系統樹の空間における形状制約つき密度推定

東京大学・情報理工学系研究科　髙澤 祐槻
東京大学・情報理工学系研究科　清 智也
Yuki Takazawa and Tomonari Sei
Graduate School of Information Science and Technology
The University of Tokyo

概要

近年，推測された系統樹の集合をある距離空間上に埋め込み統計的に分析する研究がなされている．本稿では，この系統樹空間における統計手法のいくつかを紹介した上で，ノンパラメトリック密度推定手法の一つである対数凹最尤推定を系統樹の空間に応用することを考える．この推定量の存在条件，一意性に関する結果と低次元の場合の計算アルゴリズムを示し，既存のカーネル密度推定の手法と数値的に精度比較を行う．

1　はじめに

系統樹の推定は，生物学における重要な問題の一つである．遺伝子やタンパク質のデータを用いて系統樹の復元する手法は発達しており，様々なモデルや推論手法が利用できる (Felsenstein, 2004).

しかし，推論の不確実性や水平移動などの不規則な生物学的プロセスの存在により，異なる遺伝子座（データ）が異なる進化史を示すことが通常である．この問題に対し，2001 年に Billera et al. (2001) が n 葉の系統樹空間を構成したことをきっかけに，この空間上での統計手法の開発というアプローチがとられている．この空間は，非正曲率の測地的距離空間であるという良い性質があるため，これを用いて Fréchet 平均による点推定 (Benner et al., 2014), 主成分分析 (Nye, 2011), カーネル密度推定を用いた外れ値検知 (Weyenberg et al., 2014, 2017), 信頼集合の構成 (Willis, 2019) などが開発されている．

推定される系統樹には多数の要因が影響するため，パラメトリックなアプローチによる系統樹の分布の特定は困難であり，モデルの誤指定のリスクも高い．その意味で，ノンパラメトリックアプローチは一般に分布の制約を少なく指定できるので，この場合は望ましい．Weyenberg et al. (2014, 2017) で提案されたカーネル密度推定は，この目的のために設計されたものである．

本稿では，系統樹空間における対数凹密度の最尤推定量がある条件下で存在し，その推定が 1 次元の場合に実装可能であることを示す．多次元の場合は凸包計算の難しさに起因して最尤推定量を求めることが困難であるが，2 次元の場合には最尤推定量を近似的に計算できる．

構成は以下のとおりである．2 節では，系統樹空間の基本的な概念やその幾何学的性質について概説する．また，系統樹空間を含む空間である Hadamard 空間における凸解析の重要な概念をいくつか紹介する．3 節では，系統樹空間における統計分析手法のうち，本研究の内容と関連の深いものに

ついて簡単に紹介する．4 節では系統樹空間で 1 次元および多次元の対数凹密度の最尤推定に関する主要な理論的結果を示すとともに，推定量の特徴づけを行う．5 節では低次元の場合における最尤推定量の計算方法を説明する．6 節で数値実験の結果を示し，最後に 7 節で本稿のまとめを述べる．

2　系統樹空間と Hadamard 空間

2.1　系統樹のモデル化と，系統樹空間の構成

本節では，Billera et al. (2001) による系統樹と系統樹空間のモデリングの概説をする．まず系統樹は，葉にのみラベルを付けた木としてモデル化される．$n+1$ 個のラベル付きの葉を持つ木を n-tree と呼ぶ．n-tree は，共通の祖先を表す一つの「根」と，n 個の現存する分類群を表す葉を持つ系統樹と考えることができる．内部辺とは，葉に直接接続されていない辺のことである．二分木の場合，内部辺の数は $n-2$ であることが容易にわかる．非二分木は二分木の内部辺を「縮退」して得られるので，内部辺の数は少なくなる．また，二分木である n-tree の異なるトポロジーの数は $(2n-3)!!$ 個である (Felsenstein, 1978)．ここで，異なる木のトポロジーの辺は，同じように葉を分割すれば同じと見なすことができることに注意する．

系統樹空間は以下のように定式化される．まず，各二分木のトポロジーに対し，内部辺一つの長さと一つの軸が対応するように $n-2$ 次元のユークリッド正象限を対応させる（図 1）．前述のように，非二分木は一部の内部辺を収縮させた二分木とみなすことができるので，これらの象限の境界を構成していると考えることができる．したがって，各トポロジーに対応する象限を共通の面で貼り合わせることができる．このようにして得られた n-tree の空間を以下系統樹空間と呼び，$\mathcal{T}_n$ と表記することとする．特に，内部辺を持たない n-tree は，すべての象限に接続された系統樹の空間の中心に位置する点である．本稿では，この点を原点と呼ぶ．

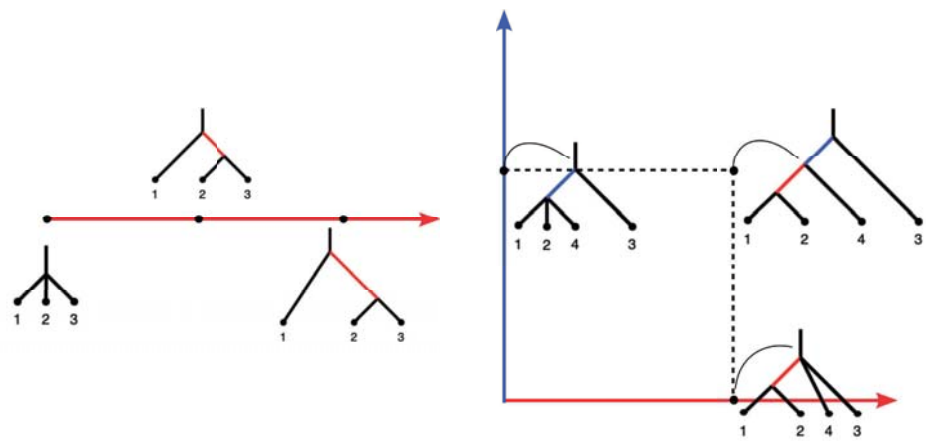

図 1　一つの二分木のトポロジーの空間への埋め込み．（左）：3-木の一つのトポロジーの埋め込み．（右）：4-木の一つのトポロジーの埋め込み．

本稿での実際の密度推定においては，1 次元または 2 次元の場合を主に考える．次元とは，各象限

の次元，または立方複体として見た場合の空間の次元を意味する．したがって，p 次元の系統樹空間とは，$\mathcal{T}_{p+2}$ を意味する．1 次元系統樹空間 $\mathcal{T}_3$ では，$3!! = 3$ 個のトポロジーしかなく，あるトポロジーを表す各象限は，三分木を表す点である原点で他の二つと接続する半直線である（図 2 左）． 2 次元系統樹空間 $\mathcal{T}_4$ は $5!! = 15$ のトポロジーを持ち，各象限は非負のユークリッド平面で，各軸に 2 種類の別の象限が接続されている（図 2 中央）．この $\mathcal{T}_4$ 全体は，複雑につながっており通常のユークリッド空間上に図示することができないが，各軸を点，各象限を 2 点を結ぶ線分で表すと，ピーターセングラフとして図示できることが分かっている（図 2 右）．詳しくは，例えば Billera et al. (2001) や Lubiw et al. (2020) を参照されたい．

