統計数理(2025)
第 73 巻 第 2 号 213–238

特集「生存時間解析と信頼性解析」
[研究詳解]

# 天文学における生存時間解析：銀河の光度関数推定

竹内 努[1,2]

(受付 2024 年 11 月 27 日；改訂 2025 年 3 月 13 日；採択 3 月 14 日)

**要　　旨**

天文学の観測データは例外なく観測装置の検出限界に対応する切断を受けている．このようなデータから天体の統計量の分布関数を推定する際には，切断データの生存時間解析を用いるのが適切である．しかし，統計学で発展した生存時間解析は天文学分野には 1980 年代までまったく知られておらず，長らく天文学者が独立に考案した推定法が用いられてきた．天文学者による方法の多くは最終的には生存時間解析に帰着するものの数学的に未整理であり，統計的に体系だった議論がなされるようになったのは 21 世紀に入ってからである．本稿では，天文学における打ち切りデータ解析の応用として，銀河の光度関数推定について議論する．まず 1 変数の場合について天文データの特徴を詳しく紹介し，生存時間解析との対応を述べる．次に 2 変数の場合を導入する．2 変数行動関数の推定は天体の標本の取り方(サンプリング)によっては切断に加え打ち切りも入り，一般的な議論は複雑になる．ここでは 2 変数光度関数推定法を最も一般的な形で構築することを試みる．

キーワード：多波長光度，銀河進化，光度関数，生存時間解析，切断，打ち切り．

## 1. はじめに

### 1.1 銀河とその形成進化

銀河とは，星と星間物質(ガスと塵の混合流体)，暗黒物質からなる巨大な天体である．その質量は $10^4$–$10^{12}$ $\mathcal{M}_\odot$($\mathcal{M}_\odot$: 太陽質量)と広い範囲にわたる．そして観測可能な宇宙の範囲に $\sim 10^{11}$ 個もの銀河が存在している．宇宙は 138 億年前に誕生したが，初期の宇宙の物質はほぼ一様に分布しており，銀河のような天体は存在していなかった．つまり，銀河は形成し，現在の姿に時間的に進化してきたのである．すなわち時間進化が銀河の本質であり，これを定量化する銀河形成進化の研究は Tinsley (1980)以来半世紀近くにわたり銀河研究の中心であり続けている．

単体としての銀河の中では，暗黒物質が形作る重力場の中で星間物質から星が誕生し，そして死滅して持っていた物質の一部は星間空間に戻る．さらに，星は中心部での核融合によって重元素を合成し，その死とともに合成した元素も星間空間に還元する．銀河誕生から連綿と繰り返されるこの過程を物質循環といい，その結果星間物質の元素組成は刻々と変化してきた．

[1] 名古屋大学 素粒子宇宙物理学専攻：〒 464–8602 愛知県名古屋市千種区不老町; tsutomu.takeuchi.ttt@gmail.com

[2] 統計数理研究所 客員：〒 190–8562 東京都立川市緑町 10–3

### 1.2 天文学観測データの普遍的問題

#### 1.2.1 天文学における「明るさ」

天文学の観測は，その長い歴史の中で常に可能な限り暗い天体を検出する挑戦の連続である．大口径の光学望遠鏡や大規模電波干渉計が建造され，最新の観測データを提供してきた．どのような波長の天文観測データでも，我々観測者から遠ければ天体からの放射は検出限界を下回ってしまい，観測サンプルから落ちてしまう．このことを厳密に理解するため，いくつか重要な物理量を導入する．天体の本来の明るさである光度(luminosity: $L$)，すなわち天体が単位時間あたりに放射するエネルギーを考える．単位振動数当たりの光度を単色光度(monochromatic luminosity)と呼び，$L_\nu$ と表す．光度は天体が持つ本来の性質であり，観測者に依存しない量である．これに対し，天体から遠く離れた観測者が単位時間あたりに受ける検出器の単位面積当たりのエネルギーを放射流束(flux)と呼ぶ．ここでは放射流束を $S$ で表記する．光度と同じく，単位振動数当たりの放射流束を考えることができる．これを放射流束密度(flux density)と呼び，$S_\nu$ と表記する．放射流束密度は天体からの距離に依存する量で，遠ければ遠いほど値が小さくなる．この性質は遠くの光が暗く見えるという我々の直感と一致する．定量的に理解するため最も単純な定常ユークリッド空間の場合を考える[1)]．距離を $d$ と置くと

$$S = \frac{L}{4\pi d^2}$$

が成り立つ．すなわち放射流束は距離の 2 乗に反比例する．よって，天文学で「明るい」「暗い」というときには光度の意味なのか放射流束の意味なのかに常に注意を払う必要がある．光度が高いことは天体本来の(intrinsic)性質であるのに対し，放射流束が高いことは光度が高いか距離が近いのかの 2 通りの理由が考えられる．

#### 1.2.2 切断による天文学データ特有の偏り

このことを踏まえ，天文学データの持つ共通の面倒な性質について考える．あらゆる天文学探査データに共通するのが，データは観測装置の検出限界(detection limit)よりも明るい(放射流束が高い)天体しか含まれないことである．すなわち $S > S^{\rm lim}$ となる天体しかサンプルとして検出されない．振動数 $\nu$ での天体の等級(magnitude) $m_\nu$ は

$$m_\nu \equiv -2.5 \log_{10} S_\nu + \text{定数} \tag{1.1}$$

と定義される(詳しくは Supplement 第 B 章を参照)[2)]．等級を用いれば，観測データに含まれるのは $m_\nu < m_\nu^{\rm lim}$ の天体のみである．これが等級選択効果(magnitude selection effect)である．放射流束(あるいは放射流束密度)では，検出限界のあるデータは生存時間解析の左側切断データに対応する．同じデータを等級単位で扱った場合には係数 $-2.5$ のため右側切断データとなる．いずれにしても，電磁波で観測される天文データは必然的に切断データとならざるを得ない．この切断を図 1 に示す．

### 1.3 光度関数推定の歴史

銀河の光度関数とは，基本的には銀河の光度の確率密度関数で，慣習的に $\phi(L)$ と表記される．ただし，$n_{\rm gal}$ を銀河の宇宙空間における個数密度(単位 [個 $\mathrm{Mpc}^{-3}$])と定義すると，多くの天文学的応用において光度関数は

$$\int_L \phi(L)\,\mathrm{d}L = n_{\rm gal}\ ,$$

すなわち全光度範囲で積分すると空間個数密度となるよう規格化される．銀河の光度関数は，銀河系外天文学と観測宇宙論にとって本質的に重要な役割を果たす．まず，銀河の基本的な統

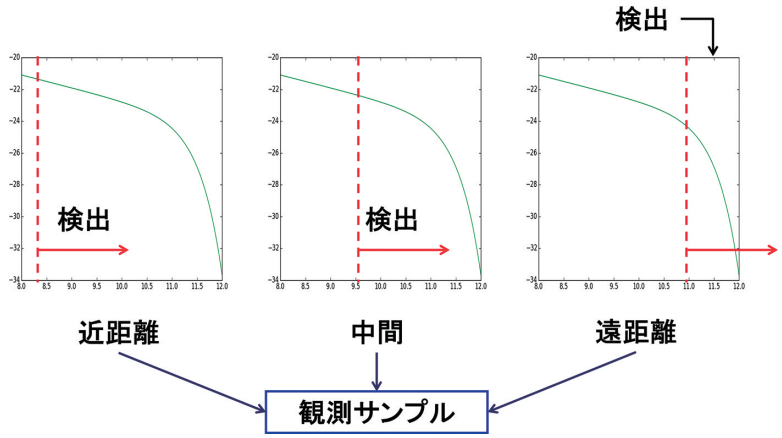


図 1. 銀河の 1 変数光度関数に対する放射流束選択効果．この図は（対数）光度を横軸に，（対数）空間密度を縦軸に取っている．

計的記述の 1 つであり，色，形態，あるいは周囲の銀河密度への依存性や，あるいは銀河の相関関数解析[3]などを通じて銀河の様々な性質を解析するのに用いられている．また光度関数は銀河形成理論の試金石としても重要である（たとえば Takeuchi et al., 2000, およびその参考文献を参照）．

