

Shape-constrained density estimation in the space of phylogenetic trees (系統樹の空間での形状制約つき密度推定)

数理情報学専攻 48206221 高澤 祐槻
指導教員 清 智也 教授

1 はじめに

系統樹の推論は生物学における重要な課題である。ただし、統計的推測の不確かさや生物学的プロセスの影響によって、使用する遺伝子の場所により異なる系統樹が推測されることがある。

近年、[1] によって導入された系統樹の空間 (tree space) 上で統計的手法を開発する取り組みがなされている。ただし、空間や系統樹推測の手法の複雑性により、系統樹のパラメトリックな分布を与えることは難しいため、ノンパラメトリックな手法が求められている。

本研究では、系統樹の空間において対数凹という形状制約を用いたノンパラメトリック最尤推定手法を考える。まず、系統樹の空間における最尤推定量の存在性と一意性に関する結果を 1 次元の場合、多次元の場合に分けて報告する。さらに、2 次元以下での計算のアルゴリズムも提案する。また、数値実験を用いてカーネル密度推定との推定精度の比較を行い、最後にクラスタリングへの応用例も示す。

2 先行研究

2.1 tree space

[1] により導入された tree space $\mathcal{T}_n$ は、 n 種の系統樹の空間をユークリッド空間の正象限を貼り合わせてモデルする。これは曲率が非正の距離空間である。特に 3 種の系統樹の空間 $\mathcal{T}_3$ は、3 つのトポロジーに対応する半直線 $\mathbb{R}_+$ が原点でつながった 1 次元の空間であり、4 種の系統樹の空間 $\mathcal{T}_4$ は、15 個のトポロジーに対応する正象限 $\mathbb{R}_+^2$ が複雑につながった 2 次元の空間である。この空間上のノンパラメトリック密度推定手法としては、カーネル密度推定を適用したもの [2] がある。

2.2 ユークリッド空間での対数凹密度の最尤推定

ユークリッド空間においては対数凹という密度の形状制約つき推定法が開発された [3]。対数凹密度は正規分布等を含むノンパラメトリックな密度のクラスである。この推定法は最尤推定を用いるため、バンド幅等のハイパーパラメータの指定が不要である。正しくモデル化され、大サンプルの場合には数値的にカーネル密度

推定より精度が良いことも報告されている [3]。

3 最尤推定量の存在性、一意性

3.1 1 次元 tree space

1 次元 tree space においては、ユークリッド空間の場合と同様の結果が成り立つことを確認した。 $\mathcal{T}_3$ 上の対数凹密度の集合を $\mathcal{F}_0$ とおく。

定理 1. $(X_1, \dots, X_n)$ ($n \geq 2$) をある密度 f_0 からの独立標本とする。この時、確率 1 で $l(f)$ を $\mathcal{F}_0$ 上で最大化する f_0 の推定量 $\hat{f}_n$ が存在し、一意である。

3.2 多次元 tree space

多次元 tree space においては、最尤推定量の存在のために追加の条件が必要となる。 $\bar{\mathcal{F}}_0$ を $\mathcal{T}_{d+2}$ 上の上半連続な対数凹密度の集合とする。

定理 2. $(X_1, \dots, X_n)$ をある密度 f_0 からの独立標本とし、その凸包を C_n とおく。 $\mathcal{T}_{d+2}$ の各正象限 O_i で以下の条件のいずれかが成り立つならば、 $l(f)$ を $\bar{\mathcal{F}}_0$ 上で最大化する f_0 の推定量 $\hat{f}_n$ が存在し、 ν -almost everywhere に一意。

- (a) $O_i \cap C_n$ が点を含まない、
- (b) $\nu(C_n \cap O_i) > 0$,
- (c) $C_n \cap O_i$ は境界での値のみを含み、点を含む各境界は (a) を満たす象限とつながっている。

4 最尤推定量の計算法

ユークリッド空間の場合と同様に、対数凹最尤推定量は次の $\mathbb{R}^n$ 上の関数

$$\sigma_n(y) = -n^{-1} \sum_{i=1}^n y_i + \int_{\mathcal{T}_3} \exp(h_y(x)) d\nu(x) \quad (1)$$

を最大化することにより得ることができる。この関数は凸関数であり、劣勾配を求めることができる。ただし、 $h_y: \mathbb{R}^n \rightarrow [-\infty, \infty]$ は、 $h_y(X_i) \geq y_i$ ($i = 1, 2, \dots, n$) という条件を満たす最小の凹関数である。最適化において難しい点は、この関数 h_y を求めることであり、その問題は $\mathcal{T}_{d+2} \times \mathbb{R}$ での凸包を求める問題に帰着される。以下、集合 $\{(X_i, y_i) \in \mathcal{T}_{d+2} \times \mathbb{R}\}$ の凸包の $d = 1$

の場合の計算と、 $d = 2$ の場合の近似について述べる。

4.1 1次元 tree space の場合

$\mathcal{T}_3$ は 3 つの半直線がつながった空間であるため、特に $\{X_i\}$ が 1 つの象限のみに値をとる場合は $\mathbb{R}^2$ の凸包計算に帰着される。そうでない場合は、凸包には必ず $(0, y)$ という形の点が含まれる。このような点を先に求めておくことにより、問題を $\mathbb{R}^2$ での凸包計算に帰着させることができる。