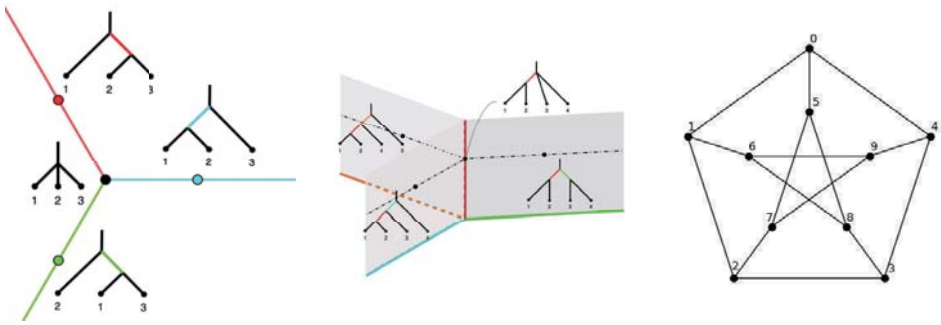

図 2 （左）：$\mathcal{T}_3$．（中央）：$\mathcal{T}_4$ の一部．（右）：$\mathcal{T}_4$ の象限の繋がり方を表すピーターセングラフ．各頂点の番号は，各象限の軸に対し便宜上インデックスをつけたものである．

2.2 系統樹空間の幾何

系統樹空間は距離空間として容易に定式化できる．この空間はユークリッド非負象限を合わせたものであるから，同じ象限にある 2 点間の距離は，通常のユークリッド距離と定義できる．また，異なる象限にある任意の 2 点については，その 2 点を端点とするパスの長さの下限として距離を設定することができる．系統樹空間は有限個の象限からなり，また象限の並びが決まれば，その並びの中で長さが最小になる唯一のパスが存在することが容易にわかるため，この下限はあるパスによって達成される．この距離が最小になるようなパスを測地線と呼ぶ．

系統樹空間は，このように構成された距離に対して非正曲率を持つ完備な距離空間となる．一般にそのような空間を Hadamard 空間と呼ぶ．証明は例えば Bačák (2014) を参照されたい．ここで，距離空間の曲率が非正であるということは，次のことが成り立つことである．$(\mathcal{H}, d)$ を距離空間とし，任意の点 $x_1, x_2, x_3 \in \mathcal{H}$ に対して $x'_1, x'_2, x'_3 \in \mathbb{R}^2$ は次の性質を持つ点であるとする．

$$d(x_i, x_j) = d(x'_i, x'_j) \qquad (i, j \in \{1, 2, 3\}). \tag{2.1}$$

三角形 $x'_1x'_2x'_3$ は比較三角形と呼ばれる．ここで，$\lambda \in [0, 1]$ に対し， $d(x_i, \gamma_{x_i,x_j}(\lambda)) = \lambda d(x_i, x_j)$ となる x_i から x_j への測線上の点を $\gamma_{x_i,x_j}(\lambda)$ と表す．非正曲率の性質は次の不等式で特徴付けられる（CAT(0) 不等式）．

$$d(x_1, \gamma_{x_2,x_3}(\lambda)) \leq \|x_1 - ((1 - \lambda)x_2 + \lambda x_3)\|_2. \tag{2.2}$$

ただし，$\|\cdot\|_2$ は $\mathbb{R}^2$ における 2 乗ノルムを表す．この性質は，系統樹空間における多くの重要な結果の鍵である．さらに Hadamard 空間では，任意の 2 点を結ぶ測地線は一意であることが保証されている．実際に，系統樹空間における測地線を $O(p^4)$ 時間（p は系統樹空間の次元）で求めるアルゴリズムも開発されている (Owen and Provan, 2011)．系統樹空間における測地線は各象限における線分をつなげたものになっており，場合によっては，端点と原点を結ぶ線分二つからなる「コーンパス」になる．特に，1 次元系統樹空間 $\mathcal{T}_3$ においては任意の別象限に存在する 2 点を結ぶパスは必然的にコーンパスとなる．2 次元系統樹空間 $\mathcal{T}_4$ では，2 点間の測地線がコーンパスになるかどうかは，2 点間の角度にのみ依存する．この性質についての詳しい説明は，例えば，Lubiw et al. (2020) を参照されたい．

2.3 Hadamard 空間における凸集合と凹関数

本節では，Hadamard 空間上の凸集合と凹関数の概念を定義する．

まず，凸集合と凹関数を Bačák (2014) に倣って定義する．Hadamard 空間を $(\mathcal{H}, d)$ とし，任意の点 $x, y \in \mathcal{H}$ に対して，その間の唯一の測地線を $[x, y]$ とする．集合 $A \subseteq \mathcal{H}$ は，任意の点 x, y に対して $[x, y] \subseteq \mathcal{H}$ が成り立つとき，凸であるという．本稿では凹関数に興味があるため，通常の凸解析の場合と異なり，関数 $f : \mathcal{H} \to [-\infty, \infty]$ のエピグラフを $\mathrm{epi}f = \{(x, \mu) \in \mathcal{H} \times \mathbb{R} \mid f(x) \geq \mu\}$ で定義する．関数 f は，そのエピグラフが凸である時，凹関数であるとする．Hadamard 空間の積も Hadamard 空間でもあり，特に $\mathcal{H} \times \mathbb{R}$ は Hadamard 空間であることに注意．

凸包の概念も同様に定義することができる．任意の集合 $S \subseteq \mathcal{H}$ に対し，S の凸包とは S を含む $\mathcal{H}$ の最小の凸集合のことである．この集合が存在することは，凸集合の共通部分が凸であることと，$\mathcal{H}$ が凸であることから導かれる．S の凸包を $\mathrm{conv}S$ と表記する．Hadamard 空間において，凸包は次の補題のように書けることが知られている．

補題 1 (Bačák (2014), Lemma 2.1.8)**.** $S \subseteq \mathcal{H}$ に対し，$C_0 = S$ とし，$n \in \mathbb{N}$ に対し再帰的に C_n を $C_n = \{x \in \mathcal{H} \mid x \text{ は } C_{n-1} \text{の } 2 \text{ 点の測地線上にある }\}$ で定義する．この時，

$$\mathrm{conv}S = \bigcup_{n=0}^{\infty} C_n. \tag{2.3}$$

この補題は，2 点間の測地線を無限回とれば凸包が得られることを示している．

次に，関数の凹包という重要な概念を定義する．$\mathcal{H}$ 上の任意の関数 f の凹包 $\mathrm{conc}f$ は f によって下から抑えられる最小の凹関数である．ユークリッド空間におけるその存在と一意性はよく知られた結果であり，例えば Rockafellar (1970) の第 5 節に詳しい説明がある．Hadamard 空間においても同様の議論により凹包の存在性と一意性を示すことができる．