第 1.1 節で述べたように，銀河は時間的に進化する（たとえば Tinsley, 1980）．銀河進化によって，銀河の光度関数も時間進化してゆく（たとえば Tinsley and Danly, 1980; Cowie et al., 1996; Sawicki et al., 1997; Pascarelle et al., 1998）．よって，銀河の光度関数推定は銀河進化研究の最も重要かつ基本的な統計的推定の一つである．しかし，観測された銀河カタログからの行動関数推定はまったく自明な作業ではない．上記のように，赤方偏移探査の検出限界によってデータは切断されており，これを扱う適切な統計手法が必要となる．銀河系外天文学の初期段階では，考えている宇宙空間体積 $V$ 内の銀河の数密度 $N/V$ という古典的推定量が光度関数推定に用いられた（たとえば Hubble, 1936）．もちろん，これは詳細な研究には不十分であり，多くの専門家が様々な方法を提案してきた．

Schmidt（1968）は，クエーサーの統計的研究の中で，有名な $1/V_{\rm max}$ 推定量を提唱した．そして Eales（1993）は宇宙年齢による光度関数の進化を調べるため，$1/V_{\rm max}$ 推定量を進化のあるサンプルに適用するための拡張を提案した．この推定量は，ある銀河 $i$ が検出可能な最大距離を半径とする「最大観測可能体積」$V_{\rm max}(i)$ を計算する．高光度の銀河ほど遠くまで観測できる，すなわち $V_{\rm max}$ が大きいことからその逆数を重みとして分布関数を求めるのが $1/V_{\rm max}$ 法である．$1/V_{\rm max}$ 推定量の詳細については Supplement 第 C 章で述べている．$1/V_{\rm max}$ 法では，銀河は最大観測可能体積 $V_{\rm max}$ のどこに存在していてもよい，すなわち銀河の分布が空間的に均一であると仮定していることが本質的である．しかし現在では銀河が大規模構造内で強くクラスタリングしていることが分かっており，これの仮定は $1/V_{\rm max}$ 法の短所と見なされている．この欠点にもかかわらず，$1/V_{\rm max}$ 推定量は現在でも銀河研究に頻繁に使用されている（たとえば Lilly et al., 1996; Ellis et al., 1996）．これは主に計算が簡単であるという理由による．

その後，主として可視光と周辺波長での光度関数を適切に表現する関数が Schechter（1976）によって提示され，これにより光度関数の関数形を仮定したパラメトリック推定の方法が提案されるようになった．Schechter 関数の仮定のもと，Sandage et al.（1979）は最尤法を導入

し，密度の非一様性による影響を取り除いた光度関数推定法を提案した．彼らの方法(STY 法)は爆発的に流行し，現在では光度関数のパラメトリック推定の標準的手法となっている（de Lapparent et al., 1989; Lin et al., 1996）．Lin et al.（1999)は光度関数のパラメータ進化を扱う拡張を行った．また Heyl et al.（1997)は光度関数と銀河密度のパラメトリックな同時推定法を開発した．これとは違った方法として Marshall et al.（1983)は，等級-赤方偏移空間上での天体のポアソン分布を仮定し，光度関数と進化パラメータの両方を同時に扱うことができる別のパラメトリック推定量を提示した．そして近代的な多波長銀河探査が主流となった現在では，光度関数推定も多変数への拡張が不可欠となっている．多波長での光度関数推定はしかし，複数の観測波長における選択効果が入りこむためより複雑な作業となる．天文学者はこれまで多波長あるいは一般の物理量を含む多変数への拡張に挑んできたが，天文分野においては統計学の方法論の知識は極めて限られていたため，往々にして提案手法は *ad hoc* であり拡張性に欠けていた（たとえば Chołoniewski, 1985）．系統的な光度関数の多波長化と選択効果の適切な議論は 2000 年代に入り，コピュラ多変数光度関数の導入によって可能となった（Takeuchi, 2010; Takeuchi et al., 2013; Takeuchi and Kono, 2020）．

銀河の光度関数の詳細な解析が進むにつれ，光度関数が銀河環境や色，形態など実に様々な物理量に依存することが分かってきた（たとえば Binggeli et al., 1988）．このような詳細な特徴を記述するには，解析的なフィッティング式では不十分である．このような背景から，本稿は光度関数のノンパラメトリック推定に焦点を絞って議論する．パラメトリックな推定法については稿を改めて紹介したい．天文学的サンプルの偏りの定量評価と銀河の光度関数推定法について，2011 年までの歴史と応用例の最も完備な解説は Johnston（2011)にある．本章の内容がより具体的に示されているので，天文学的内容に興味を持たれた読者は Johnston（2011)を参照されたい．

なお，これらの方法で推定する分布関数としては，銀河の持つ様々な種類の天体の総質量 $\mathcal{M}$ の分布，いわゆる質量関数も重要な研究対象である．質量は原理的にはある波長での銀河の光度から換算して得られる量であり，統計的解析方法は光度関数の場合と本質的に同一である．このため，以降の議論では光度関数と質量関数は区別せず，光度関数で統一する．

本論文の構成は以下のようになっている．第 2 章では，Lynden-Bell（1971)の $C^-$ 法と呼ばれる推定法を導出する．続いて，Woodroofe（1985)による生存時間解析の切断(truncated)データ解析の観点からの再定式化を示す．この章の後半では Woodroofe（1985)の議論と，2 変数への拡張のための一般化を展開する．第 3 章では 2 変数光度関数推定問題を議論する．天体探査データの場合，2 変数への拡張，すなわちある探査データに別の波長での観測を組み合わせる際，銀河の放射流束測定に切断に加えて打ち切り(censoring)が生じる．まず伝統的に天文学で用いられてきた 2 変数光度関数推定の手続きについて述べる．次に van der Laan（1996)によって導入された，切断データの 2 変数分布関数の最尤推定量を導く．これを踏まえ，最終的に目指す天文学観測データの完全な 2 変数光度関数推定法を示す．第 4 章では，本論文の内容のまとめを示し，天文学における光度関数推定で今後検討すべき潜在的課題について言及する．Dabrowska（1988)の与えた 2 変数打ち切りデータにおける分布関数推定法について，本論文で用いる式とその導出を付録 A に示す．

本論文は天文学・宇宙物理学および生存時間解析の横断的研究であるため，必要な基礎知識を Supplement にまとめている．Supplement A 章では宇宙論の基礎的な量を紹介する．天文学で長く用いられている等級という量については混乱しがちな点が多いため，第 B 章で解説する．本文では簡単に紹介している Schmidt の $1/V_{\max}$ 推定量については，第 C 章で正確な定義を示す．第 D 章では，打ち切りデータの分布の最尤推定量であるカプラン-マイヤー(Kapkan–Meier)推定量の導出を示す．一般に生存関数のノンパラメトリック最尤推定量であ

る積極限推定量の分散は影響関数(influence function)で表せる．影響関数の基礎的説明は第 E 章で述べる．

明示的には扱わないが，本論文では宇宙年齢の算出などにおいて宇宙論パラメータ $h = H_0/(100\ [\mathrm{km\ s^{-1}\ Mpc^{-1}}]) = 0.7$, $\Omega_{\Lambda 0} = 0.7$, $\Omega_{\mathrm{M}0} = 0.3$, 曲率パラメータ $\Omega_{K0} = 0$ を採用している(Supplement D 章参照)[4]．

## 2. 1 変数光度関数推定

### 2.1 銀河のクラスタリングと $1/V_{\max}$ 法の限界

既に議論したように，$1/V_{\max}$ の適用限界を決めている短所は，天体の分布が空間的に一様であるという仮定である．銀河を含むぼんやりとした天体である “星雲(nebula)” の天球上での分布が一様ではなくゆらぎを示すという事実は，それらが銀河系外天体であると証明されるずっと前，19 世紀からウィリアム・ハーシェル(William Herschel)やシャルル・メシエ(Charles Messier)の研究によって知られていた．Shane and Wirtanen (1967)はパロマー天文台のシュミット望遠鏡で撮像された写真乾板を用いて 19 等より明るい銀河の天球上での地図を作製し，銀河の広範囲にわたるクラスタリングを初めて明らかにした．このような研究を通じて，銀河の空間分布における銀河群，銀河団，超銀河団[5]といった銀河分布の階層構造が明らかになった．そして de Lapparent et al. (1986)以降，宇宙論的赤方偏移を距離として用い，銀河の 3 次元空間分布を直接探査する赤方偏移探査(redshift survey)プロジェクトが大規模におこなわれ，銀河の 3 次元クラスタリングの性質は精密に測定されている．銀河分布の非一様性を無視して $1/V_{\max}$ 法を用いると，光度関数推定に顕著なバイアスが生じることがよく知られている．たとえば Takeuchi et al. (2000)はこのバイアスが極めて深刻になり得ることを数値実験によって明確に示した．クラスタリングは $1/V_{\max}$ の枠組みでは極めて厄介な問題だが，意外にも宇宙物理学においてこの問題を回避できる代替手法が開発されるまでにさほど時間はかからなかった．これが Lynden-Bell (1971)によって導入された “$C^-$ 法” である．

