Di Marzio et al. (2011) は， sin 型 p 次オーダーカーネルは必ずしもバイアスを修正できるわけではないと指摘した．つまり，sin 型 p 次オーダーカーネルの次数と MISE の収束レートは必ずしも対応しているわけではない．

4　高次オーダーカーネル密度推定量

我々はカーネル関数は主要項 $L(\cdot)$ と基準化定数 $C_\kappa(L)$ を用いて，$K_\kappa(\theta) := C_\kappa^{-1}(L)L(\kappa\{1-\cos(\theta)\})$ とする．また，モーメントを $\mu_l(L) := \int_0^\infty L(r)r^{(l-1)/2}dr$ と定義する．ただし，$r = \kappa\{1-\cos(\theta)\}$ とする．

$K_\kappa(\theta)$ が p 次オーダーカーネル関数であるとは，正の偶数 l に対して，

$$\mu_0(L) \neq 0, \qquad \mu_l(L) = 0, \quad 0 < l < p, \qquad \mu_p(L) \neq 0,$$

を満たすことである．p 次オーダーカーネル密度推定量の MISE は以下の式で表せる．

$$\text{AMISE} = \frac{\mu_p^2(L)R\left(\sum_{t=1}^{p/2} \frac{b_{p,2t}f^{(2t)}(\cdot)}{2t!}\right)}{\mu_0^2(L)\kappa^p} + \frac{d(L)\kappa^{1/2}}{n}, \tag{4.1}$$

ただし，$b_{p,2t}$ と $d(L)$ は定数である．式 (4.1) の第 1 項と第 2 項は，それぞれバイアスの 2 乗と分散に対応している．$\kappa = h^{-2}$ とおけば，(4.1) は台が実数直線上で定義された密度推定量の MISE の形によく似ている．我々が導入した高次オーダーカーネル関数は，次数に対応して最適な MISE の収束レートが $O(n^{-2p/2(p+1)})$ となる．この収束レートは実数直線上で定義された p 次オーダーカーネル密度推定量の収束レートに対応する．

5　高次オーダーカーネルの構成法

$K_{\kappa,[p]}(\theta) := C_\kappa^{-1}(L_{[p]})L_{[p]}(\kappa\{1-\cos(\theta)\})$ は前章で定義した p 次オーダーカーネル関数を表す．

Jones and Foster (1993) に対応する加法型構成法は，L とその導関数の和として次式で表せる：

$$L_{[p]}(r) := \frac{p+1}{p}L_{[p]}(r) + \frac{2}{p}rL'_{[p]}(r),$$

ただし，$r = \kappa\{1-cos(\theta)\}$，$L'_{[p]}(r) = dL'_{[p]}(r)/dr$.　加法型構成法を p 次オーダーカーネルに適用することで，$p+2$ 次オーダーカーネルを新しく生成できる．

Terrell and Scott (1980) に対応する乗法型構成法は，異なるバンド幅を持つ 2 次オーダーカーネル密度推定量 $\hat{f}$ の積として次式で表せる：

$$\hat{f}_\kappa^{[\text{TS}]}(\theta) := \hat{f}_\kappa^{4/3}(\theta)\hat{f}_{\kappa/4}^{-1/3}(\theta),$$

ただし，$\text{bias}_f^2[\hat{f}_\kappa^{[\text{TS}]}(\theta)] = O(\kappa^{-4})$, $\text{Var}_f[\hat{f}_\kappa^{[\text{TS}]}(\theta)] = O(n^{-1}\kappa^{1/2})$ かつ $\text{MISE} = O(n^{-8/9})$ となる．つまり，$\hat{f}_\kappa^{[\text{TS}]}$ は 4 次オーダーカーネル密度推定量となっている．

2 次オーダーカーネル，加法型 (4 次オーダー) と乗法型 (4 次オーダー) を比較した数値実験の結果は当日発表する．

6　平滑化パラメータの選択法

2 次オーダーカーネル密度推定量の (4.1) を最小にする集中度パラメータ κ_* は以下の式で与えられる:

$$\kappa_* = \beta \psi_4^{2/5} n^{2/5}, \tag{6.1}$$

ただし β は定数，$\psi_r := \int_{-\pi}^{\pi} f^{(r)} f(\theta) d\theta$．本稿では κ_* の推定法として CV 法と PI 法を挙げている．これらの推定法の定義と漸近的性質について述べる．

