系統樹の空間におけるモード推定とその性質

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

系統樹の内部枝長とトポロジーが埋め込まれた距離空間である BHV tree space では，フレシェ平均やフレシェメジアンは空間の象限の境界に貼り付きやすいという性質が知られている．本研究では，このような性質を持たないモードを推定する問題を考え，4 つの推定量を与える．このうち 2 つは密度推定を伴う間接的な推定量であり，残りはそうでない直接的な推定量である．これらの推定量の真のモードへの一致性と，直接的な推定量のロバストネスの性質を示し，数値実験において精度を検証する．

1　はじめに

系統樹は生物の進化の歴史を表す基本的なモデルである．数学的には，葉にラベルがついた有根または無根の木構造として捉えることができる．以下，有根の系統樹のみを考えるが，無根の系統樹の場合は一つの葉を根として考えることで有根の場合に帰着させることができる．

系統樹の内部頂点はある共通祖先に対応している．系統樹自体を観測することは原則できない．現代では，現存種（葉）の遺伝子やアミノ酸配列のアラインメントをデータとし，木に従う進化のモデルを仮定することで最尤法やベイズ法などでその推定を行うことができる．詳しくは，例えば Kapli et al. (2020) を参照されたい．

本研究では，系統樹の集合が与えられた時，その中心を求めるという問題を考える．例えば，ベイズ法で系統樹を求めた場合には，マルコフ連鎖モンテカルロ法で得られた木のサンプルの代表点を作るという用途が考えられる．また，推定されたいくつかの遺伝子樹から種の系統樹を求めるという問題もこのような枠組みに収まる．古典的な方法としては，合意樹の形成が挙げられ，特に多数決合意樹はその簡易な計算方法や直感的な定義，また Robinson-Foulds 距離に対するメドイドになっている (Holder et al., 2008) という点で多く用いられる．一方，系統樹間の距離には様々なものが提案されており，特に Billera et al. (2001) が提案した BHV tree space は木のトポロジーと内部枝の長さの情報を埋め込んだ距離空間である．この空間は全域で非正の曲率を持つ（CAT(0) 性）という性質を持っていることが知られており，近年ではこの空間上での統計的手法の開発がなされている．本研究では，系統樹の集合をこの BHV tree space 上のサンプルと捉えて統計解析をすることを考える．

BHV tree space 上での自然なサンプルの代表点としては，算術平均の一般化に対応するフレ

シェ平均や幾何メジアンに対応するフレシェメジアンがある．特に有限サンプルのフレシェ平均はCAT(0) 性から一意に定まることが知られている．一方で，フレシェ平均の性質として，stickinessというものが知られている．これはフレシェ平均が BHV tree space の各象限の境界面に張り付きやすいという性質であり，木の形の意味では多分岐を含む木となりやすいという意味である．このstickiness の問題を解決するため，本研究では中心の推定問題としてモード推定を考え，4 つの推定量を与える．またそれらの推定量の理論的な性質について考察する．

構成は以下のとおりである．まず 2 節では系統樹空間の定義と，stickiness の現象について説明する．3 節で 4 つのモード推定量を定義し，4 節では 1 次元の系統樹空間上での数値実験の結果を示す．最後に 5 節でまとめをする．

2　系統樹空間と Stickiness

2.1　系統樹空間

まず，N 葉の有根系統樹の BHV tree space $\mathcal{T}_N$ (Billera et al., 2001) の定義をする．BHV tree space では，系統樹がそのトポロジーと内部枝の長さで規定されると考える．内部枝以外の長さも後にこの空間 $\mathcal{T}_N$ と N 次元ユークリッド空間との直積空間を考えることで埋め込むことができるが，本稿では考えない．また，無根系統樹はその葉の一つを根としてとることにより，有根系統樹の場合に帰着させることができる．

N 葉の二分木のトポロジーは $(2N-3)!!$ 個であることが知られている．また，N 葉の二分木は $N-2$ 本の内部枝を持つ．ここで，内部枝とは根や葉に直接接続していない枝のことである．そこで，まずある固定された二分木のトポロジーを持つ木の内部枝 $N-2$ 本の長さがなす非負ベクトルと，$N-2$ 次元ユークリッド空間の非負象限 $\mathbb{R}^{N-2}_{\geq 0}$ 上の点を対応させる．すると，ある二分木のトポロジー i を持つ木全体のなす空間 $\mathcal{O}_i$ は，$\mathbb{R}^{N-2}_{\geq 0}$ のコピーとして表すことができる．特に，非負象限の境界においてはいくつかの枝長が 0 になるため，境界は非二分木のトポロジーの空間に対応していることに注意する．このような非二分木のトポロジーはいくつかの二分木の象限の境界として現れるため，共通の境界を用いて各象限を貼り合わせることができる．このようにしてできた全体の空間が N 葉の BHV tree space $\mathcal{T}_N$ である．$N=3$ の場合，二分木のトポロジーは 3 個であり，二分木に含まれる内部辺は 1 本であるため，BHV tree space $\mathcal{T}_3$ は 3 本の半直線が原点で繋がったような空間となる（図 1）．このような空間を 3-spider とも呼ぶ．4 葉以上の場合は多次元のユークリッド空間の非負象限が複雑に繋がった形をしている．詳細は Billera et al. (2001) などを参照せよ．