### 2.2 Lynden-Bell の $C^-$ 法

Lynden-Bell (1971)は統計学分野での生存時間解析の存在を知らないまま，天文学分野で独立にその一種である切断データ解析の方法を構築した．彼はこの方法を Schmidt (1967)のクェーサー探査データに適用してクラスタリングの影響を受けない光度関数を求めることに成功した．近代的な視点から見ると，Lynden-Bell の方法は切断(truncated)データに対する生存時間解析の応用例となっている．統計分野では，Woodroofe (1985)による Lynden-Bell の方法の紹介以降研究が活発になり，議論が大きく発展した．本章ではオリジナルの議論と生存時間解析からの再構成を詳しく紹介する．なお，天文学的表記に不慣れな読者は本節(第 2.2 節)の式(2.1)–(2.6)の議論は飛ばして次節(第 2.3 節)に進まれてもよい．

まずいくつかの基本的な量を定義する．銀河の絶対等級を $M$, 見かけの等級を $m$, 赤方偏移 $z$ に対応する光度距離を $d_{\mathrm{L}}(z)$ (単位 [Mpc])とする．

$$M = m - 5\log d_{\mathrm{L}}(z) - 25$$

である．また，銀河の光度距離は非常に大きな範囲を取るため，その対数を用いた距離の指標である距離指数(distance modulus)

$$D(z) \equiv m - M = 5\log d_{\mathrm{L}}(z) + 25$$

を導入する．赤方偏移 $z$ と距離指標 $D(z)$ は等価な情報であり，往々にして引数 $z$ は省略して

単に $D$ と表記する．距離指数の単位は等級 [mag] である（詳細は Supplement 第 B 章を参照）．絶対等級で表現した光度関数を $\phi(M)$ [$\mathrm{Mpc}^{-3}\,\mathrm{mag}^{-1}$]) と書く．また $N_{\mathrm{obs}}$ を考えている銀河探査で検出された銀河の総数とする．$C^-$ 法は積分分布関数を Dirac の $\delta$ 関数 $\delta_{\mathrm{D}}$ の重み付き和として表現する．すなわち，等級の測定誤差は仮定しない．見かけの等級の誤差を考慮するためには，たとえば等級に対してスムージングカーネルを導入することもできる（例えば Caditz and Petrosian, 1993; SubbaRao et al., 1996）．

仮想的に観測装置の検出限界に制約されない状況を考えると，Hubble (1936) で用いられた初期の議論そのものに帰着し，積分光度関数 $\Phi(M)$ の構築は容易である．しかし，現実では検出限界による見かけの等級 $m$ への制約により，現実には偏りのある部分サンプルのみが観測される．具体的には，みかけの等級が限界等級 $m^{\mathrm{lim}}$ よりも明るい銀河のみがサンプルに含まれる．これを $X(M)$ と表記すると，

$$\frac{\mathrm{d}\Phi(M)}{\Phi(M)} > \frac{\mathrm{d}X(M)}{X(M)} \tag{2.1}$$

となる．後述のように，限界等級 $m^{\mathrm{lim}}$ への依存性は絶対等級 $M$ を通じて $X(M)$ に入っている．

次に，観測された銀河の分布を絶対等級 $M$ と距離指標 $D$ の平面 $M$–$D$ で考え，$M$ と $D$ は変数分離可能と仮定する（図 2）．図 2 において，探査の検出限界 $m^{\mathrm{lim}}$ は斜めの線で示されている．観測された母集団での $M$ と $D$ の確率密度関数を次のように書くことができる．

$$\mathrm{d}P = \rho(D)\mathrm{d}D\,\phi(M)\mathrm{d}M\,\Theta(m^{\mathrm{lim}} - m)\,,$$

すなわち

$$\frac{\mathrm{d}^2P}{\mathrm{d}D\mathrm{d}M} = \rho(D)\phi(M)\Theta(m^{\mathrm{lim}} - m)\,.$$

ここで $\Theta(m^{\mathrm{lim}} - m)$ は等級による選択効果（打ち切り）を記述する Heaviside 関数で，

$$\Theta(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

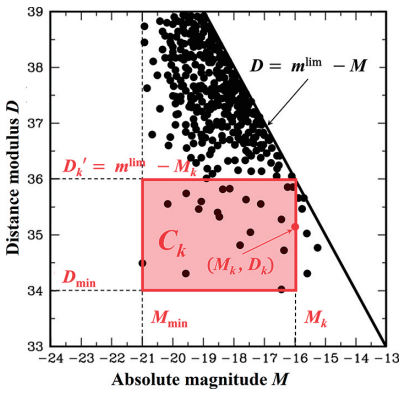


図 2. Lynden-Bell (1971) によって導入された $C^-$ 法の概念図．黒丸シンボルがひとつひとつの銀河を表す．サンプル内の各銀河 $(M_k, D_k)$ に対して領域 $C_k$ を定義する．この領域内において，$(M_k, D_k)$ 以外の全ての銀河をカウントし，計数を $C^-(M_k)$ と表す．$M_{\mathrm{min}}$ はサーベイデータ中で最も明るい銀河によって決定される．同様に，$D_{\mathrm{min}}$ は最も近い赤方偏移 $z_{\mathrm{min}}$ の銀河によって定義される最小距離指数である．

と定義される．

観測された $X(M)$ の部分領域 $C^-(M)$ を

$$\frac{d\Phi}{\Phi} = \frac{dX}{C^-}$$

となるように取れば，累積光度関数(累積分布関数)はこれを積分することにより

$$\Phi(M) = A \exp\left[\int_{-\infty}^{M} \frac{\mathrm{d}X(M')}{C^-(M')}\right]$$

という形で求められる．ここで $A$ は規格化定数，$X(M)$ は $C^-(M)$ に存在する部分サンプルである．

しかし，実際の解析では微分光度関数(確率密度関数に対応)および空間数密度が必要であり，これはそれぞれ Dirac の $\delta$ 関数 $\delta_{\mathrm{D}}$ の系列として表すことができる．実際に観測されたサンプルを $\{(M_i, D_i)\}$ $(i = 1, \ldots, N)$ とする．データは明るい順(絶対等級 $M_i$ の数値が小さい順)にソートされているとする．光度関数と空間密度はそれぞれ次のように表される．

$$\phi(M) = \sum_{i=1}^{N} \psi_i \delta_{\mathrm{D}}(M - M_i), \tag{2.2}$$

$$\rho(D) = \sum_{i=1}^{N} \rho_i \delta_{\mathrm{D}}(D - D_i), \tag{2.3}$$

ここで，$\psi_i$ と $\rho_i$ はそれぞれのステップサイズ，すなわちサンプル $i$ と $i+1$ の積分光度関数および空間数密度の差である．次に，積分光度関数を構築するために以下の量を定義する．得られた観測サンプルに含まれるある銀河 $k$ のデータを $(M_k, D_k)$ とし，この銀河が仮想的に検出される最大の距離指標

$$D'_k = m^{\lim} - M_k$$

を考える．関数 $C^-(M_k)$ を，図 2 の影つきの領域で示された領域 $C_k$

$$\begin{cases} M_{\min} \leq M_i < M_k, \\ D_{\min} \leq D_i \leq D'_k \end{cases}$$

内に存在する銀河の数として定義する．

$$C^-(M_k) \equiv \#\{(M_i, D_i) \in C_k\} \quad (k = 1, \ldots, N)\,.$$

Jackson (1974)によると，上付きの $-$ は $C^-(M_k)$ を評価するときに $(M_k, D_k)$ の点が含まれないことを強調するためのものである．光度関数の係数は次の漸化式から決定される．

$$\psi_{k+1} = \psi_k \frac{C^-(M_k) + 1}{C^-(M_{k+1})}. \tag{2.4}$$

以上より，積分光度関数は

$$\Phi(M_k) = \int_{M_{\min}}^{M_k} \phi(M) dM = \psi_1 \prod_{k:M_k<M} \frac{C^-(M_k) + 1}{C^-(M_k)} \tag{2.5}$$