補題 2. 任意の関数 $f : \mathcal{H} \to [-\infty, \infty]$ に対しその凹包 $\mathrm{conc}f$ は存在し一意である．

証明（スケッチ）．f を $\mathcal{H}$ 上の任意の関数とし，関数 g を $g(x) = \sup\{\mu \mid (x, \mu) \in \mathrm{conv}(\mathrm{epi}f)\}$ と定義する．すると，g は凹であり，g は f によって下から抑えられる．g の最小性は凸包の最小性から従う． □

前の補題の証明の重要な含意は，関数の凹包の導き方である．すなわち，エピグラフの凸包を計算し，その上限を取るという操作である．

特に本稿では，関数が次のような形で書ける場合に注目する．

$$f_y(x) = \sum_{i=1}^{N} I(x = X_i) + y_i. \tag{2.4}$$

ただし，任意の命題 P に対し, I は次のように定義される指示関数である：

$$I(P) = \begin{cases} 0 & (\text{if } P) \\ -\infty & (\text{otherwise}) \end{cases}. \tag{2.5}$$

この場合，関数 f_y のエピグラフは N 本の「垂直な半直線」の集合となる．

$$\mathrm{epi} f_y = \{(X_i, \mu) \mid \mu \leq y_i, i = 1, \ldots, N\}. \tag{2.6}$$

これから，$h_y = \mathrm{conc} f_y(x)$ をこの凹包を表す記号として用いて，$h_y(X_i) \geq y_i \quad (i = 1, \ldots, n)$ である最小の凹関数などと呼ぶことにする．

凹性と共に仮定されることが多い条件の一つとして，上半連続性がある．関数 $f : \mathcal{H} \to [-\infty, \infty]$ が点 x で上半連続であるとは，$\limsup_{y \in \mathcal{H}: d(x,y) \to 0} f(y) \leq f(x)$ が成り立つことである．凹包と同様に，関数 f の上半連続包を，f によって下から抑えられる最小の上半連続関数として定義することが出来る．さらに，関数 f の上半連続凹包も，f によって下から抑えられる最小の上半連続凹関数として定義することが出来る．この関数を $\overline{\mathrm{conc}} f$ とかき，式 (2.4) に対して $\bar{h}_y = \overline{\mathrm{conc}} f_y$ とおく．4.3 節において，対数凹密度のクラスにおける最尤推定量の対数はこの $\bar{h}_y$ という形で与えられることを述べる．

3　系統樹空間での統計

本節では，系統樹空間において既に開発されている統計手法のいくつかについて概説する．

3.1　Fréchet 平均

サンプルの Fréchet 平均は，与えられたサンプル $X_1, \ldots, X_n$ と重み $w_1, \ldots, w_n$ に対してその二乗距離和の最小化として定義される：

$$\mu = \arg\min_{x \in \mathcal{H}} \sum_{i=1}^{n} w_i d(x, X_i)^2. \tag{3.1}$$

ただし，w_i は適当な正の重みを表す．Fréchet 平均はヒルベルト空間においては算術平均と一致するため，これは算術平均の一般化として見なすことが出来る．Hadamard 空間において，Fréchet 平均は一意に存在することがわかっている．これは，関数 $d(\cdot, X_i)^2$ の強凸性と連続性によるものである．Fréchet 平均の計算に関しても，近接点法を用いたアルゴリズムが知られている (Bačák, 2013)．Hadamard 空間における Fréchet 平均に関する結果の証明は，例えば Bačák (2014) を参照されたい．

3.2 カーネル密度推定と外れ値検知

Weyenberg et al. (2014, 2017) では，主に外れ値検知の目的のために，系統樹空間上のカーネル密度推定量を提案している．系統樹空間 $\mathcal{T}_{p+2}$ 上のサンプル $\{X_1, \ldots, X_n\}$ が与えられた時，彼らは次のような形の推定量を提案した：

$$\hat{f}(X) \propto \frac{1}{n} \sum_{i=1}^{n} k(X, X_i). \tag{3.2}$$

ただし，k はカーネル関数を表す．

特に，彼らは次のような正規型のカーネル関数を用いている：

$$k(X, X_i) \propto \exp\left(- \left(\frac{d(X, X_i)}{h_i} \right)^2 \right). \tag{3.3}$$

ユークリッド空間の場合，これらのカーネル関数に対する正規化定数は次のようにかける．

$$c(X, h_i) = (2\pi h_i)^{-p/2}. \tag{3.4}$$

しかし，系統樹空間においては，そのように正規化定数を陽に書くことが難しい．実際，正規化定数はカーネル関数の中心の位置によって変化する．例えば，中心 X が原点の場合には正規化定数が $2^p\{(2\pi h_i)^{p/2} \cdot (2p+1)!!\}^{-1}$ となるのに対し，中心 X が全ての象限の境界から十分遠くにある場合は，正規化定数は式 (3.4) に近づく．Weyenberg et al. (2017) では正規化定数のある下限を用いることでこの問題の解決を試みている．この下限は，ホロノミック勾配法あるいは古典的なガウス求積の方法を用いて求められるものとなっている．

3.3 系統樹空間上の確率測度

Willis (2019) は系統樹空間上の信頼集合の構成法を与える中で，確率測度の可能な構築法の一つについて考察している．以下，その与え方について説明する．

$(\mathcal{T}_{p+2}, d)$ を系統樹空間とする．この空間は $(2p+1)!!$ 個の p 次元の非負象限からなることに注意する．このことを利用して，$\mathcal{T}_{p+2}$ の任意の集合 A を各象限における集合の和として $A = \cup_{i=0}^{(2p+1)!!} A_i$ と書くことができる．ここで，A_0 は A に含まれる木のうち $\mathcal{T}_{p+2}$ の境界にも含まれる木の集合，A_i は A に含まれる木のうち i 番目の正象限に含まれる木の集合を示す．ここで，ν_B を $\mathbb{R}^p$ 上のルベーグ測度とし，ν を $\nu(A) = \sum_{i=1}^{(2p+1)!!} \nu_B(A_i)$ で定義する．すると，この ν はシグマ加法性を保存するので，完備化により ν は完備測度となる．

3.2 節で紹介したカーネル密度推定量の文献 (Weyenberg et al., 2014, 2017) においてはどのような確率測度を考えているかに関して明示をしていないが，このようなものを暗に仮定していると考えることができる．残りの節では，測度 ν を基底測度として，これに関する密度を考える．