$\mathcal{T}_N$ は，ユークリッド空間の非負象限の貼り合わせであるため，ユークリッド距離を用いて距離空間と見ることができる．具体的には，同象限の 2 点間の距離は通常のユークリッド距離で定義し，別の象限の 2 点間の距離は 2 点を結ぶパスの最小の長さで定義する．ただしここで，2 点を結ぶパスの長さは，その各象限内への制限の長さの和で定義する．このようにしてできた距離空間 $(\mathcal{T}_N, d)$ は曲率が非正であるという性質（CAT(0)) を持つことが知られている (Billera et al., 2001)．CAT(0) 性は任意の 2 点間の測地線の一意存在性を担保し，これによって例えば測地凸性の定義が可能である．

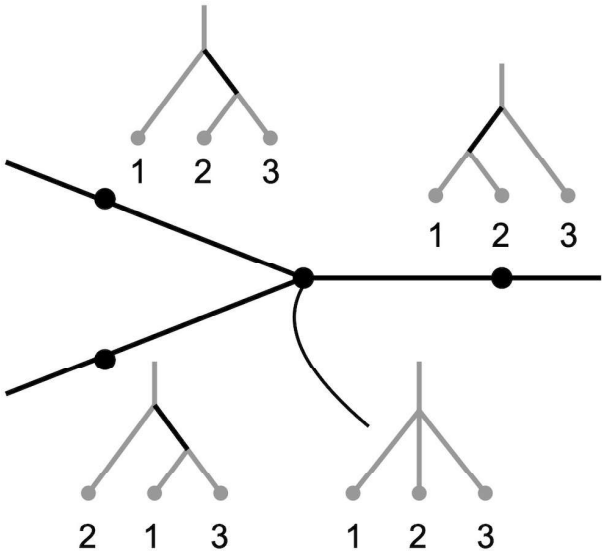

図 1　3 種の系統樹空間（3-spider）

2.2　フレシェ平均，フレシェメジアンと Stickiness

距離空間 $(\mathcal{M}, d)$ 上のサンプル $\mathcal{X} = \{X_1, \ldots, X_n\}$ が与えられたとする．フレシェ平均とは次のように定義される量である：

$$\mu(\mathcal{X}) \coloneqq \arg\min_{X} \sum_{i=1}^{n} d(X, X_i)^2. \tag{2.1}$$

同様に，フレシェメジアンは次のように定義される：

$$m(\mathcal{X}) \coloneqq \arg\min_{X} \sum_{i=1}^{n} d(X, X_i). \tag{2.2}$$

一般の距離空間においては，これらの存在性や一意性は必ずしも保証されない．一方で，CAT(0) 空間上のサンプルであればフレシェ平均は一意存在性を持ち，フレシェメジアンは（一意とは限らないが）存在することが知られている．

フレシェ平均やフレシェメジアンは直感的で自然なサンプルの代表点となるが，これらは stickiness という性質を持つことが知られている (Barden et al., 2018)．Stickiness とは，系統樹空間の境界に貼り付きやすいという性質である．以下はその例である．

例 1. 3-spider 上の 3 点のサンプル $\mathcal{X} = \{X_1, X_2, X_3\}$ が図 2 左のように与えられたとする．このサンプルのフレシェ平均やフレシェメジアンは原点 O である．一方，サンプルの一つ X_1 をさらに原点から 1 だけ遠くした場合（図 2 右）でもフレシェ平均やフレシェメジアンは原点 O に留まり続ける．特に，フレシェメジアンは X_1 を原点からどれだけ離しても原点 O に留まり続ける．