によって推定できる．これが Lynden-Bell の公式と呼ばれる推定量である．ここで，式(2.5)に

おいて，

$$\frac{C^-(M_1)+1}{C^-(M_1)}=1 \tag{2.6}$$

とし，積が $k=2$ から始まるように設定することを注意しておく（Chołoniewski, 1987; Takeuchi et al., 2000）．

Lynden-Bell の $C^-$ 法は，データセット内の天体の空間分布に関する仮定を必要としない点において従来の $1/V_{\max}$ 法よりも本質的に優れている．天文学者の考案したその他のノンパラメトリックな光度関数推定法は，本質的には全て各ビン当たりの銀河数を $\{0,1\}$ にした極限を取れば $C^-$ 法に帰着する（たとえば Takeuchi et al., 2000; Johnston, 2011, および参考文献）．

Chołoniewski（1987）はオリジナルの $C^-$ 法を再検討し，難解な原論文を簡略化しただけでなく，光度関数を適切に規格化して銀河の密度進化を同時に推定できるよう改良した．Chołoniewski の方法の証明は Takeuchi et al.（2000）によって与えられた．しかしこの改良版 $C^-$ 法はこれまでのところ，応用上はごく希にしか用いられていない（たとえば Takeuchi et al., 2000, 2006）．

### 2.3 Woodroofe（1985）による $C^-$ 法の再定式化

Woodroofe（1985）は Lynden-Bell の $C^-$ 法を生存時間解析の立場から再構成し，推定量の漸近的性質を導いた．これにより，Lynden-Bell（1971）の業績は Woodroofe によって統計学分野に知られることとなった．本節では Woodroofe の定式化について述べる．

仮想的な母集団サンプル $(X_1,Y_1),\ldots,(X_N,Y_N)$ を考える．$X$ と $Y$ の分布関数をそれぞれ $X\sim F, Y\sim G$ とし，$X$ と $Y$ は独立であるという基本的仮定を置く．サンプル $i(\leq N)$ が観測されるのは次の条件

$$Y_i\leq X_i \tag{2.7}$$

が満たされる場合に限られるとする．観測されたサンプルのデータの集合を $(x_1,y_1),\ldots,(x_n,y_n)$ と表し，$\{x_i\}$ $(i=1,\ldots,n)$ および $\{y_j\}$ $(j=1,\ldots,n)$ の経験分布関数を，それぞれ $F_n^*$ と $G_n^*$ とする．すなわち，$0\leq t<\infty$ について次のように定義する．

$$F_n^*(t)=\mathbb{P}(X\leq t|Y\leq X)=\frac{1}{n}\#\{i\leq n\,:\,x_i\leq t\}\,, \tag{2.8}$$

$$G_n^*(t)=\mathbb{P}(Y\leq t)=\frac{1}{n}\#\{j\leq n\,:\,y_j\leq t\}\,. \tag{2.9}$$

これらの量の直感的意味を図 3 に示す．$F_n^*$ および $G_n^*$ は条件(2.7)の下での分布関数 $F^*$ および $G^*$ の標本推定量となる(すなわち $n\longrightarrow\infty$ において $F_n^*\longrightarrow F^*$, $G_n^*\longrightarrow G^*$)．

ここで次の量

$$C_n(t)\equiv G_n^*(t)-F_n^*(t-)\qquad(0\leq t<\infty)$$

を定義する．すると $^\forall i\leq n$ について

$$C_n(x_i)\geq\frac{1}{n}$$

となる．分布 $F$ の累積ハザード関数は次で推定される．

$$\hat{\Lambda}_n(t)\equiv\int_0^t\frac{\mathrm{d}F_n^*(x)}{C_n(x)}=\sum_{i:\,x_i\leq t}\frac{1}{nC_n(x_i)}\qquad(0\leq t<\infty)\,. \tag{2.10}$$

式(2.10)を用いると，分布関数 $F$ の標本推定量 $\hat{F}_n(t)$ は次のようになる．

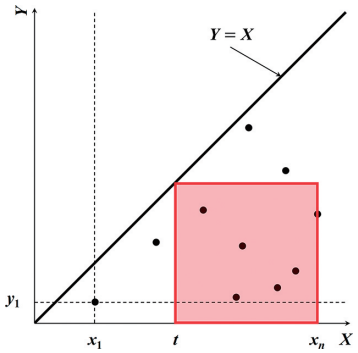


図 3. Woodroofe (1985)の方法におけるデータ構造の概略図.

$$\hat{F}_n(t) = 1 - \prod_{i:x_i \le t} \left[1 - \frac{r(x_i)}{nC_n(x_i)}\right] \qquad (0 \le t < \infty)\,, \tag{2.11}$$

ここで $r(x_i) \equiv \#\{k \le n : x_k = x_i\}$ である．また

$$\hat{G}_n(t) = \prod_{j:y_j > t} \left[1 - \frac{s(y_j)}{nC_n(y_j)}\right] \qquad (0 \le t < \infty) \tag{2.12}$$

が成り立つ．ただし $s(y_j) \equiv \#\{k \le n : y_k = y_j\}$.

さらに

$$\alpha \equiv \mathbb{P}\{Y \le X\} \tag{2.13}$$

の最尤推定量は

$$\hat{\alpha}_n = \int_0^\infty \hat{G}_n \mathrm{d}\hat{F}_n \tag{2.14}$$

で与えられる．この $\hat{\alpha}_n$ を用いて，母集団サイズ $N$ の推定量は

$$\hat{N}_n \equiv \frac{n}{\hat{\alpha}_n} \tag{2.15}$$

のように求められる.

データに同値(タイ)がない場合，次が成り立つ.

$$\hat{G}_n(t) = \prod_{j:y_j > t} \left[\frac{nC_n(y_j) - 1}{nC_n(y_j)}\right] \equiv \prod_{j:y_j > t} \left(\frac{n_j - 1}{n_j}\right)\,. \tag{2.16}$$

天文学における実践では同値が生じることは基本的にないため，式(2.16)の形が最尤推定量として用いられている．これは Woodroofe の公式，あるいは Lynden-Bell–Woodroofe の公式と呼ばれる．$\hat{G}_n$ の漸近分散推定量は

$$\hat{\mathbb{V}}[\hat{G}_n(t)] = \hat{G}_n(t)^2 \sum_{j:y_j > t} \frac{1}{n_j(n_j - 1)} \tag{2.17}$$

で与えられる(Wang et al., 1986)．式(2.14)，(2.15)，(2.16)は最尤推定量であることが示されている（Woodroofe, 1985)．また収束性，一致性，および漸近正規性の証明は Woodroofe (1985)，Wang et al. (1986)，Keiding and Gill (1990)，および van der Vaart (1991)によって

与えられている．

ここで Lynden-Bell (1971) と Woodroofe (1985) の光度関数推定量の関係について考察する．Lynden-Bell の公式(2.5)の積の部分は次のように表される．

$$\prod_{k:M_k\le M}\frac{C^-(M_k)+1}{C^-(M_k)}=\frac{1}{\prod_{k:M_k>M}\frac{C^-(M_k)+1}{C^-(M_k)}}=\prod_{k:M_k>M}\frac{C_k^-(M)}{C^-(M_k)+1}\,. \tag{2.18}$$

ここで $n_k=C^-(M_k)+1$ と置くと，Woodroofe の公式(2.16)が得られる．

この関係は視覚的に理解できる．極限等級 $m^{\lim}$ より明るい銀河が実際の観測サンプルに含まれることを想起する．検出の境界に対応する絶対等級 $M^{\lim}$ は次のように表される．

$$M^{\lim}(D)=-D+m^{\lim}\,. \tag{2.19}$$

等級の定義 [式(1.1)] により，光度が高い銀河ほど絶対等級 $M$ は小さくなる．したがって，図 2 の横軸は銀河の光度の順序としては反転していることになる．