4　対数凹最尤推定量とその性質

ここからは，ユークリッド空間におけるノンパラメトリック手法の一つである対数凹密度推定について系統樹空間で考える．

4.1　対数凹密度

ユークリッド空間におけるノンパラメトリック密度推定手法として対数凹密度という形状制約を用いるものが研究されている．対数凹密度は，対数が凹関数である確率密度のクラスである．ユークリッド空間上の対数凹密度のクラスは正規分布等の多くの実用的な分布を含む．ノンパラメトリック手法の一つではあるが，推定には最尤推定を用いることができるため，カーネル密度推定などの古典的な平滑化手法と異なり，平滑化パラメータやバンド幅などのハイパーパラメータを指定する必要がない．ユークリッド空間の場合，Cule et al. (2010) が，1 次元の場合と多次元の場合の両方で，十分なサンプルサイズのもと確率 1 で最尤推定量が存在し，その推定量の計算は n 個の標本点を持つ場合に $\mathbb{R}^n$ における凸最適化問題に帰着することを示している．

系統樹空間においては，パラメトリックな分布があまり考案されていないためどのような分布を含むクラスであるかを説明することは難しいが，3.2 節で紹介したカーネル密度推定に用いられている，等方向の分散を持つような正規分布型の分布を含むことは容易にわかる．

4.2　系統樹空間における対数凹密度の存在性と一意性

本節では，対数凹最尤推定量の存在性・一意性についての結果を述べる．まず，最も単純な 1 次元の系統樹空間 $\mathcal{T}_3$ を考える．この空間において，先に定義した基底測度 ν に関する対数凹密度の集合を $\mathcal{F}_0$ とする．次の定理は，$\mathcal{T}_3$ における最尤推定がユークリッド空間の場合と同様に存在することを示すものである．

定理 3. $(X_1, \ldots, X_n)$ を $\mathcal{T}_3$ 上の密度 $f \in \mathcal{F}_0$ からの独立標本とする（$n \geq 2$）．この時，確率 *1* で，f の最尤推定量 $\hat{f}_n$ が存在し一意である．すなわち，$\hat{f}_n$ は対数尤度関数 $l(f) = \sum_{i=1}^{n} \log f(X_i)$ の $\mathcal{F}_0$ 上における一意の最大解である．

続いて，多次元の系統樹空間 $\mathcal{T}_{p+2}$ を考える． $\bar{\mathcal{F}_0}$ を $\mathcal{T}_{p+2}$ 上の上半連続対数凹密度の集合とし，$l(f) = \sum_{i=1}^{n} \log f(X_i)$ を対数尤度関数とする．

この場合の最尤推定量は，1 次元系統樹空間やユークリッド空間の場合とは異なり，存在しない可能性がある．最尤推定量の存在に関する十分条件を導く前に，まず，最尤推定量の存在を仮定すると，一意性が以下の定理のように述べられることを示す．

定理 4. $(X_1, \ldots, X_n)$ を $\mathcal{T}_{p+2}$ 上のある密度からのサンプルとする．$\bar{\mathcal{F}_0}$ 上で $l(f)$ の最大解が存在すると仮定する．すると，最大解は $\nu-a.e.$ で一意である．

証明. $f_1, f_2 \in \bar{\mathcal{F}_0}$ がどちらも $l(f)$ を最大化すると仮定する．この時，関数 f を次のように定義す

る： $f(x) = \{f_1(x)f_2(x)\}^{1/2} / \int_{\mathcal{T}_3} \{f_1(z)f_2(z)\}^{1/2} d\nu(z)$．すると，$f$ は凹関数であり，

$$l(f) = \frac{1}{2n}\sum_{i=1}^{n}\{\log f_1(X_i) + \log f_2(X_2)\} - \log \int_{\mathcal{T}_{p+2}} \{f_1(z)f_2(z)\}^{1/2} d\nu(z)$$
$$\geq l(f_1) - \log \int_{\mathcal{T}_{p+2}} \frac{f_1(z) + f_2(z)}{2} d\nu(z) = l(f_1). \tag{4.1}$$

ここで，最後の不等式は算術平均と幾何平均の関係式から導かれ，等号は $f_1(z) = f_2(z)$ $\nu-$a.e. の時のみ成り立つ． □

次の定理は，最尤推定量が存在し，測度 ν に関してほとんど至る所で一意となるようなサンプルの十分条件を与えるものである．

定理 5. $(X_1, \ldots, X_n)$ を $\mathcal{T}_{p+2}$ 上のある密度 f からのサンプルとする（$n \geq p+1$）．$C_n = \mathrm{conv}\{X_1, \ldots, X_n\}$ とし，各非負象限 $\{O_i\}_{i=1}^{(2p+1)!!}$ において，以下のいずれかの条件が成り立つと仮定する：

(a) $C_n \cap O_i$ は空集合である．
(b) $C_n \cap O_i$ は O_i の d 次元集合となる*:* すなわち，$\nu(C_n \cap O_i) > 0$.
(c) $C_n \cap O_i$ は境界での値のみを含む．この境界を B_j $(j = 1, \ldots, J)$ とすると，各 B_j が面している象限 O_j の中で *(b)* が成り立つものが少なくとも一つ存在する．

この時，$\bar{\mathcal{F}}_0$ 上で f の最尤推定量 $\hat{f}_n$ が存在する．すなわち，$\hat{f}_n \in \bar{\mathcal{F}}_0$ は対数尤度関数 $l(f) = \sum_{i=1}^{n} \log f(X_i)$ の最大解である．さらに，これは $\mathrm{int}(C_n)$ を含まないある測度 *0* の集合を除いて一意に定まる．

なお，この十分条件は n が無限大になるにつれて 1 に近づく確率で成立する条件である．また，この条件は $\mathcal{T}_{p+2}$ のサンプルの凸包が計算できれば，簡単に確認できる．

4.3 定理の証明と最尤推定量の特徴づけ

定理 3，定理 5 の証明は，ユークリッド空間の場合 (Cule et al., 2010) と似た議論により行うことが出来る．以下，定理 5 の証明の概略を与える．

定理 *5* の証明の概略． まず，$C_n = \mathrm{conv}\{X_1, \ldots, X_n\}$ とすると，これは有界な集合となる．これは系統樹空間の非正曲率の性質から導くことが出来る．一方，補題 2.3 より，$D_0 = \{X_1, \ldots, X_m\}$，$D_l = \{D_{l-1}$の任意の 2 点を結んだ測地線上にある点 $\}$ $(l = 1, 2, \ldots)$ とすると，$C_n = \cup_l D_l$ となる．

ここで，$\mathcal{T}_{p+2}$ 上の上半連続な対数凹密度の代わりに，上半連続な対数凹関数の集合を考え，これを $\bar{\mathcal{F}}$ とする．ユークリッド空間での議論と似たように，目的関数として $\psi_n(f) = n^{-1}\sum_{i=1}^{n} \log f(X_i) - \int_{\mathcal{T}_{p+2}} f(x) d\nu(x)$ を考え，これを $\bar{\mathcal{F}}$ 上で最大化することを考える．すると，以下のことが順にわかる．

1. 任意の $g \in \mathcal{H}$ に対し，ある $f \in \mathcal{H}$ が存在し，以下の条件を満たす．

$$f(x) > 0 \ (x \in C_n), \quad f(x) = 0 \ (x \notin \mathrm{cl}(C_n)), \quad \psi_n(f) \geq \psi_n(g). \tag{4.2}$$