このような現象は有限サンプルの場合だけでなく，連続な分布に関しても起こる．ある距離空間上

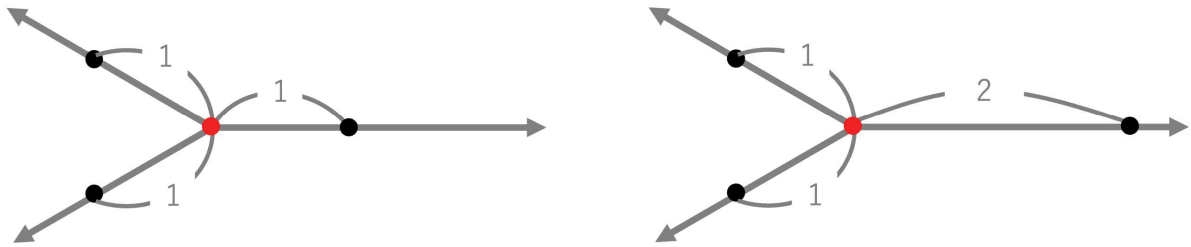

図 2　例 1 における，3-spider 上の 3 点 X_1, X_2, X_3．赤点は 3 点のフレシェ平均並びにフレシェメジアンを表す．

の確率分布 P に対して，そのフレシェ平均とフレシェメジアンは同様に次のように定義される：

$$\mu(P) := \arg\min_X \int d(X,Y)^2 dP(Y), \tag{2.3}$$

$$m(P) := \arg\min_X \int d(X,Y) dP(Y). \tag{2.4}$$

距離空間が CAT(0) であり，可積分性の条件 $\int d(X,Y)^2 dP(X) < \infty$ が成り立てば，フレシェ平均は一意に存在する[*1]．フレシェメジアンは同じ条件下で一意とは限らないが存在性がわかる．以下は，ある 3-spider 上の分布に対する stickiness の例である．

例 2. 合祖過程は，集団遺伝学における遺伝子進化の最も基本的なモデルである．標準的な合祖過程においては，時間を逆向きに見た際に，2 つの系統が共通祖先へ「合祖」するまでの時間（合祖時間）を指数分布でモデルする．このモデルを多種の場合に拡張し，ある種の系統樹が与えられた下での遺伝子系図の分布を求める問題に応用したものが多種合祖モデル（Multispecies Coalescent Model）(Rannala and Yang, 2003) である．

多種合祖モデルでは，種の系統樹上で複数の系統が共存する集団において標準的な合祖モデルを適用する．詳細は割愛するが，3 種のある二分系統樹で，内部枝の長さ T を持つようなものを固定し，各種に対し，1 つの系統をサンプルする．この時，3 つの系統がなす遺伝子樹の分布は，3-spider 上で次のように表される：3-spider の半直線 j $(j = 1, 2, 3)$ 上の遺伝子樹の確率密度関数 f_j は，

$$f_j(t;T) = \begin{cases} -\frac{1}{6}e^{-t-T} + \frac{1}{2}e^{-|T-t|} & \text{if } j = 1 \\ \frac{1}{3}e^{-T-t} & \text{if } j \in \{2,3\}. \end{cases} \tag{2.5}$$

ただし，$j = 1$ の半直線に対応するトポロジーが種の系統樹のトポロジーを表しているものとした．式 (2.5) より，一定の確率で種の系統樹と別のトポロジーの遺伝子系図が観測されることがわかる．また，その確率は内部枝の長さ T が小さいほど大きくなる．

次の条件が満たされる時，この分布に対するフレシェ平均は原点となる：

$$-\frac{1}{3}e^{-T} + T \leq 0. \tag{2.6}$$

[*1] 実際は，フレシェ平均の定義を修正することで 1 次の可積分性の元で一意存在性が言える (Sturm, 2003).

また，フレシェメジアンは次の条件が満たされるときに原点となる：

$$T \leq \log \frac{4}{3}. \tag{2.7}$$

例 2 のように，フレシェ平均やフレシェメジアンが推定したい木へ一致性を持たない場合がある．本研究では，このような問題への対処として，モード推定の問題を考える．

3　4 つのモード推定量とその性質

本節では，系統樹空間上のモード推定の手法を 4 つ提案する．モード推定量は，密度推定を伴う間接的なものと，そうでない直接的なものに分類することができる．本研究で与える初めの 2 つの推定量は直接的なものであり，多次元ユークリッド空間におけるモード推定の手法を系統樹空間へ拡張したものとなっている．残りの 2 つは間接的なものであり，カーネル密度推定と対数凹密度推定を用いたものである．

以下，この 4 つの推定量それぞれの定義を説明し，その一致性の性質について論じる．また，直接的な推定量に関してはロバストネスの性質として Finite Sample Breakdown Point (FSBP) について考察する．FSBP は次のように定義される：

定義 3. ある距離空間 (M, d) 上のサイズ n のサンプル $\mathcal{X} = \{X_1, \ldots, X_n\}$ に対して，そのうちの m 個だけを任意に変更してできる集合の族を $\mathcal{Y}_m$ とおく．この時，ある推定量 $\hat{\theta}(\mathcal{X})$ の（サンプル $\mathcal{X}$ に対する）Finite Sample Breakdown Point $\varepsilon(\hat{\theta}, \mathcal{X})$ は次の量である：

$$\varepsilon(\hat{\theta}, \mathcal{X}) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} \middle| \sup_{\mathcal{Y} \in \mathcal{Y}_m} d(\theta(\hat{\mathcal{X}}), \theta(\hat{\mathcal{Y}})) = \infty \right\}. \tag{3.1}$$