6.1　Least square cross validation 法

CV 法とは，

$$\mathrm{CV}(\kappa) = R(\hat{f}) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(\Theta_i),$$

を最小にする推定量 $\hat{\kappa}_{\mathrm{CV}}$ を求める手法である．$\hat{\kappa}_{\mathrm{CV}}$ の漸近的な性質を示す：

$$\begin{aligned} \hat{\kappa}_{\mathrm{CV}}/\kappa_* &\xrightarrow{p} 1, \\ n^{1/10}(\hat{\kappa}_{\mathrm{CV}}/\kappa_* - 1) &\xrightarrow{d} \mathrm{N}(0, \sigma_{\mathrm{CV}}^2). \end{aligned} \tag{6.2}$$

(6.2) の収束レートは実数直線上の CV 法で推定したバンド幅推定量 $\hat{h}_{\mathrm{CV}}$ の収束レートの対応している．

7　Direct plug-in (PI) 法

PI 法とは，ψ_4 をカーネル密度推定法で推定し κ_* を推定する方法である．カーネル密度推定量 $\hat{\psi}_4(g)$ を定義する：

$$\hat{\psi}_4(g) := n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} T_g^{(r)}(\Theta_i - \Theta_j),$$

ただし，$T_g(\theta) = C_g^{-1}(S) S(g\{1 - cos(\theta)\})$ は p 次オーダーカーネル関数，g は集中度パラメータ.

$\hat{\psi}_4(g)$ のバイアスを最小にする g を g_* と表す：

$$g_* := c n^{2/(p+5)},$$

ただし，c は定数．このとき，$\hat{\psi}_4(g)$ の平均二乗誤差 (MSE) は，

$$\inf_{g>0} \mathrm{MSE}[\hat{\psi}_4(g_*)] = \begin{cases} O(n^{-(2p+1)/(p+5)}) & p < 4, \\ O(n^{-1}) & p \geq 4, \end{cases}$$

となる．したがって，4 次以上の高次オーダーカーネルを用いたとき，$\hat{\psi}_4(g)$ はパラメトリックな収束レートである $O(n^{-1})$ を達成する．

また，PI 推定量は，

$$\hat{\kappa}_{\mathrm{PI}} := \beta\hat{\psi}_4(g_*)^{2/5}n^{2/5}, \tag{7.1}$$

で与えられる．$\hat{\kappa}_{\mathrm{PI}}$ は T_g が 2 次オーダーカーネル関数のとき，以下のような漸近正規性を持つ．

$$n^{5/14}(\hat{\kappa}_{\mathrm{PI}}/\kappa_* - 1) \xrightarrow{d} \mathrm{N}(0, \sigma^2_{\mathrm{PI}}). \tag{7.2}$$

(7.2) の収束レートは実数直線上の PI 法で推定したバンド幅推定量 $\hat{h}_{\mathrm{PI}}$ の収束レートに対応する．CV 法と PI 法の収束レートを比較すると PI 法の方が収束スピードが速いことが分かる．一般的に PI 法は，CV 法に比べて分散が小さくより安定した推定量を与える利点を持つ．

CV 法と PI 法を比較するための数値実験を当日に報告する．

参考文献

[1] Di Marzio, M., Panzera, A. and Taylor, C. C. (2011). *Journal of Statistical Planning and Inference* **141**, 2156-2173.

[2] Hall, P., Watson, G. S. and Cabrera, J. (1987). *Biometrika*, **74**, 751-762.

[3] Jones, M. C. and Foster, P.J. (1993). *Journal of Nonparametric Statistics* **3**, 81-94.

[4] Terrell, G. R. and Scott, D. W. (1980). *The Annals of Statistics* **8**, 1160-1163.

[5] Tsuruta, Y., and Sagae, M. , 2016. *Asymptotic Property of Wrapped Cauchy Kernel Density Estimation* on the Circle. Discussion Paper, No.28 in school of Economics, Kanazawa University.

[6] Tsuruta, Y., and Sagae, M. , 2016. *Higher order kernel density estimation on the circle* Discussion Paper, No.30 in school of Economics, Kanazawa University.

[7] Scott, D. W. and Terrell G. R. *Jounal of the American Statistical Association*, **82**, 1131-1146.

[8] Sheather S. J. and Jones M. C. *Jounal of the Royal Statistical Society. Series B*, **53**, 683-690.

[9] Stephens, M. A. (1969). Techniques for directional data (No. TR-150). *STANFORD UNIV CA DEPT OF STATISTICS.*