2. $H = \{\bar{h_y} \mid y = (y_1, \ldots, y_n) \in \mathcal{R}^n\}$ とすると，任意の $g \in \mathcal{F}$ に対しある $\bar{h_y} \in H$ が存在し，
$$\psi_n(\exp(\bar{h_y})) \geq \psi_n(g). \tag{4.3}$$
また，$\mathrm{dom}\ \bar{h_y} = \mathrm{cl}(C_n)$ であり，$\bar{h_y}$ は $\mathrm{int}(C_n)$ 上で連続である．
3. 任意の $\bar{h_y} \in H$ に対して，C_n の有界性から $\int_{C_n} \exp(h_y(x))dx < \infty$ である．そこで，$\int_{C_n} \exp(\bar{h_y}(x))dx = c$ とおき，$f_0 = \exp(\bar{h_y})/c = \exp(\bar{h}_{y-\log c})$ とおけば，
$$\psi_n(f_0) \geq \psi_n(\exp(\bar{h_y})) \tag{4.4}$$
である．
4. $\|y\|_\infty \to \infty$ で $\psi_n(\exp(\bar{h_y})) \to -\infty$ である．よって，ある $M > 0$ を用いて，
$$\sup_{y \in \mathbb{R}^n} \psi_n(\exp(\bar{h_y})) = \sup_{y:\|y\|_\infty \leq M} \psi_n(\exp(\bar{h_y})) \tag{4.5}$$
となる．これと $\bar{h_y}$ の y に関する連続性から，最大解が存在する．

以上より，証明が完了する． □

証明から，最尤推定量に関する性質がわかる．まず，対数尤度関数の最大化を $\bar{\mathcal{F}}_0$ 上で考える代わりに，n 次元ユークリッド空間上のベクトル y に関する以下の関数の最大化を考えれば良い：
$$\tilde{\psi_n}(y) = n^{-1} \sum_{i=1}^{n} \bar{h_y}(X_i) - \int_{\mathcal{T}_{p+2}} \exp(\bar{h_y})(x)d\nu(x). \tag{4.6}$$
これは，以下の y に関する凸関数の最小化と等しい：
$$\sigma_n(y) = -n^{-1} \sum_{i=1}^{n} y_i + \int_{\mathcal{T}_{p+2}} \exp(\bar{h_y})(x)d\nu(x). \tag{4.7}$$
σ_n の最小解（$\tilde{\psi_n}$ の最大解）が与えられると，対応する密度の推定量は
$$\hat{f_n}(x) = \exp(\bar{h_y})(x) \tag{4.8}$$
で与えられる．したがって，最尤推定量のサポートは $\mathrm{dom}\ \bar{h_y} = \mathrm{cl}(C_n)$ となる．この性質は，カーネル密度推定の手法とは大きく異なる点である（6 節の数値例も参照）．

最後に，定理 5 の仮定が成り立たず，最尤推定量が存在しない例を示す．

例 6. 図 *3* において，$X_1, X_2, X_3 \in \mathcal{T}_4$ をサンプルとする． 今，図の二つの象限の点の間の測地線は必ずコーンパスになることが保証されている．したがって， X_1, X_2, X_3 の凸包は図の茶色の部分であり，定理 *5* の仮定は満たされていない．今，線分 X_1O と X_2O の長さは等しいとし，また線分 X_3O の長さはそれらの *3* 倍であるとする．$\log f$ を $\log f(X_3) = y + \Delta$, $\log f(X_2) = \log f(X_1) = y - \Delta/3$ となるような線形な関数であるとする．すると $\log f(O) = y$ で，
$$\psi_n(f) = y + \frac{\Delta}{9} - \int_{\mathrm{conv}\{X_1, X_2, X_3\}} f(x)dx. \tag{4.9}$$
このことから，Δ が $+\infty$ になると，第二項は線形に $+\infty$ になり，最後の積分項は *0* に収束することが分かる．したがって，ψ_n は Δ に対して単調増加する関数であり，最大値をとらない．したがって，この場合，最尤推定量は存在しない．

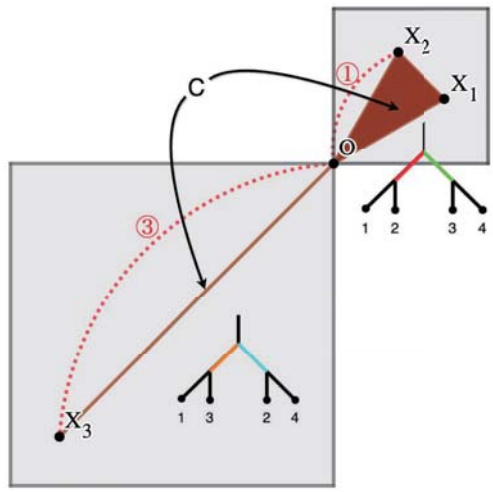

図 3　例 6 におけるサンプルとその凸包．描かれている二つの木は，二つの象限のトポロジーを表す．これらの象限の間の測地線は必ずコーンパスになる．

5　最尤推定量の計算法

4.3 節の結果から，最尤推定量の計算のためには $\sigma_n(y)$ という n 次元凸関数を最小化すれば良い．したがって，$\sigma_n(y)$ の値が各 y について計算できれば，通常の凸関数最小化のアルゴリズムを用いて計算ができる．ただし，$\sigma_n(y)$ の計算のためには，最小の上半連続凹関数 $\bar{h_y}(x)$ を求める必要がある．これは，集合 $S_0 = \{(X_i, y_i) \mid i = 1, \ldots, n\} \subseteq \mathcal{T}_{p+2} \times \mathbb{R}$ の凸包の閉包を求める操作に帰着する．ユークリッド空間の場合，凸包を求めるアルゴリズムは多数存在するが，系統樹空間ではそうでない．以下，1 次元と 2 次元の場合に限って，この問題への一つのアプローチを示す．

5.1　1 次元の場合

1 次元の系統樹空間 $\mathcal{T}_3$ は，3 つの半直線が原点でつながった空間であった．この時，$S_0 = \{(X_i, y_i) \mid i = 1, \ldots, n\} \subseteq \mathcal{T}_3 \times \mathbb{R}$ の凸包は以下のステップで簡単に求めることが出来る．

1. S_0 が $\mathcal{T}_3$ の二つ以上の非負象限に属すかを判断する．属する場合，凸包は必ず $(0, z) \in \mathcal{T}_3 \times \mathbb{R}$ の形の点を含むため，このような点のうち z が最大（最小）になる点を求める必要がある．そのためには，S_0 の，別の象限にある 2 点全ての組み合わせに対し測地線を取り，z が最大（最小）になるものを探せば良い．ここで求めた原点上の点（$(0, z)$ の形の点）を原点での最大点（最小点）と呼ぶことにする．
2. S_0 の点が存在する各象限において，$\mathbb{R}^2$ 上の凸包をとる．この時，先のステップで原点上の点が追加された場合は各象限に原点での最大点，最小点を含めた上で凸包を取ることに注意する．最後に各象限で作られた凸包を合わせることで全体の凸包を得る．

