修士論文 要旨

カーネル指数分布族による非負値データの密度関数推定

数理情報学専攻 48176228　渡邉 拓

指導教員　清 智也 准教授

1　はじめに

確率変数の分布をサンプルから推定する問題は統計における基本的な課題であり，連続値のデータに対しては分布の情報を豊富に含む密度関数が重要となる．密度推定の代表的な手法である最尤法は，正規化定数の計算が困難で実用的でない場合が多い．この困難を回避する手法としてスコアマッチングが知られている．近年，無限次元の指数分布族に適用する提案がなされた．

本研究では非負値データの密度関数推定を扱う．既存手法で用いられた分布族を拡張することによって柔軟なモデリングを可能にする．さらに，目的関数の修正によって既存手法における境界条件を緩和することで，より広い状況で使用できる手法を提案する．また，提案した手法と代表的なノンパラメトリック密度推定手法である Kernel Density Estimation (KDE) の性能を数値実験により比較する．

2　既存研究

ここではカーネル指数分布族の定式化とスコアマッチング推定に関する既存の結果を説明する．

2.1　カーネル指数分布族

Fukumizu [1] は $\Omega \subset \mathbf{R}^m$ に値をとる確率変数の分布がなす無限次元指数分布族の定式化を与えた．Ω 上の関数がなすヒルベルト空間 $\mathcal{H}$ が再生核 k を持つとし，

$$\Phi(f) = \log \int_\Omega \mathrm{e}^{f(x)} q_0(x)\,\mathrm{d}x < \infty$$

なる $f \in \mathcal{H}$ 全体を $\mathcal{F}$ とする．このとき

$$\mathcal{P} = \left\{ p_f(x) = \mathrm{e}^{f(x)-\Phi(f)} q_0(x) \;\middle|\; f \in \mathcal{F} \right\}$$

を q_0 と k が定めるカーネル指数分布族という．

カーネル指数分布族 $\mathcal{P}$ において，正規化定数 $\Phi(f)$ は一般には計算困難な量である．このため，最尤推定などの正規化定数の計算を伴う推定法を用いることができない点が課題となる．

2.2　スコアマッチング

密度関数 $p(x)$ のスコア関数を $\nabla \log p(x)$ と定義し，2 つの密度関数 p, q のスコア関数の 2 乗誤差の期待値

$$J(p\|q) = \frac{1}{2}\mathbf{E}_p \left\| \nabla \log p(x) - \nabla \log q(x) \right\|^2 \quad (1)$$

を Fisher ダイバージェンスと定める．真の密度 p に対して (1) を最小化する q を推定量とする手法をスコアマッチングという．Hyvärinen [2] はパラメトリックモデルにおいて，スコアマッチングが正規化定数の計算を回避する有用な手法であることを示した．

Sriperumbudur et al. [3] は $\mathcal{P}$ を用いたスコアマッチング推定を提案した．p と p_f の Fisher ダイバージェンスは

$$J(p\|p_f) = \frac{1}{2}\langle f, Cf\rangle_{\mathcal{H}} + \langle f, \xi\rangle_{\mathcal{H}} + \text{const.}$$

と表される．ただし，作用素 C および関数 ξ は

$$\begin{aligned} Cf &= \mathbf{E}_p\left[\Sigma_{i=1}^m \partial_i f(x) \partial_i k(x,\cdot)\right] \\ \xi &= \mathbf{E}_p\left[\Sigma_{i=1}^m \left(\partial_i k(x,\cdot)\partial_i \log q_0(x) + \partial_i^2 k(x,\cdot)\right)\right] \end{aligned}$$

で定義される．C, ξ の定義中の期待値をサンプル $X_1, \ldots, X_n$ による経験平均に置き換えたものを $\hat{C}, \hat{\xi}$ をすると，Fisher ダイバージェンスは定数の差を無視して $\hat{J}(f) = \frac{1}{2}\langle f, \hat{C}f\rangle_{\mathcal{H}} + \langle f, \hat{\xi}\rangle_{\mathcal{H}}$ と推定され，これの正則化項付き最適化が線形方程式の求解に帰着される．

命題 1 (Sriperumbudur et al., 2017)**.** スコアマッチング推定量を $f_{\lambda,n} = \mathrm{argmin}_{f\in\mathcal{H}} \hat{J}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ と定める．$\beta \in \mathbf{R}^{nd}$ を線型方程式

$$(G + \lambda I)\beta = \frac{1}{\lambda} h$$

の解とすると $f_{\lambda,n} = -\frac{1}{\lambda}\hat{\xi} + \sum_{a,i} \beta_{(a-1)d+i} \partial_i k(X_a, \cdot)$ と書ける．G および h の成分は次で与えられる：

$$\begin{aligned} G_{(a-1)m+i,(b-1)m+j} &= \partial_i \partial_{j+m} k(X_a, X_b), \\ h_{(a-1)m+i} &= \left\langle \partial_i k(X_a,\cdot), \hat{\xi} \right\rangle_{\mathcal{H}}. \end{aligned}$$

ただし，命題 1 の $f_{\lambda,n}$ が $\mathcal{F}$ に属すことを保証するためには，再生核の有界性 $\sup_{x\in\Omega} k(x,x) < \infty$ などが要求される．スコアマッチング推定量は一致性をもち，収束レートの保証も与えられている．