ここで新しい変数 $W$ を

$$W\equiv-(M-m^{\lim})$$

と定義すると，限界等級を表す線[式(2.19)]は

$$D=W$$

と表される．この変換によって $(W,D)=(X,Y)$ とみなせ，Woodroofe (1985) の定式化が得られることが分かる．

### 2.4 2 変数推定量への拡張のための一般的定式化

ここでは，推定量の 2 変数への拡張を目指して，上記の議論をより一般的な形で記述する．確率変数 $X$ と $Y$ を考え，分布関数を $X\sim F$ および $Y\sim G$ とする．前節と同様，$X$ と $Y$ は独立と仮定する．それぞれの経験分布関数を前節同様 $F_n^*$, $G_n^*$ とし，その極限形式を $F^*$, $G^*$ とおく．なお，前節の Woodroofe (1985) の議論では絶対等級の符号を変えることで左側切断データとして扱ったが，天文学データでは混乱のもととなるためこれ以降は符号は変えず，右側切断データとして取り扱う．即ち $X\le Y$ のときにサンプルとして観測されるとする．$X_i$ および $Y_i$ $(i=1,\ldots,n)$ を観測したとき，生存関数 $S(x)$,

$$S(x)\equiv1-F(x) \tag{2.20}$$

のノンパラメトリック最尤推定量(NPMLE)は積極限推定量(product limit estimator)として次の式(2.21)で定式化される．

$$S_{\rm PL}(x)\equiv\prod_{(0,x]}[1-\Lambda_n(\mathrm{d}s)]. \tag{2.21}$$

ここで，

$$\Lambda_n(\mathrm{d}s)=\frac{\sum_{i=1}^n\mathbb{1}(X_i\in\mathrm{d}s)}{\sum_{i=1}^n\mathbb{1}(X_i\ge s,Y_i\ge s)} \tag{2.22}$$

はハザード確率

$$\Lambda(\mathrm{d}s)\equiv\mathbb{P}(X\in\mathrm{d}s|X\ge s,X\le Y) \tag{2.23}$$

の推定量である．積極限推定量は，漸近線型かつ有効推定量であることが示されている．漸近

線型性とは

$$\sqrt{n}[S_{\mathrm{PL}}(x) - S(x)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{IF}_{\mathrm{PL}}(X_i, Y_i | F, G, x) + o_p(1) \tag{2.24}$$

を意味する．ここで，$\mathbb{IF}_{\mathrm{pl}}(X_i, Y_i|F, G, x)$ は推定量の影響関数である．積極限推定量の影響関数は次で与えられる．

$$\begin{aligned} &\mathbb{IF}_{\mathrm{PL}}(X_i, Y_i|F, G, x) \\ &= -S(x) \left[ \frac{\mathbb{1}(X \le s | X \le Y)}{(G^* - F^*)(X)} - \alpha \int_0^{x \wedge X} \frac{F(\mathrm{d}s)}{S^2(s)G(s)} \alpha \int_0^{c \wedge Y} \frac{F(\mathrm{d}s)}{S^2(s)G(s)} \right]. \end{aligned} \tag{2.25}$$

(Bickel et al., 1993)．ここで $\alpha$ は前節で定義した式(2.13)であり，

$$(G^* - F^*)(s) = \mathbb{P}(X \ge s, Y \le s | X \le Y) \tag{2.26}$$

である．また $a \wedge b \equiv \min\{a, b\}$ と表記している．したがって，$\hat{S}_{\mathrm{PL}}$ の分散は式(2.24)を用いて評価できる．一般に，NPMLE の分散は影響関数を用いることで評価可能である．影響関数の基本的事項は Supplement 第 E 章で述べている．

さらに，右側打ち切りデータが存在する場合について議論する．具体的には，$Y^*$ を右側打ち切り変数，$X$ を関心のある変数とし，$\Delta \equiv \mathbb{1}(X \le Y^*)$ とすると，条件 $X \le Y$ のもとで $(\tilde{X} = X \wedge Y^*, Y, \Delta)$ を観測する．この場合の $S$ の積極限推定量は次で与えられるハザード関数推定量を式(2.21)に代入することで得られる．

$$\Lambda_n(\mathrm{d}s) = \frac{\sum_{i=1}^{n} \mathbb{1}(\tilde{X}_i \in \mathrm{d}s, \Delta_i = 1)}{\sum_{i=1}^{n} \mathbb{1}(\tilde{X}_i \ge s, Y_i \ge s)}, \tag{2.27}$$

ここで $\Delta_i = \mathbb{1}(X_i \le Y_i^*)$ である．上記の影響関数は，この場合にも拡張可能である．

## 3. 2 変数光度関数推定

### 3.1 2 変数データにおける選択効果

現在，大規模天文サーベイはすべて多波長(マルチバンド：multiband)で行われている．ここでは 2 変量の場合(すなわち，2 つのバンドで選択されたサンプル)を考え，その 2 変数光度関数を $\phi^{(2)}(M_1, M_2)$ で表す．より多くのバンドの場合(あるいは，より一般的に任意の物理量で選択される場合)への拡張は直接的である．放射流束によるセレクションは，光度-光度($L_1$–$L_2$)平面上に下限 $L^{\lim}$ を設定することとして記述される(図 4 を参照)．同様に，絶対等級を用いれば $M_1$–$M_2$ 平面上に上限 $M^{\lim}$ を設定することとして表される．上限絶対等級 $M^{\lim}$ は，距離指数の関数としての限界等級 $m^{\lim}$ によって定義される[式(2.19)]．一般的なサーベイでは，ある特定の波長を主選択バンドと設定する．例えば $B$-バンド(可視光)，$K$s-バンド(近赤外)，60 $\mu$m (遠赤外)などである．

バンド 1 で天体(この場合は銀河)のサンプルを選択すると，距離指数 $D$ において $M_1 \ge {M^{\lim}}_1(D)$ である天体はサンプルに含まれない．したがって，検出される天体は $M_1 < {M^{\lim}}_1(D)$ かつ $M_2 < {M^{\lim}}_2(D)$ である必要がある．したがって，$M_1$–$M_2$ 平面上での検出された天体の 2 次元分布

$$\begin{aligned} &\Sigma^{\mathrm{det}}(M_1, M_2, D) \\ &\equiv \int_0^D \frac{\mathrm{d}^2 V}{\mathrm{d}D' \mathrm{d}\Omega} \phi^{(2)}(M_1, M_2) \{1 - \Theta[{M^{\lim}}_1(D')]\} \{1 - \Theta[{M^{\lim}}_2(D')]\} \mathrm{d}D' \end{aligned} \tag{3.1}$$

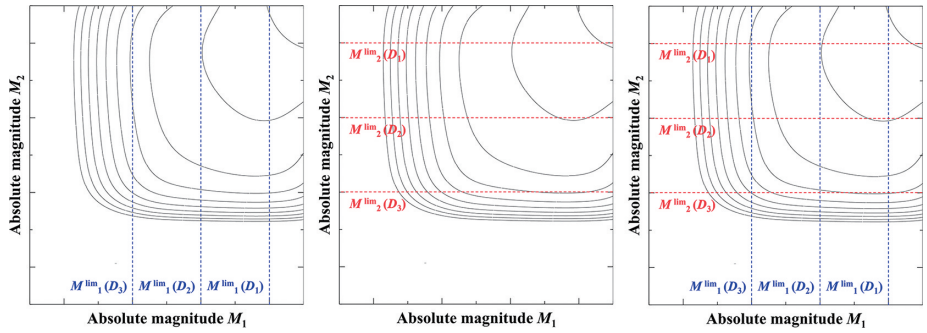


図 4. 2 バンドサーベイにおける選択効果の概略図．等高線は銀河の 2 変数光度関数のモデルを表す．ここでは銀河の光度は絶対等級で表している．$M_1$, $M_2$ がそれぞれバンド 1, 2 での絶対等級，$M^{\lim}{}_1(D)$, $M^{\lim}{}_2(D)$ が距離指数$(D)$でのそれぞれのバンドにおける検出限界光度を表す．左：バンド 1 に基づくサンプルセレクションの場合，中央：バンド 2 に基づくサンプルセレクションの場合，右：バンド 1 および 2 の 2 つのセレクションがかかる場合をそれぞれ示している．バンド 1 では垂直線よりも左側，バンド 2 では水平線よりも下側にある銀河が検出される．

のように表される．ここで $\Omega$ は立体角である．$\Sigma^{\rm det}$ は $M_1$–$M_2$ 平面上で両バンドで検出された天体の面密度に比例する量である．

バンド 1 を主選択バンドとすると，バンド 1 で検出されているがバンド 2 では検出されない天体が存在することになる．

この場合，これらの天体については見かけの等級の下限値(放射流束の上限値)のみが得られる．バンド 2 における等級の下限値の 2 次元分布も同様に次のように定式化される．

$$\Sigma^{\rm LL2}(M_1, M_2, D) \equiv \int_0^D \frac{{\rm d}^2V}{{\rm d}D'{\rm d}\Omega}\phi^{(2)}(M_1, M_2)\{1 - \Theta[M^{\lim}{}_1(D')]\}\Theta[M^{\lim}{}_2(D')]{\rm d}D'. \tag{3.2}$$