これは元のサンプルのうちどの割合の観測を汚染すれば推定量を任意に悪くできるかということを示す指標であり，よって小さい値を取るとそのような意味におけるロバスト性がないと解釈できる．例えばユークリッド空間においての算術平均の FSBP は $1/n$ である．

3.1　共通する設定

p 次元の系統樹空間 $\mathcal{T}_{p+2}$ 上の観測 $\mathcal{X} = \{X_1, \ldots, X_n\}$ は，ある連続な密度 f からの独立サンプルであるとする．さらに，この密度 f は有界であり，一意なモード θ を持つとする．ここでのモード θ は，任意の $x \neq \theta$ に対して，$f(\theta) > f(x)$ を満たすようなものと定義する．従って，多峰の分布であっても密度が最大となる点が一意に存在すれば，一意なモードを持つこととなる．さらに，病的な状況を除くため，次を仮定する：任意の $\varepsilon > 0$ に対し，ある $\delta > 0$ が存在し，$d(x, \theta) > \varepsilon$ の時 $f(x) + \delta < f(\theta)$.

3.2　直接的なモード推定量

3.2.1　Minimum Volume Ball モード推定量

まず，Minimum Volume Ball 推定量の定義をする．n に依存する正の整数 $r(n) < n$ をとる．この時，観測 $X_1, \ldots, X_n$ に対して，そのうち少なくとも $r(n)$ 個を含む球で最も体積の小さいものを

選ぶ．この球を S_n とした時，その中から一つ任意に代表点を取り，これを推定量 $\hat{\theta}_n^{\mathrm{MVB}}$ とする．代表点の取り方としては例えば S_n に含まれるサンプルのフレシェ平均やフレシェメジアンを取ることが考えられる．

計算の観点においては，BHV tree space 上では，球の体積が半径のみで決まらず，中心に依存する問題がある．2 次元程度までであれば，計算は正確にすることが可能である一方，3 次元以上の場合は凸集合の体積の近似アルゴリズム等を用いて計算をする必要がある．さらに，ユークリッド空間の場合においても，r 点以上含む最小体積の球を見つける問題は一般に難しいことが知られている (Shenmaier, 2013)．次節の数値実験では，サンプル点を中心とする球 n 個のなかで最小のものを選んでいる．また，代表点の選び方はフレシェ平均を採用している．

推定量の一致性は，選ぶ点の個数 $r(n)$ を適当に取ることによって，代表点の選び方によらずに保証することができる．

定理 4 (MVB モード推定量の一致性)**.** 列 $r(n)$ は $r(n)/n = o(1)$, $\sqrt{n \log n}/r(n) = o(1)$ を満たすとする．この時，$\hat{\theta}_n^{\mathrm{MVB}} \to \theta$ a.s..

定理の条件は，例えば $r(n) = O(n^{\alpha})$ $(1/2 < \alpha < 1)$ などと選ぶことで満たされる．

FSBP は次のようになる．

定理 5 (MVB モード推定量の Finite Sample Breakdown Point)**.** 確率 1 で，

$$\varepsilon(\hat{\theta}_n^{\mathrm{MVB}}, \mathcal{X}) \geq \min\left\{\frac{n - r(n)}{n}, \frac{r(n) - 1}{n}\right\}. \tag{3.2}$$

また，フレシェ平均を代表点として選んだ時，式 (3.2) は等号で成立する．

3.2.2 Minimum Volume Convex Peeling モード推定量

Minimum Volume Convex Peeling 推定量は，次のように計算される．まず，ある定数 $q \in (0, 1)$ を取り，$r(n) : \mathbb{N} \to \mathbb{N}$ を，$r(n)/n \to q$ $(n \to \infty)$ が成り立つように取る．例えば，$r(n) = \lfloor nq \rfloor$ などとすれば良い．また，$\sigma(n) : \mathbb{N} \to \mathbb{N}$ を $p + 1 < r^{\sigma(n)}(n) \coloneqq r(r^{\sigma(n)-1}(n)), \sigma(n) \to \infty$ $(n \to \infty)$ を満たすようにとる．この時，観測 $X_1, \ldots, X_n$ に対して，そのうち少なくとも $r(n)$ 個を含む凸包の中で最も体積の小さいものを選び，これを $M_{1,n}$ とする．同様に，$M_{m,n}$ は，$M_{m-1,n}$ 内の観測点のうち，少なくとも $r^m(n)$ 個を含む凸包（$\mathcal{C}_{m,n}$ の要素）の中で最も体積の小さいものとする．最後に，$M_{\sigma(n),n}$ の中から代表点を選び，これを MVCP モード推定量 $\hat{\theta}_n^{\mathrm{MVCP}}$ とする．代表点は，例えば $M_{\sigma(n),n}$ に含まれるサンプル点のフレシェ平均やフレシェメジアンを採用できる．MVB モード推定量との違いは，基本的な集合族として閉凸包全体を用いていることと，繰り返し最小体積の凸包を求めるところである．