4.2 2次元 tree space の場合

$\mathcal{T}_4 \times \mathbb{R}$ における凸包の近似アルゴリズムを説明する。

まず、 $S_0 = \{(X_i, y_i)\}_{i=1}^n$ と初期化する。また、 S_0 の各非負象限 O への制限に対して凸包をとったものを $H_0(O)$ とおき、 $H_0 = \cup_O H_0(O)$ と定義する。ここで、集合 S_l, H_l ($l = 1, 2, \dots$) を次のように定義する。第一に、 H_{l-1} の点を結ぶ測地線が原点を通る軸上と交差する点 $(0, y)$ を見つけ、その y の最大値と最小値を y_{l1}, y_{l0} とおく。集合 T_l を $S_{l-1} \cup \{(0, y_{l1}), (0, y_{l2})\}$ と定義する。 $y_{l1}(y_{l0})$ の計算は非凸最適化問題となるため、数値計算上は下限(上限)で代用をしている。

続いて、全ての T_l の点の組み合わせに対し測地線を取り、その境界面との交差点全てを T_l に追加する。ただし境界面とは、 $\mathcal{T}_4$ の各象限の軸と $\mathbb{R}$ がなす 2 次元平面のことである。さらに、各境界面 B に対して、 $T_l \cap B$ の凸包を取り、その頂点を成さない点は T_l から削除し、 $S_l = T_l$ とする。最後に S_l の各象限 O への制限に対して凸包を取ったものを $H_l(O)$ とし、 $H_l = \cup H_l(O)$ とする。次の結果は、 H_l を凸包の近似とすることの妥当性を示している。

定理 3.

$$\cup_{l=0}^{\infty} H_l = \text{conv } S_0$$

5 数値実験

5.1 カーネル密度推定との比較

次のような 2 つの $\mathcal{T}_4$ 上の対数凹密度を考える。

密度 (a) $\mathcal{T}_4$ の全ての象限でサポートされる平均 0, 分散 I の切断正規分布

密度 (b) $\mathcal{T}_4$ のある 6 つの連結した象限でサポートされる平均 0, 分散 I の切断正規分布

サイズ 50, 100, 200, 300, 500 のサンプルを生成し、提案手法並びにカーネル密度推定により密度推定を行った。10 サンプルの平均積分二乗誤差の比較が、図 1 である。

密度 (a) に関してはユークリッド空間の場合同様、サ

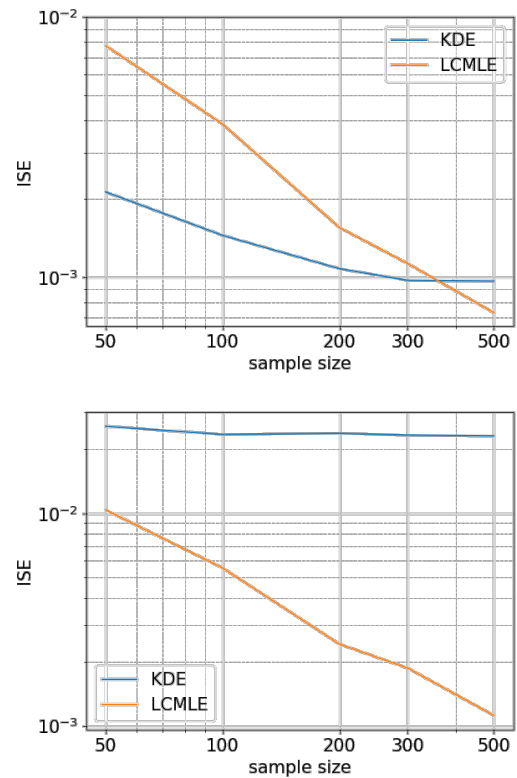


図 1. LCMLE (青) が提案手法、KDE (橙) がカーネル密度推定の平均積分二乗誤差 (ASE) を表す。(上) 密度 (a) の推定結果; (下) 密度 (b) の推定結果。

ンプルサイズが増えるにしたがって対数凹最尤推定量の精度が上回る様子が確認された。一方、tree space に特有なサポートになっている密度 (b) に関しては、対数凹最尤推定量が全サンプルサイズでより良い推定精度を持った。これは、対数凹最尤推定量がサポートをデータ点の凸包の閉包に限るという性質によるものと考えられる。

参考文献

- [1] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. “Geometry of the space of phylogenetic trees”. In: *Advances in Applied Mathematics* 27.4 (2001), pp. 733–767. ISSN: 01968858. DOI: 10.1006/aama.2001.0759.
- [2] Grady Weyenberg, Ruriko Yoshida, and Daniel Howe. “Normalizing Kernels in the Billera-Holmes-Vogtmann Treespace”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14.6 (2017), pp. 1359–1365. ISSN: 15455963. DOI: 10.1109/TCBB.2016.2565475.
- [3] Madeleine Cule, Richard Samworth, and Michael Stewart. “Maximum likelihood estimation of a multidimensional log-concave density”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 72.5 (2010), pp. 545–607. ISSN: 14679868. DOI: 10.1111/j.1467-9868.2010.00753.x. arXiv: 0804.3989.