5.2　2 次元の場合

2 次元の場合，すなわち $S_0 = \{(X_i, y_i) \mid i = 1, \ldots, n\} \subseteq \mathcal{T}_4 \times \mathbb{R}$ の凸包の閉包を求める場合は，これを正確に求めることは難しい．Lubiw et al. (2020) は $\mathcal{T}_4$ を含む 1 頂点の 2 次元 CAT(0) 空間における凸包の閉包を線形計画により求めるアルゴリズムを構築したが，今回対象とする空間は $\mathcal{T}_4 \times \mathbb{R}$ であるためこのアルゴリズムをそのまま用いることはできない．ただし，$\mathcal{T}_4 \times \mathbb{R}$ における境界部分での値を繰り返し更新することにより，凸包の近似を求めることが出来る．以下，そのアルゴリズムの概略を説明する．

ここでは原点が凸包内にある場合のみを考える（標本点の凸包が原点を含まない場合は，ユークリッドの場合に帰着する）．ここで，0 は $\mathcal{T}_4$ における原点を表し，$(0, y_l)$ の形の点を横切る測地線もコーンパスと呼ぶことにする．また，$\mathcal{T}_4$ における任意の非負象限 O に対して，$S_0(O) = S_0 \cap (O \times \mathbb{R})$ と定義する．また，Lubiw et al. (2020) の用語にならって，アルゴリズムで帰納的に定義していく $S_l \subseteq \mathcal{T}_4 \times \mathbb{R}$ $(l = 0, 1, 2, \ldots)$ を l 番目のスケルトンと呼ぶことにする．また，$H_l(O) = \mathrm{conv}(S_l(O))$ and $H_l = \cup_O H_l(O)$ とする．以下，S_{l-1}, H_{l-1} からの S_l の構成法を説明する．

第一のステップでは，H_{l-1} の点間の測地線でコーンパスであるものを探し，その原点での値 $(0, y_k)$ を求める．次に，$\{y_k\}$ の最大値，最小値を求め，それぞれ y_{l1}, y_{l0} とする．そして T_l を $S_{l-1} \cup \{(0, y_{l1}), (0, y_{l0})\}$ で初期化する．

第二のステップでは，T_l の全ての 2 点の組み合わせについて，その間の測地線をとり，その境界との交点をすべて T_l に加える．ここで，「境界」とは，$\mathcal{T}_4$ と $\mathbb{R}$ 上の軸が形成する 2 次元平面のことである．実際は得られた各境界上の点全てを保存する必要はなく，各境界上での 2 次元の凸包の頂点をなす点だけを T_l に残し，それ以外の点は T_l から削除する．最後に $S_l = T_l$ としてスケルトンの更新を終える．なお，$H_l(O)$ の計算は $\mathbb{R}^3$ 上の凸包を求める問題であるため，可能であることに注意．

この構成は 2 次元 CAT(0) の場合 (Lubiw et al., 2020) と似ているが，異なる点は，原点での最大値と最小値を最初に決定しなければならないことと，1 次元軸上のでの最大，最小点のみを保存するのではなく，2 次元境界面上の点が生成する凸包の全ての頂点を保存する必要があることである．

以下の定理は，H_l を凸包の近似とすることの正当性を与えるものである．

定理 7. $p, q \in H_{l-1}$ とする．すると，p から q への測地線は H_l に含まれる．さらに，任意の l に対して，$H_l \subseteq \mathrm{conv}A_0$ である．したがって，$\cup H_l = \mathrm{conv}(S_0)$ である．

証明は割愛するが，測地線がコーンパスの場合とそうでない場合に分けて議論すれば良い．コーンパスの場合は定義よりそれが H_l に含まれることが明らかである．コーンパスでない場合は，測地線は必ず隣りあう 3 つの象限の中に埋め込むことが出来るため，ユークリッド幾何を用いて測地線が H_l に入ることを証明することが出来る．

実際は，第一ステップにおいて正確な最大値，最小値を求めることは現実的に困難である．なぜなら，H_{l-1} は有限集合ではなく，さらにこの問題は実は非凸な最適化問題となるからである．ただし，詳細は省略するが，ある線形計画問題を用いて近似値を得ることによって，最大値（最小値）のある下限（上限）を見つけることができる．この近似値で代用することの影響に関しては 5.3 節で議論する．

5.3 アルゴリズムの実装に関する注意点

原点での最大値，最小値を近似値で代用することに関しての理論的な保証はまだないが，サンプル点同士の測地線全ては第二ステップで考慮に入れているため，サンプルサイズが増えるにしたがって原点での最大値，最小値の近似精度が良くなることは期待できる．また，この過程を用いると，H_{l-1} の 2 点間の測地線は，それがコーンパスでない限りは必ず H_l に入ることを保証できるため，全体の凸包の推定に大きくは影響しないと予想される．実際に，6 節で紹介する数値実験においては，集合列 S_l が有限回のイテレーションで収束することが多かった．これらの事実から，原点での最大値や最小値の近似が一定の正当性を持つと思われる．

$\bar{h_y}$ の近似を求めるために集合列 S_l を帰納的に構成するが，実際には S_l が収束することがなくても有限回のイテレーションで打ち切ることが必要である．数値実験の際には，打ち切りの基準として，定数回のイテレーションで打ち切るというルールを使用したが，少なくとも 6 節の数値実験の対象としたデータに関しては目的関数 σ_n の収束性に問題は見られなかった．これは，先にも述べたように有限回で収束する場合が多いこと，また，収束しない場合でも多くの場合最初の数回のイテレーションの後は既存の点に非常に近い点の追加になることがほとんどであることに寄与すると考えられる．このような点は今後の研究で理論的に解析が進むことが望ましい．

本アルゴリズムの欠点は，非常に多くの計算時間を必要とすることである．その要因の一つは，サンプルサイズに比例して解くべき最適化問題の次元が上がることである．これはユークリッド空間の場合と同様の現象であり，ノンパラメトリックに最適化を行っている以上制約の追加等がなければ避けられないものである．一方，系統樹空間特有の現象として，凸包計算に時間がかかることが挙げられる．特にユークリッド空間の場合と比べると，収束までに必要な繰り返しの分だけ多く凸包計算に時間を取る必要があり，実際にはこの部分が時間を消費しているようである．

このように計算時間が長くなってしまう性質から，目的関数の値のみを用いて最小化をするアルゴリズムを用いるのは好ましくない．ユークリッド空間の場合は，関数 σ_n の劣勾配を求めることができた (Cule et al., 2010)．今回の場合も，与えられた $\bar{h_y}(x)$ の近似関数は，ユークリッド空間における最小の凹関数の組み合わせの形になっているため，同様の手順で近似的な劣勾配を求めることが出来る．ただし，その場合，求めた凸包に含まれる $\mathcal{T}_4 \times \mathbb{R}$ の境界上の頂点がどのサンプル点を用いて生成されたかを保存しておく必要がある．

6 数値実験

本節では，数値実験の結果を与える．まず，対数凹密度が 1 次元と 2 次元で正しく推定できることを示す．そして，この結果を Weyenberg et al. (2017) が提案したカーネル密度推定量と比較する．

6.1 1 次元の対数凹密度の推定

1 次元の例として，以下の 2 種類を考える．O_i $(i = 1, 2, 3)$ を $\mathcal{T}_3$ 上の 3 つの象限とする．

ケース 1. 平均 x_0 が象限 O_1 の原点から 1 離れた点で，分散が 1 の正規分布型の密度 f_1: すなわち，

$x \in \mathcal{T}_3$ に対して，$f_1(x) \propto \exp(-d(x, x_0)^2/2)$.