計算の観点においては，まず BHV tree space 上の閉凸包の計算の難しさが挙げられる．BHV tree space では，2 次元の場合までのみ多項式時間での計算アルゴリズムが知られており，3 次元以上での計算は難しい．また，閉凸包の計算ができたとしても，最小の体積のものを見つける問題は MVB モード推定量同様に難しく，ユークリッド空間上においてはヒューリスティックなアルゴリズムが提案されている．次節における数値実験では，このアルゴリズムをさらに簡略化したものを BHV tree

space 上で適用することで対処している．また，代表点としてはフレシェ平均を用いた．

推定量の一致性は，密度のレベルセットが凸集合になっているという仮定を追加することで得られる．

定理 6 (MVCP モード推定量の一致性)**.** 密度 f は次を満たすとする：$0 \leq h \leq f(\theta)$ に対し，$I(h) = \{x \mid f(x) \geq h\}$ は閉凸集合であり，等高線 $\{x \mid f(x) = h\}$ はその境界であるとする．また，$0 \leq h < f(\theta)$ において，$I(h)$ は以下の条件を満たすとする：

- 任意の点 $x \in I(h)$ に対して，系統樹空間 $\mathcal{T}_{p+2}$ の非負象限 $\mathcal{O}_i$ が存在し，$x \in \mathcal{O}_i$ かつ $\nu(x \cap \mathcal{O}_i) > 0$（ただし，$\nu$ は $\mathcal{O}_i$ 上のルベーグ測度）．

この時，$\hat{\theta}_n^{\mathrm{MVCP}} \to \theta$ a.s..

ユークリッド空間の場合，$q = 1/2$ と取ることで，FSBP は n によらず 1/2 に近い値に抑えられることが知られている (Kirschstein et al., 2016)．一方，BHV tree space においては，FSBP が低くなる一般的な点配置が存在する．

定理 7 (MVCP モード推定量の Finite Sample Breakdown Point)**.** $p \geq 2$ の場合を考える．代表点としてフレシェ平均を取った場合の *MVCP* モード推定量 $\theta_n^{\hat{\mathrm{MVCP}}}$ に対し，ある一般的な点配置 $\mathcal{X}$ が存在し，$\varepsilon(\hat{\theta}, \mathcal{X}) = 1/n$ とできる．

以下はその例である．

例 8. 2 次元の系統樹空間 $\mathcal{T}_4$ を考える．$\mathcal{T}_4$ は 15 個の 2 次元の非負象限が複雑につながった形をしている．ここで，2 つの象限のペア $\mathcal{O}_1, \mathcal{O}_2$ であって，それぞれの象限上の点を結ぶ測地線（最短パス）が必ず原点を通るようなものが存在する（図 3 左）．今，与えられた $q, r(n), \sigma(n)$ に対し，次の条件を満たすような n 点の点配置を考える：

- 象限 $\mathcal{O}_1$ に 2 つの点 (X_1, X_2 とする) があり，象限 $\mathcal{O}_2$ にそれ以外の点がある．
- X_1, X_2 が $M_{\sigma(n),n}$ に含まれる．

この時，X_1 または X_2 を，三角形 OX_1X_2 の面積を保ちながら同象限内の原点 O から十分離れた点に動かすことによって，汚染されたサンプルでのモード推定量は元の推定量から任意に遠くにできる（図 3 右）．