2.3　既存手法の課題

既存手法では再生核の有界性が要求される．このため，正規分布全体を含むモデルを用いることができない等，カーネル指数分布族に著しい制限が生じる．

また，Fisher ダイバージェンスの変形に要求される境界条件は，非負値データの密度推定を行う際には真の

密度が指数分布や切断正規分布のような基本的な分布に対しても成立しない強い仮定である．

3 提案手法

まず，非負実数値データに対する密度推定法を提案する．

推定に用いる分布族は

$$\left\{ p_{\theta,f}(x) = \mathrm{e}^{\theta_1 x + \cdots + \theta_d x^d + f(x) - \Phi(\theta,f)} \middle| \theta \in \Theta, f \in \mathcal{H}_\sigma \right\}$$

とする．$\mathcal{H}_\sigma$ は再生核 $k(x,y) = \mathrm{e}^{-(x-y)^2/2\sigma^2}$ を持つヒルベルト空間，Θ は $\int_0^\infty \mathrm{e}^{\theta_1 x + \cdots + \theta_d x^d}\,\mathrm{d}x < \infty$ なる θ 全体である．また，$\Phi(\theta,f)$ は正規化定数である．この分布族は非有界な再生核 $k(x,y) = (xy+1)^d + \mathrm{e}^{-(x-y)^2/2\sigma^2}$ から定まるカーネル指数分布族である．

一般化 Fisher ダイバージェンスを

$$J_s(p\|q) = \frac{1}{2}\mathbf{E}_p \left(s(X)\partial \log p(X) - s(X)\partial \log q(X) \right)^2$$

と定めると，$s(x)^2\sqrt{\partial_1\partial_2 k(x,x)}p(x) \to 0\ (x \to 0, \infty)$ という境界条件のもとで既存手法と同様の変形ができる．よって適切に関数 s をとれば，J_s の最小化によって推定量を構成する手法により境界条件を緩和できる．

実際に推定量を構成する手順は以下のようになる．まず，$f \in \mathcal{H}_\sigma$ を固定すると，$\{p_{\theta,f}\}_{\theta\in\Theta}$ は有限次元指数分布族であるから，通常のスコアマッチングにより θ を推定できる．一方，$\theta \in \Theta$ を固定したとき，$\{p_{\theta,f}\}_{f\in\mathcal{H}_\sigma}$ は有界な再生核によるカーネル指数分布族であるから，既存手法と同様に線形方程式の求解によって f の推定量を構成できる．したがって，θ, f の一方を固定し他方のスコアマッチング推定量を計算するという操作を反復することによって，$\{p_{\theta,f}\}$ 全体の中で一般化 Fisher ダイバージェンスを最小化する元を探索することが可能である．図 1 に推定量のプロットの一例を示す．

次に，多次元データへの拡張を考える．この場合はパラメータ空間 Θ が複雑でそのまま扱うのは難しい．そこで，分布族を

$$p_{\theta,f}(x) = \exp\left(\sum_{i=1}^{m}\sum_{k=1}^{d_i} \theta_{ik} x_i^k + f(x) - \Phi(\theta,f) \right)$$

に制限する．この分布族における推定量を次のように構成する．まず $f = 0$ とした上で各 θ_i をサンプルの第 i 成分のみを用いてスコアマッチングにより推定する．次に $\theta_1, \ldots, \theta_m$ を固定して f を推定するが，これは 1 次元の場合と同様である．

4 数値実験

提案手法と Gaussian カーネル密度推定 (KDE) による密度関数の推定量の精度を数値実験により比較する．

多次元データ $X = (X_1, \ldots, X_m)$ を，各成分 X_i の平均が 2，分散が 2，X_i と X_j の共分散が 1 であるような多次元正規分布を第 1 象限で切断して得られる分布から 500 点サンプリングし，それを元に密度関数の推定量を構成した．図 2 は真の密度と推定量の一般化 Fisher ダイバージェンスがデータの次元の増加に伴い悪化する様子である．提案手法は次元が増加すると KDE より高精度な推定量が得られることが確認される．

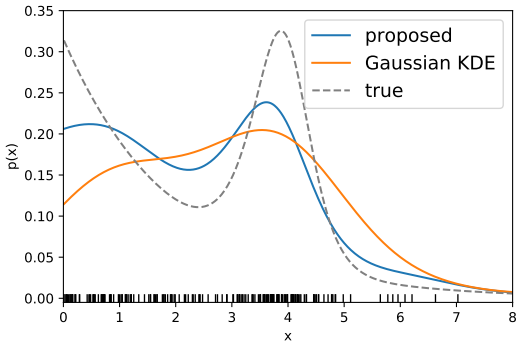

図 1. 提案手法 (青) と KDE (橙) による密度推定の一例．下段の黒点はサンプルを表す．

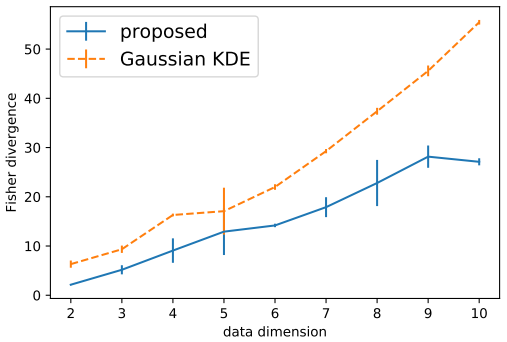

図 2. データの次元と推定精度の関係．提案手法 (青実線) と KDE(橙点線) の比較．

参考文献

[1] K. Fukumizu: Exponential manifold by reproducing kernel Hilbert spaces. In P. Gibilisco, E. Riccomagno, M.-P. Rogantin, and H. P. Wynn, editors, *Algebraic and Geometric Methods in Statistics*, 291–305. Cambridge University Press, 2009.

[2] A. Hyvärinen: Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, **6**, 695–709, 2005.

[3] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar: Density estimation in infinite dimensional exponential families. *J. Mach. Learn. Res.*, **18**(57), 1–59, 2017.