上付き添字 LL2 は ‘バンド 2 における下限値 (lower limit at band 2)’ を表す．このように，天体の存在は確定しているが特定の量の上限(または下限)のみが得られる場合は統計学では ‘打ち切り’(censored)と呼ばれる．式(3.2)で分布 $\Sigma^{\rm LL2}(M_1, M_2)$ を定義できるが，このカテゴリに属するサンプル天体はプロット上で下限値としてのみ現れる．一方，バンド 1 で天体を選択しているため，バンド 1 には下限値は存在しない．下限値以上の等級に天体が存在するかどうかはわからないからである．この場合は統計学の ‘切断(truncated)’ に対応する．

バンド 2 で天体を選択する場合も，バンド 1 で選択されたサンプルと全く同様に，検出された天体と下限値の 2 次元分布を定式化できる．この場合も両バンドで検出された天体の 2 次元分布は式(3.1)で表される．バンド 2 で検出され，バンド 1 では検出されない天体は

$$\Sigma^{\rm LL1}(M_1, M_2, D) \equiv \int_0^D \frac{{\rm d}^2V}{{\rm d}D'{\rm d}\Omega}\phi^{(2)}(M_1, M_2)\Theta[M^{\lim}{}_1(D')]\{1 - \Theta[M^{\lim}{}_2(D')]\}{\rm d}D' \tag{3.3}$$

のように表される．天文学の多波長サーベイでは，様々な $D$ におけるサンプルが同時に含まれたデータが得られる．このデータから 2 変数光度関数を推定する問題は複雑だが，2 変数生存時間分析の手法を用いることで統一的に扱うことができる．

### 3.2 生存時間解析の方法による 2 変数光度関数の推定

#### 3.2.1 天文学で用いられてきた手法

2 変量光度関数の推定は，長らく天文学者の関心を集めてきた重要な問題である．にも拘ら

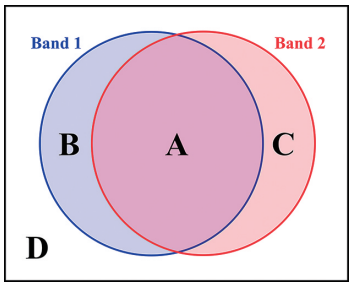


図 5. 2 変数光度選択でのデータのカテゴリを示すヴェン図．それぞれ，領域 A: バンド 1 と 2 の両方で検出された天体，領域 B: バンド 1 で検出されるがバンド 2 では検出されていない天体，領域 C: バンド 2 で検出されるがバンド 1 では検出されていない天体，領域 D: バンド 1 と 2 のどちらでも検出されていない天体に対応する．バンド 1 で選択されたサンプルは A と B の和集合に対応し，バンド 2 で選択されたサンプルは A と C の和集合に対応する．

ず，実際にはこの問題におけるデータと推定量の構造は天文学および天体物理学では適切に理解されてこなかった．天文学における膨大な数の論文が ‘2 変数分布関数’ に言及しているが，データが 1 つの波長バンドで選択されたと暗黙のうちに仮定している．つまり，天文学で ‘2 変数分布関数推定’ として知られるほぼすべての手法は，主選択バンドで測定された観測量の周辺分布が完全に既知であるか推定可能であり，上限値または下限値を含む第 2 の観測量がある場合の問題について扱っている（e.g., Mobasher et al., 1993; Keres et al., 2003; Wall and Jenkins, 2003）．

しかし，図 4 からも理解できるように，これは 2 変数分布推定のうちの 1 つの特殊な場合に過ぎない．図 5 は，2 つの波長バンド（バンド 1 とバンド 2）で選択されたサンプルのカテゴリを示している．

バンド 1 で選択されたサンプルのバンド 2 光度関数は通常次のように推定される．

(1) 何らかの推定法，たとえば $C^-$ 法を用いてバンド 1 の 1 変数光度関数を推定する．
(2) サンプルのバンド 1 光度をビンに分割する．
(3) 各バンド 1 光度ビンについて，打ち切りデータ解析（例：Kaplan–Meier 法: Kaplan and Meier 1958）を用いてバンド 2 の 1 変数光度関数を推定する．
(4) バンド 1 光度関数と条件付きバンド 2 光度関数を組み合わせ，最終的な 2 変数光度関数を構築する．

この方法の応用例としては，例えば Buat et al. (2009) がある．Kaplan–Meier 推定量については Supplement 第 D 章で説明されている．ここで注意すべき点は，この方法で得られる 2 変数光度関数は，天文学者が本来求めているものとは一致しないということである．この混乱は，いくつかの天文学文献において銀河の母集団の物理的特性について偏った結論をもたらした．以下ではこの問題を明確に定義し，より適切に扱うための方法について議論する．

### 3.2.2 切断（truncated）データの 2 変数生存時間分析

まず van der Laan (1996) の提案した，2 バンド切断データの分布関数推定法を構築する．van der Laan (1996) は最尤法のスコア方程式の解として推定量を導出したが，ここでは Huang et al. (2001) に従い，構成的に推定量を求める．このために第 2.4 節の議論を直接拡張すると，2 変数切断データは以下のようなモデルで表現できる．解析対象の確率変数を $(X_1, X_2)$, 切断を決め

る確率変数を $(Y_1, Y_2)$ とする．それぞれの分布関数を $(X_1, X_2) \sim F(x_1, x_2)$, $(Y_1, Y_2) \sim G(y_1, y_2)$ と置く．ここでも $(X_1, X_2)$ と $(Y_1, Y_2)$ は独立と仮定する．

第 2.4 節の議論と同様，データには $X_1 \leq Y_1$ かつ $X_2 \leq Y_2$ の場合のみが含まれる．観測された $(X_1, X_2)$ および $(Y_1, Y_2)$ の同時分布は次で与えられる．

$$\mathbb{P}(X_1 \leq x_1, Y_2 \leq x_2, Y_1 \leq y_1, Y_2 \leq y_2 | X_1 \leq Y_1, X_2 \leq Y_2)\,. \tag{3.4}$$

$x_1 \geq 0$ および $x_2 \geq 0$ が固定されているとする．観測された $(X_1, X_2)$ の 2 変数分布を次のように考える．

$$F^*(x_1, x_2) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2 | X_1 \leq Y_1, X_2 \leq Y_2)\,, \tag{3.5}$$

観測された(換算)絶対等級 $(y_1, y_2)$ の分布関数を $G^*(x_1, x_2)$ とすると，次が成り立つ．

$$\begin{aligned} G^*(x_1, x_2) &= \mathbb{P}(Y_1 \leq x_1, Y_2 \leq x_2 | X_1 \leq Y_1, X_2 \leq Y_2) \\ &= \frac{1}{\alpha^{(2)}} \mathbb{P}[(Y_1 \leq x_1, Y_2 \leq x_2), (X_1 \leq Y_1, X_2 \leq Y_2)] \\ &= \frac{1}{\alpha^{(2)}} \int_{x_1}^{\infty} \int_{x_2}^{\infty} F(y_1, y_2) \mathrm{d}G(y_1, y_2)\,. \end{aligned} \tag{3.6}$$

したがって，

$$\mathrm{d}G^*(x_1, x_2) = \frac{1}{\alpha^{(2)}} F(x_1, x_2) \mathrm{d}G(y_1, y_2) \tag{3.7}$$

が成り立つことより，式(3.5)は次のように書き換えられる．

$$\begin{aligned} F^*(x_1, x_2) &= \frac{1}{\alpha^{(2)}} \int \mathbb{P}(X_1 \leq x_1 \wedge y_1, X_2 \leq x_2 \wedge y_2) \mathrm{d}G(x_1, x_2) \\ &= \int \frac{F(x_1 \wedge y_1, x_2 \wedge y_2)}{F(y_1, y_2)} \mathrm{d}G^*(y_1, y_2) \\ &= G^*(x_1, x_2) + \Xi_1^*(x_1, x_2) + \Xi_2^*(x_1, x_2) + F(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} \frac{\mathrm{d}G^*(y_1, y_2)}{F(y_1, y_2)}\,, \end{aligned} \tag{3.8}$$

ここで

$$\begin{aligned} \Xi_1^*(x_1, x_2) &= \int_{x_1}^{\infty} \int_{0}^{x_2} \frac{\mathrm{d}G^*(y_1, y_2)}{F(y_1, y_2)} \\ &= \mathbb{P}\left(Y_1 > x_1, X_1 \leq x_1, Y_2 \leq x_2 | X_1 \leq Y_1, X_2 \leq Y_2\right)\,, \end{aligned}$$