3.3 間接的なモード推定利用

本小節では，密度推定を伴う間接的なモード推定量を考える．間接的なモード推定量の計算アルゴリズムは，以下の 2 ステップで構成される：

1. サンプルから密度推定量 $\hat{f}$ を構成し，
2. $\hat{f}$ のモードをモード推定量 $\hat{\theta}$ とする．

以下，最初のステップである密度推定に対し，カーネル密度推定を用いる手法 (KDE モード推定

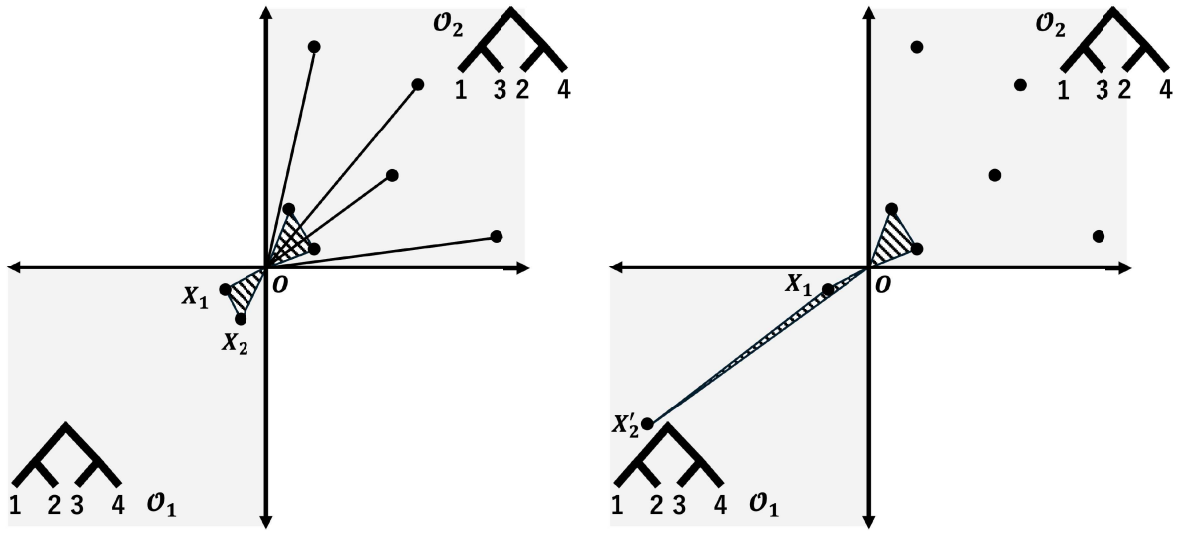

図 3　（左）：象限をまたぐすべての測地線が原点を通る 2 つの象限の例．斜線部は，描かれた 8 点のうち 4 点を含む凸集合のうち最も体積が小さいものを示す．（右）：凸包の面積を保ちながら X_2 を原点から離れた場所（X_2'）へ移動させる例．

量) と，対数凹最尤推定を用いる手法 (LCMLE モード推定量) を与える．

3.3.1　KDE モード推定量

BHV tree space 上におけるカーネル密度推定量 (Kernel Density Estimator, KDE) は，Weyenberg et al. (2014, 2017) で提案されている．カーネル密度推定量は次のように定義される：

$$\hat{f}_n^{\mathrm{KDE}}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x; X_i, h_n). \tag{3.3}$$

ただし，k はカーネル関数で，これらの文献では，次の形の正規型のカーネルを考えている：

$$k(x; y, h) = C(y, h)\exp\left(-\frac{d(x,y)^2}{2h^2}\right). \tag{3.4}$$

ここで，$C(y,h)$ は正規化定数であり，以下で定義される：

$$C(y,h) = \left(\int \exp\left(-\frac{d(x,y)^2}{2h^2}\right) dx\right)^{-1}. \tag{3.5}$$

正規化定数がバンド幅 h のみならず，中心 y に依存していることに注意されたい．ここで，次のような少し異なるカーネルを考えることも可能である：

$$k(x; y, h) = C(x, h)\exp\left(-\frac{d(x,y)^2}{2h^2}\right). \tag{3.6}$$

式 (3.6) で定義されるカーネル関数を用いた場合，カーネル密度推定量 (3.3) は x に関する密度関数となっているとは限らないし，各点での密度推定量の値を求めるために毎回正規化定数の計算が必要となる．多次元で正規化定数を求めることは容易ではなく，Weyenberg et al. (2017) では，あ

る正規化定数の下限を用いることで対処している．その一方で，式 (3.6) による定式化は後の一致性の性質を確認することを容易にする．本稿では式 (3.6) で定義されるカーネル関数を考えることとする．また，数値実験では，Weyenberg et al. (2014) に従って，経験上パフォーマンスの良い適応的なバンド幅選択手法を用いている．

KDE モード推定量は $\hat{\theta}_n^{\mathrm{KDE}} \in \arg\max_x \hat{f}_n^{\mathrm{KDE}}(x)$ として選ばれる．一方，実用的にはこの計算も難しい場合があるため，サンプル点の中で最も密度が高いものを選ぶことが考えられる．次節における数値実験ではこの方法を採用している．

カーネル密度推定量及び KDE モード推定量の一致性は，密度 f の一様連続性の仮定と式 (3.6) の形のカーネルを用い，バンド幅を適切に選ぶことによって達成される．

定理 9 (KDE モード推定量の一致性)**.** 式 (3.3), (3.6) で定義されるカーネル密度推定量に対し，バンド幅 h_n が次の条件を満たすとする：

- $h_n \to 0$
- $\frac{nh_n^d}{|\log h_n|} \to \infty$
- $\frac{\log\log n}{|\log h_n|} \to \infty$
- ある $c > 0$ に対し，$h_n^d \leq c h_{2n}^d$.