ケース 2. 以下のように定義される指数分布型の分布 f_2:

$$f_2(x) \propto \begin{cases} \exp(-d(x, x_0)) & \text{if } x \in O_2 \cup O_3 \text{ or } (x \in O_1 \text{ and } d(x, 0) < 1) \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

なお，$\mathcal{T}_3$ は原点でつながった 3 本の半直線に過ぎないので，前者については標準正規分布の累積分布関数を使うことを許せば，密度 f_1 と f_2 の正規化定数を計算することは可能である．具体的には，$N(0, 1)$ の累積分布関数を $\Phi(x)$ とすると，f_1 と f_2 の正規化定数 c_1，c_2 は次のようになる．

$$c_1 = \frac{1}{\sqrt{2\pi}(1 + \Phi(-1))}, \tag{6.2}$$

$$c_2 = \frac{1}{1 + e^{-1}}. \tag{6.3}$$

真の密度から 100, 200, 300, 500,1000 個のサンプルをそれぞれ 10 回生成し，積分二乗誤差（ISE）を算出した．図 4 に推定密度の一例を，図 5 に各サンプルサイズにおける平均 ISE を示す．

シミュレーションの結果，対数凹最尤推定量はすべてのサンプルサイズにおいてカーネル密度推定量より優れた推定精度を持つことが示された．

6.2 2 次元の対数凹密度の推定

$\mathcal{T}_4$ の各象限を，図 2 右のピーターセングラフの頂点のインデックスを用いて表す．例えば，軸が 0 と 1 の象限は $\{0, 1\}$ で表される．ここでは，以下の二つの種類の密度を考える．

ケース 3. $\mathcal{T}_4$ 全域をサポートとする，平均 0, 分散 I の正象限に切断された正規分布型の分布 f_3:

$$f_3(x) \propto \exp(-d(x, 0)^2/2). \tag{6.4}$$

ケース 4. $\mathcal{T}_4$ の 6 つの象限（$\{0, 1\}, \{1, 6\}, \{6, 8\}, \{3, 8\}, \{3, 4\}, \{0, 4\}$）のみでサポートされた，平均 0, 分散 I の正象限に切断された正規分布型の分布 f_4:

$$f_4(x) \propto \begin{cases} \exp(-d(x, 0)^2/2) & x \text{ が上の 6 つの象限のいずれかに入っている} \\ 0 & \text{otherwise.} \end{cases} \tag{6.5}$$

2 次元の場合，これらの密度の正規化定数は解析的に計算することができる．具体的には，正規化定数をそれぞれ c_3 と c_4 とすると，次のようになる．

$$c_3 = \frac{1}{2\pi} \times \frac{4}{15} = \frac{2}{15\pi}, \tag{6.6}$$

$$c_4 = \frac{1}{2\pi} \times \frac{4}{6} = \frac{1}{3\pi}. \tag{6.7}$$

また，系統樹空間の非正曲率の性質から，これらの密度は対数凹であることもわかる．

この対数凹最尤推定量とカーネル密度推定量 (Weyenberg et al., 2017) を，真の密度に対する ISE の観点から比較する．真の密度から 50，100，200，300，500 個のサンプルを生成し，最尤推定量並びにその ISE を 10 回計算し，それぞれのサンプルサイズで，平均 ISE を計算した．その推定結果を図 6 に示す．

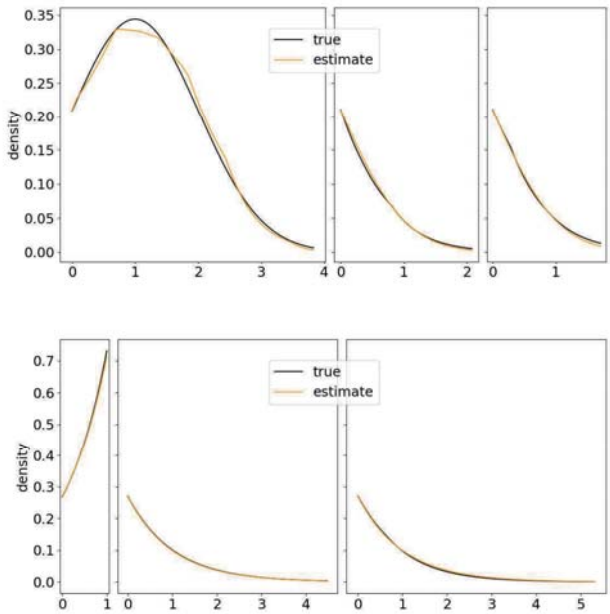

図 4　ケース 1（上）とケース 2（下）における対数凹最尤推定量の例と真の密度（$n = 1000$）.

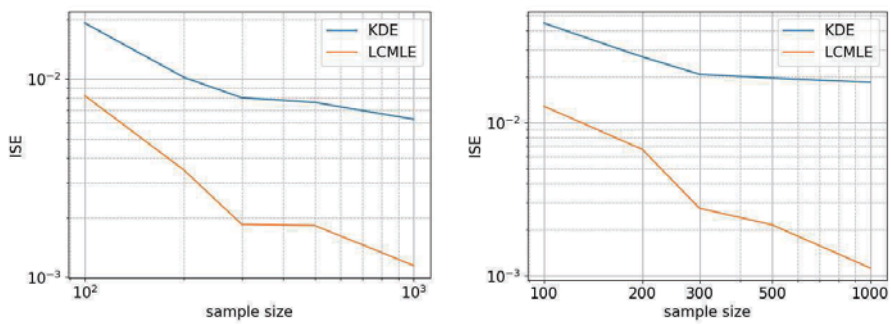

図 5　サンプルサイズ 100, 200, 300, 500, 1000 における平均積分二乗誤差.（左）：ケース 1,（右）：ケース 2. LCMLE は対数凹最尤推定量，KDE はカーネル密度推定量を表す.