および

$$\begin{aligned} \Xi_2^*(x_1, x_2) &= \int_{0}^{x_1} \int_{x_2}^{\infty} \frac{\mathrm{d}G^*(y_1, y_2)}{F(y_1, y_2)} \\ &= \mathbb{P}\left(Y_2 > x_2, X_2 \leq x_2, Y_1 \leq x_1 | X_1 \leq Y_1, X_2 \leq Y_2\right) \end{aligned}$$

である．式(3.8)を評価するため，次の量を定義する．

$$K^*(x_1, x_2) \equiv F^*(x_1, x_2) - G^*(x_1, x_2) - \Xi_1^*(x_1, x_2) - \Xi_2^*(x_1, x_2)\,. \tag{3.9}$$

$(X_1, X_2)$ および $(Y_1, Y_2)$ が独立であるという仮定のもと，式(3.8)を用いると次が得られる．

$$
\begin{aligned}
K^*(X_1, X_2) &= \mathbb{P}(X_1 \le x_1, x_1 < Y_1, X_2 \le x_2, x_1 < Y_1 | X_1 \le Y_1, X_2 \le Y_2) \\
&= \frac{1}{\alpha^{(2)}} \mathbb{P}(X_1 \le x_1, x_1 < Y_1, X_2 \le x_2, x_2 < Y_2) \\
&= \frac{1}{\alpha^{(2)}} F(x_1, x_2)\, \mathbb{P}(x_1 < Y_1, x_2 < Y_2) \\
&= \frac{1}{\alpha^{(2)}} F(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} \mathrm{d}G(y_1, y_2)\ .
\end{aligned} \tag{3.10}
$$

式(3.10)と式(3.7)を組み合わせると

$$
K^*(x_1, x_2) = F(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} \frac{\mathrm{d}G(y_1, y_2)}{F(y_1, y_2)} \tag{3.11}
$$

が得られる．$(X_{1,i}, X_{2,i}, Y_{1,i}, Y_{2,i})$ $(i = 1, \ldots, n)$ を，条件 $X_{1,i} \le Y_{1,i}$ および $X_{2,i} \le Y_{2,i}$ を満たす観測データとする．2 変数関数 $K^*$ は，その経験的関数

$$
K_n^*(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(X_{1,i} \le x_1 < Y_{1,i}, X_{2,i} \le x_2 \le Y_{2,i}\right) \tag{3.12}
$$

によって直接推定できる．$G^*$ の標本推定量 $G_n^*(x_1, x_2)$ は，観測された $(Y_1, Y_2)$ の単純な経験分布関数として次で与えられる．

$$
G_n^*(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(Y_{1,i} < x_1, Y_{2,i} \le x_2\right)\ . \tag{3.13}
$$

したがって，ノンパラメトリック推定量 $F_n^{(0)}$ は次のように導入できる．

$$
K_n^*(x_1, x_2) = F_n^{(0)}(x_1, x_2) \int_{x_1}^{\infty} \int_{x_2}^{\infty} \frac{\mathrm{d}G_n^*(y_1, y_2)}{F_n^{(0)}(y_1, y_2)}\ . \tag{3.14}
$$

しかし，式(3.14)には一意な解が存在しない．もし $F_n^{(0)}$ が式(3.14)を満たすならば，定数 $k$ に対して $kF_n^{(0)}$ もまたこの式を満たす．$F_n^{(0)}$ が次の点

$$
(X_{1,n+1}, X_{2,n+1}) \equiv \left( \max_{1 \le i \le n} \{X_{1,i}\} + \frac{1}{n}, \max_{1 \le i \le n} \{X_{2,i}\} + \frac{1}{n} \right)
$$

で 1 となるように指定する．つまり

$$
F_n^{(0)}(X_{1,n+1}, X_{2,n+1}) = 1\ .
$$

**繰り返し演算のアルゴリズム**

推定量 $K^*(x_1, x_2)$ は，$2n$ 個の点 $(X_{1,i}, X_{2,i})$, $i = 1, \ldots, n$ および $(Y_{1,i}, Y_{2,i})$ $(i = 1, \ldots, n)$ で決定される．

Step 1. (1)次を定義する．

$$
(X_{1,n+1}, X_{2,n+1}, Y_{1,n+1}, Y_{2,n+1}) = \left(0, 0, \max_{1 \le i \le n} \{X_{1,i}\} + \frac{1}{n}, \max_{1 \le i \le n} \{X_{2,i}\} + \frac{1}{n}\right) \tag{3.15}
$$

そして，これを $(n+1)$ 番目の観測点として追加する．

(2)この新しい観測点における $F_n^{(0)}$ の値を割り当てる．

$$
F_n^{(0)}(X_{1,n+1}, X_{2,n+1}) = 0, \tag{3.16}
$$

$$
F_n^{(0)}(Y_{1,n+1}, Y_{2,n+1}) = 0\ . \tag{3.17}
$$

Step 2. 式(3.12)を使用して，$(s_1, s_2)$=$(Y_{1,i}, Y_{2,i})$, $(i$=$1, \ldots, n)$ および $(X_{1,i}, X_{2,i})$, $(i$=$1, \ldots, n)$ における $K_n^*(s_1, s_2)$ を計算する．

$$K_n^*(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{1,i} \leq x_1 < Y_{1,i}, X_{2,i} \leq x_2 \leq Y_{2,i}) \, .$$

Step 3. (1) $\{(Y_{1,i}, Y_{2,i}), (X_{1,i}, X_{2,i})\ i = 1, \ldots, n\}$ の中から，次の条件を満たす点 $(x_1, x_2)$ を探索する．

$\{(Y_{1,i}, Y_{2,i}) : Y_{1,i} > x_1, Y_{2,i} > x_2,\ i = 1, \ldots, n+1\}$ におけるすべての $F_n^{(0)}$ の値が既知である．

条件を満たさない場合，$\{(Y_{1,i}, Y_{2,i}) : Y_{1,i} > x_1, Y_{2,i} > x_2,\ i = 1, \ldots, n+1\}$ 内の点 $(s_1, s_2)$ で $F_n^{(0)}$ の値が未知であるものが存在する．$(s_1, s_2)$ を条件に合致するか確認する．

$F_n^{(0)}(Y_{1,n+1}, Y_{2,n+1}) = 1$ が既知であり，$(Y_{1,n+1}, Y_{2,n+1}) = 1$ も既知であるため，上記の手順で必要な $(x_1, x_2)$ が得られる．

(2) $(x_1, x_2)$ における $F_n^{(0)}$ の値は次のようにして求められる．

$$F_n^{(0)}(x_1, x_2) = \frac{(n+1)K_n^*(x_1, x_2)}{1 + \displaystyle\sum_{\substack{(Y_{1,i}, Y_{2,i}): \\ Y_{1,i} > x_1, Y_{2,i} > x_2}} \frac{1}{nF_n^{(0)}(Y_{1,i}, Y_{2,i})}} \, . \tag{3.18}$$

Step 4. $F_n^{(0)}$ の値が $^\forall 2n$ 点 $(Y_{1,i}, Y_{2,i})$ および $(X_{1,i}, X_{2,i})$, $i = 1, \ldots, n$ に対して既知になるまで，Step 1–3 を繰り返す．

Step 5. 最後に，$F_n$ およびその周辺分布 $F_{1n}$ と $F_{2n}$ を次で計算する．

$$F_n(x_1, x_2) = \frac{1}{n} \sum_{\substack{X_{1,i} \leq x_1, \\ X_{2,i} \leq x_2}} \frac{F_n^{(0)}(X_{1,i}, X_{2,i})}{K_n^*(X_{1,i}, X_{2,i})} \, ,$$

$$F_{1n}(x_1) = F_n(x_1, \infty) \, ,$$

$$F_{2n}(x_2) = F_n(\infty, x_2) \, .$$

このようにして，2 変数右切断データにおける分布関数とその周辺分布関数を推定できる．van der Laan (1996)や Quale and van der Laan (2000)は $F_n$ を

$$F_n(x_1, x_2) = \int_0^{x_1} \int_0^{x_2} \frac{\mathrm{d}F_n^*(x_1', x_2')}{\displaystyle\int_{x_1'}^{\infty} \int_{x_2'}^{\infty} \frac{\mathrm{d}F_n^*(y_1, y_2)}{F_n(y_1, y_2)}} \tag{3.19}$$