さらに，密度 f は一様連続であるとする．この時，$\sup_x |\hat{f}_n^{\mathrm{KDE}}(x) - f(x)| \to 0$ a.s. であり，$\hat{\theta}_n^{\mathrm{KDE}} \to \theta$ *a.s.*

3.3.2 LCMLE モード推定量

BHV tree space 上における対数凹最尤推定量 (Log-Concave Maximum Likelihood Estimator, LCMLE) は Takazawa and Sei (2024) によって提案されている．密度 f が対数凹であるとは，$\log f$ が凹関数となることであり，このクラス内で最尤推定が可能である．対数凹最尤推定量 $\hat{f}_n^{\mathrm{LCMLE}}$ に対し，LCMLE モード推定量は $\hat{\theta}_n^{\mathrm{LCMLE}} \in \arg\max_x \hat{f}_n^{\mathrm{LCMLE}}(x)$ として選ばれる．現在，対数凹最尤推定量は 2 次元以下の場合のみ計算が可能であり，モード推定量も同じ制約を受ける．これは，BHV tree space 上の閉凸包計算の難しさに起因する．一方で，推定量 $\hat{f}_n^{\mathrm{LCMLE}}$ が得られれば，サンプル点のいずれかで必ず密度が最も大きくなることが保証される．この性質によって，密度推定量のモードを取ることは容易である．

対数凹最尤推定量並びに，LCMLE モード推定量の一致性は，密度 f の対数凹性の仮定の下で成り立つ．

定理 10 (LCMLE モード推定量の一致性)**.** 密度 f は対数凹であるとする．この時，

$$\sup_x |\hat{f}_n^{\mathrm{LCMLE}}(x) - f(x)| \to 0 \ \ a.s.$$

さらに，$\hat{\theta}_n^{\mathrm{LCMLE}} \to \theta$ *a.s.* である．

対数凹性の仮定が破られると，このような一致性は成り立たない．一方で，そのような場合でも，適当な可積分性の条件の下で密度 f の対数凹射影と呼ばれる量への対数凹最尤推定量の一致性が保

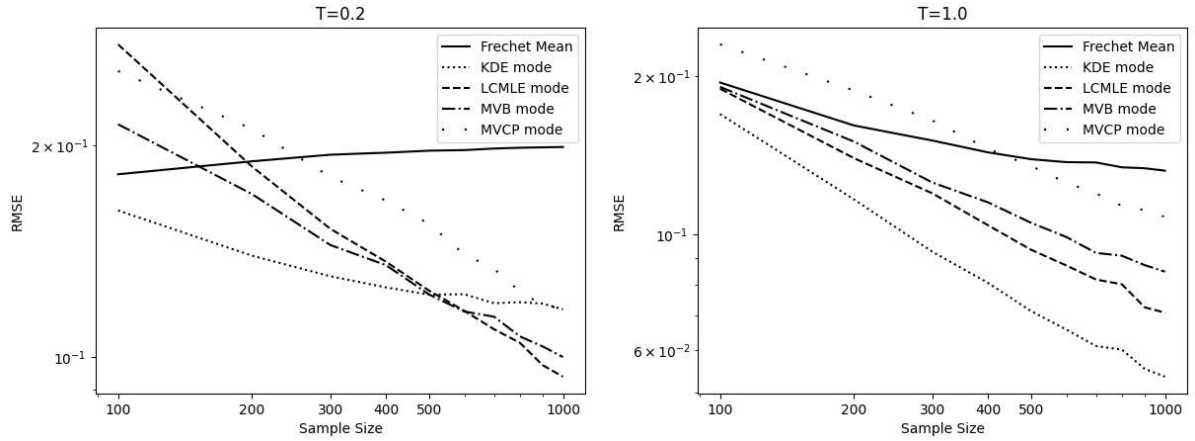

図 4　各モード推定量・フレシェ平均の真のモードに対する RMSE の平均をサンプルサイズごとに示した両対数グラフ．（左）：$T = 0.2$ の場合，（右）：$T = 0.1$ の場合．

証される．ただし，一般に対数凹射影のモードと元の密度のモードは異なるため，LCMLE モード推定量の一致性は言えないことに注意する．

4　数値実験

本節では，モード推定量の有効性を示す簡単な数値実験の結果を 1 次元系統樹空間の場合で与える．真の密度として，例 2 で挙げた多種合祖過程に基づいた遺伝子樹の分布を考え，このような分布からサンプルされた遺伝子樹から，真の遺伝子樹 T を求める問題を考える．すなわち，サンプルされた遺伝子樹 $t_1, \ldots, t_n$ に対してモード推定を行うことを考える．