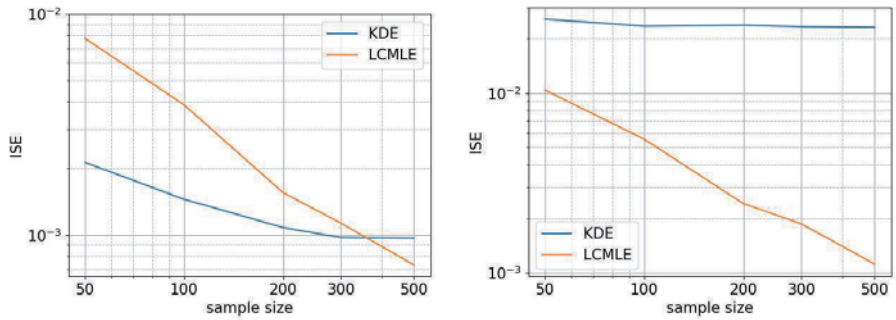

図 6　サンプルサイズ 100, 200, 300, 500 における平均積分二乗誤差．（左）：ケース 3，（右）：ケース 4．LCMLE は対数凹最尤推定量，KDE はカーネル密度推定量を表す．

ケース 3 では，小さいサンプルサイズではカーネル密度推定の方が性能が良いが，サンプルサイズが大きくなると対数凹最尤推定量がカーネル密度推定より性能が良くなり始める．この結果は，ユークリッド空間 (Cule et al., 2010) と同様である．一方，ケース 4 では，対数凹最尤推定量がカーネル密度推定量よりも全てのサンプルサイズにおいて優れている．これは，対数凹最尤推定量は十分なデータの下，真の密度のサポートに近づけることができるが，カーネル密度推定量ではできないためである．

7　まとめ

本稿では，ある条件下で対数凹最尤推定量が系統樹空間上に存在することを示した．1 次元の場合，最尤推定量は確率 1 で存在し，一意である．これはユークリッドの場合と同じ結果である．多次元では，最尤推定量が存在しない場合もある．しかし，十分条件が成立すれば最尤推定量が存在し，測度-a.e. で一意であることを示すことができた．また，1 次元の場合は正確に，2 次元の場合は近似的に凸包を計算することで，最尤推定量を計算するアルゴリズムを示した．先に開発したカーネル密度推定量と比較した結果，十分大きなサンプルサイズを持ちモデルが正しいケースでは，我々の推定量が優位に立つことが確認された．

ここで導かれた方法は，系統樹空間上の密度推定に新しいノンパラメトリックアプローチを与えるという点で意義がある．また，2 次元の例で見たように，真の分布がある凸集合であればそのサポートを尊重することができることが大きな特徴である．これは，数値実験で見たように，状況によっては推定精度にかなりの影響を与える可能性がある．実際の系統樹のデータにおいても，推測される可能性のあるトポロジーや内部辺の長さにはある程度制約があることが予想される．

また，カーネル密度推定とは異なり，平滑化パラメータの決定や正規化定数の計算が必要ない点も本手法のメリットである．計算時間はカーネル密度推定よりはるかに遅いが，対数凹の仮定が真の密度の性質から大きく離れていない場合には，密度推定の精度は高くなることが期待できる．開発した手法は，精度の向上が困難な完全無制約ノンパラメトリックアプローチと，系統樹空間に対する正しいモデルの指定が困難なパラメトリックアプローチの中間的な選択肢を与えるものである．

今後の研究課題として，系統樹空間における推定量の理論的特性を導き出すことが重要である．ユークリッド空間では，対数凹最尤推定量は一致性を持つことが知られており，モデルの指定が誤っていても，真の密度の「対数凹射影」（真の密度からのカルバック・ライブラーダイバージェンスを最小化する対数凹密度）に収束することが知られている．これらの性質が系統樹空間でも成り立つかどうかを調べることは非常に重要である．また，対数凹の仮定が成り立たない場合においての数値的なパフォーマンスの確認をさらに行うことも，実用上必要である．

系統樹の空間をモデル化するために系統樹空間が構築されたので，他の研究課題として，生物学的データや問題への適用性が挙げられる．本稿では最大 4 つの分類群が存在する場合に対応する，1 次元と 2 次元の計算アルゴリズムのみを与えたため，今のところ適用可能性は限られているように思われる．しかし，例えば，複数の木の不整合性の観点から，一部の分類群を無関係とすることで，大きな木の低次元表現を求めることが可能かもしれない．また，高次元データに関しては計算時間や収束速度の観点からも同じフレームワークを用いることは現実的ではないため，分布の形により多くの制約を導入するという方向性もあるように思われる．このような修正を加えた上で本手法が，生物データを用いた既存の手法と比較して，どのようなパフォーマンスを示すかは興味深いところである．また，データ分布に関する対数凹性の仮定を慎重に評価することも興味深い課題である．

参考文献

Bačák, M. (2013), ‘The proximal point algorithm in metric spaces’, *Israel Journal of Mathematics* **194**(2), 689–701.

Bačák, M. (2014), *Convex analysis and optimization in Hadamard spaces*, Walter de Gruyter GmbH & Co KG.

Benner, P., Bačak, M. and Bourguignon, P. Y. (2014), ‘Point estimates in phylogenetic reconstructions’, *Bioinformatics* **30**(17), i534–i540.

Billera, L. J., Holmes, S. P. and Vogtmann, K. (2001), ‘Geometry of the space of phylogenetic trees’, *Advances in Applied Mathematics* **27**(4), 733–767.

Cule, M., Samworth, R. and Stewart, M. (2010), ‘Maximum likelihood estimation of a multidimensional log-concave density’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **72**(5), 545–607.

Felsenstein, J. (1978), ‘The Number of Evolutionary Trees’, *Systematic Zoology* **27**(1), 27–33.

Felsenstein, J. (2004), *Inferring phylogenies*, Vol. 2, Sinauer associates Sunderland, MA.

Lubiw, A., Maftuleac, D. and Owen, M. (2020), ‘Shortest paths and convex hulls in 2D complexes with non-positive curvature’, *Computational Geometry: Theory and Applications* **89**(0), 1–42.

Nye, T. M. (2011), ‘Principal components analysis in the space of phylogenetic trees’, *Annals of Statistics* **39**(5), 2716–2739.

Owen, M. and Provan, J. S. (2011), ‘A fast algorithm for computing geodesic distances in tree space’, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**(1), 2–13.

Rockafellar, R. T. (1970), *Convex analysis*, number 28, Princeton university press.

Weyenberg, G., Huggins, P. M., Schardl, C. L., Howe, D. K. and Yoshida, R. (2014), 'kdetrees: non-parametric estimation of phylogenetic tree distributions', *Bioinformatics* **30**(16), 2280–2287.

Weyenberg, G., Yoshida, R. and Howe, D. (2017), 'Normalizing Kernels in the Billera-Holmes-Vogtmann Treespace', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **14**(6), 1359–1365.

Willis, A. (2019), 'Confidence Sets for Phylogenetic Trees', *Journal of the American Statistical Association* **114**(525), 235–244.