例 2 にも述べたように，このような場合に真の分布のフレシェ平均は T とは一致しない．特に，条件 (2.6) を満たす際には，真のフレシェ平均は原点となる．このような性質に留意した元で，T の推定量としてのサンプルフレシェ平均とモード推定量の精度を二乗平均平方根誤差（RMSE）の意味で比較する．

フレシェ平均が sticky な場合とそうでない場合での精度を評価するために，次の 2 つの設定で数値実験を行った．

1. 真の系統樹の内部辺の長さが $T = 0.2$ の時（sticky な場合）
2. 真の系統樹の内部辺の長さが $T = 1$ の時（sticky でない場合）

それぞれの設定のもと，サンプルサイズは 100 から 1000 まで 100 刻みで，それぞれに対し 1000 サンプルをランダムに生成した．各推定量の RMSE の平均を計算した結果が図 4 である．

どちらの場合においても，大サンプルの下ではモード推定量のパフォーマンスがフレシェ平均のものを上回る．一方で，小サンプルの下では，このようなバイアスがあるにもかかわらずフレシェ平均のパフォーマンスがいくつかのモード推定量と同等の精度を持つことも見て取れる．また，どちらの設定においても，間接的な推定量 2 つと MVB モード推定量が良いパフォーマンスを示しているのに対し，MVCP モード推定量はこれらに比べ劣ったパフォーマンスを示している．計算の難しさや

ピーリングの存在がこの結果に影響していることも考えられる．

5　まとめ

本稿では，フレシェ平均やフレシェメジアンの stickiness に対処するため，系統樹の集合に対しての中心の推定問題としてモード推定の問題を考えた．モード推定量は直接的なものと密度推定を伴う間接的なものを与え，それぞれの一致性の条件を確かめた．また，直接的な推定量に関しては Finite Sample Breakdown Point の解析によってロバストネスの性質についても考察した．数値実験では，これらの推定量の一致性が確認され，フレシェ平均を用いると不適切な場合に，特に十分なサンプルサイズのもとでモード推定量が良いパフォーマンスを示すことが確認された．

一方で，モード推定量はその推定の効率性は一般的な推定量よりも低いものとなりやすい．実際に本研究の数値実験上でも，バイアスを持つフレシェ平均のパフォーマンスが低サンプルの下では他のモード推定量のパフォーマンスとあまり変わらないことが観測された．ノンパラメトリックな中心推定の問題において，モード推定を介さずに stickiness の問題に対処できるような手法の開発が可能かの検討が必要である．また，それぞれの推定量は高次元での計算は難しく，高次元でも効率的に動く計算アルゴリズムの開発も求められる．

参考文献

Barden, D., Le, H. and Owen, M. (2018), 'Limiting behaviour of fréchet means in the space of phylogenetic trees', *Ann. Inst. Stat. Math.* **70**(1), 99–129.

Billera, L. J., Holmes, S. P. and Vogtmann, K. (2001), 'Geometry of the space of phylogenetic trees', *Adv. Appl. Math.* **27**(4), 733–767.

Holder, M. T., Sukumaran, J. and Lewis, P. O. (2008), 'A justification for reporting the majority-rule consensus tree in bayesian phylogenetics'.

Kapli, P., Yang, Z. and Telford, M. J. (2020), 'Phylogenetic tree building in the genomic age', *Nat. Rev. Genet.* **21**(7), 428–444.

Kirschstein, T., Liebscher, S., Porzio, G. C. and Ragozini, G. (2016), 'Minimum volume peeling: A robust nonparametric estimator of the multivariate mode', *Comput. Stat. Data Anal.* **93**, 456–468.

Rannala, B. and Yang, Z. (2003), 'Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci', *Genetics* **164**(4), 1645–1656.

Shenmaier, V. V. (2013), 'The problem of a minimal ball enclosing k points', *J. Appl. Ind. Math.* **7**(3), 444–448.

Sturm, K.-T. (2003), Probability measures on metric spaces of nonpositive curvature, *in* 'Heat kernels and analysis on manifolds, graphs, and metric spaces: lecture notes from a quarter program on heat kernels, random walks, and analysis on manifolds and graphs', American Mathematical Society, pp. 357–390.

Takazawa, Y. and Sei, T. (2024), 'Maximum likelihood estimation of log-concave densities on

tree space', *Stat. Comput.* **34**(2), 84.

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 Trans. Comput. Biol. Bioinform.* **14**(6), 1359–1365.