という反復解で表現している．2 変数の場合も推定量の分散は影響関数を用いて評価できるが (van der Laan, 1996)，この場合は分散が閉じた式で表現されないため，例えばブートストラップリサンプリングを用いる方が簡単である(たとえば Quale and van der Laan, 2000)．変数 $(X_1, X_2)$ を $(M_1, M_2)$ に変換することで，図 6 の領域 A に属する銀河サンプルから目的の 2 変数光度関数を得ることができる．

### 3.2.3 天文学における 2 変数選択問題への拡張

しかし，上で求めた推定量について注意しておかねばならない重要な問題がある．第 3.2.1 節で概略を見たように，van der Laan (1996)および Huang et al. (2001)で扱われているデータ選択は，そのままでは天文データの 2 バンド選択とは異なっている．これを理解するため，ある距離指標 $D$ での銀河の 2 バンド選択を考える(図 6)．van der Laan (1996)の 2 変数切断は，$M_{1,i} \leq M^{\lim}{}_1(D)$, $M_{2,i} \leq M^{\lim}{}_2(D)$ という条件が満たされている場合であり，図 6 における領

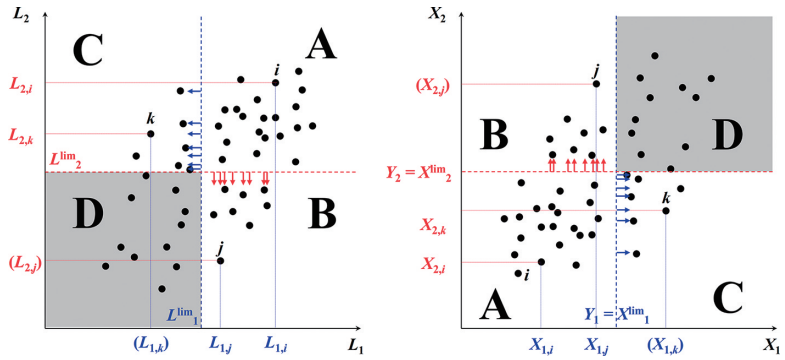


図 6. 光度 $L$ および(換算)絶対等級 $X$ で表した 2 変数光度選択．左パネルが光度による表現，右パネルが換算絶対等級による表現である．領域 A, B, C, D は図 5 に対応している．領域 D は天体は本来は存在するものの，ある距離指標 $D$ での検出限界によって切断されており，データには含まれないため影をつけて示している．銀河 $i$ はバンド 1, 2 ともに検出され，測定値を持つ．銀河 $j$ はバンド 1 では検出されているが，バンド 2 の検出限界を下回るため，得られるのは上限値 $L^{\lim}{}_2(D)$ ないし下限値 $Y_2 = X^{\lim}{}_2(D)$ のみである．一方銀河 $k$ はバンド 2 で検出されているがバンド 1 では上限値 $L^{\lim}{}_1(D)$ ないし下限値 $Y_1 = X^{\lim}{}_1(D)$ のみが得られる．即ち銀河 $j, k$ は打ち切り(censored)データである．

域 A に対応する．しかし，前節で述べたように，多波長銀河探査ではあるバンドで検出されている銀河が別のバンドでは検出されないことも多い．この場合は銀河が存在することは分かっているが，別のバンドでは光度の上限値しか得られない．これは切断ではなく打ち切りデータに対応する状況である．図 5 および図 6 の領域 B および C に入るデータ点が打ち切りを受けている．しかし，これらのデータも天体の光度に関する情報を持っており，推定から除外してしまうことでデータに偏りが生じる惧れがある．前節で紹介した天文学で用いられる方法では，主選択バンドを設定し，もう 1 つのバンドの上限値を含む情報を付加する．これは Quale and van der Laan (2000) の示した以下の手順で定式化できる．

まず，データが主選択バンド 1 でのみ切断されていると仮定する．これは $Y_2 = \infty$ を設定することと同等である．このとき，測度 $\mathrm{d}G_n^*$ は

$$\mathrm{d}G_n^* = g_1(y_1)\delta_{\mathrm{D}}(\infty - y_2)\mathrm{d}y_1\mathrm{d}y_2$$

のように表される．これにより次が導かれる．

$$\mathrm{d}F_n(x_1, x_2)\int_{x_1}^{\infty}\int_{x_2}^{\infty}\frac{\mathrm{d}G_n^*(y_1, y_2)}{F_n(y_1, y_2)} = \mathrm{d}F_n(x_1, x_2)\int_{x_1}^{\infty}\frac{g_1(y_1)\mathrm{d}y_1)}{F_n(y_1, \infty)} = \mathrm{d}F_n^*(x_1, x_2)\,.$$

したがって

$$\mathrm{d}F_n(x_1, \infty)\mathrm{d}F_n(x_1, x_2)\int_{x_1}^{\infty}\frac{g_1(y_1)\mathrm{d}y_1}{F_n(y_1, \infty)} = \mathrm{d}F_n^*(x_1, \infty)\,.$$

これは 1 変数の場合と全く同じであり，解は積極限推定量である．これは生存関数 $S$ を通じて導出できる．

周辺サンプル $(X_1, X_1)$, $(X_1 \le Y_1)$ に対する積極限推定量を $S_n^{\mathrm{PL}}(x_1, \infty)$ とすると，

$$F_n(x_1, \infty) = 1 - S_n^{\mathrm{PL}}(x_1, \infty)\,,$$

より

$$F_n(x_1,x_2) = \int_0^{x_1}\int_0^{x_2} \frac{\mathrm{d}F_n^*(x_1,x_2)}{\int_0^{x_1}\int_0^{x_2} \frac{\mathrm{d}G_{n1}^*(y_1)}{1-S_n^{\mathrm{PL}}(y_1,\infty)}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}(X_{1,i}\le x_1, X_{2,i}\le x_2, X_{1,i}\le Y_{1,i}, X_{2,i}\le Y_{2,i})}{\frac{1}{n}\sum_{j=1}^{n}\frac{\mathbb{1}(X_{1,i}\le Y_{1,j}, X_{1,i}\le Y_{1,i})}{1-S_n^{\mathrm{PL}}(Y_{1,j},\infty)}} = \sum_{i=1}^{n} \frac{\mathbb{1}(X_{1,i}\le x_1, X_{2,i}\le x_2, X_{1,i}\le Y_{1,i}, X_{2,i}\le Y_{2,i})}{\sum_{j=1}^{n}\frac{\mathbb{1}(X_{1,i}\le Y_{1,j}, X_{1,i}\le Y_{1,i})}{1-S_n^{\mathrm{PL}}(Y_{1,j},\infty)}}\,. \tag{3.20}$$

が得られる．1 バンド切断の場合に注目すべきは，2 バンド切断の場合とは対照的に，ノンパラメトリック最尤推定量(NPMLE)が陽に得られることである．

天文学における応用では，主バンドの切断と二次バンドの打ち切りを考慮する必要がある．この場合，以前に研究された天文学の例では，打ち切りはバンド 2 での二次選択でのみ発生することに注意する．推定量は式(3.20)における $S_n^*$ に，打ち切りデータに対応する生存関数 $S$ を単純に代入することで実現できる（e.g., van der Laan, 1996; Quale and van der Laan, 2000）．打ち切りデータを扱うためのプラグイン推定量としては，例えば Dąbrowska 推定量 $S_n^{\mathrm{Dab}}$ が適切である（Dabrowska, 1988）[付録 A の式(A.18)を参照]．打ち切りはバンド 2 における $Y_2$ を通じて生じる．変数 $X_1$ と $X_2$ が分離可能であるという事実により，この方法は第 3.2.1 節での天文学的 2 変数分布の古典的推定法の厳密な導出となっている．

### 3.2.4 2 変数光度関数の最終的な「真の」推定量

ここでついに 2 変数光度関数の「真の」推定量をどのように推定するかについて議論し，本論を締めくくる．「真の」2 変数光度関数とは何を意味するかを明確にする．図 5 および図 6 に戻ると，van der Laan 推定量ではこれらの図における領域 A のみを考慮することになる．つまり，バンド 1 選択サンプルとバンド 2 選択サンプルの和集合に含まれる領域 B および C の標本を落とすことで，情報量を大幅に損失している．一方で，古典的な天文学的手法を採用した場合，バンド 1 またはバンド 2 選択サンプルのどちらかに依存することになる．この場合，領域 B または C のサンプル情報を失わないが，サンプル選択は構造上バイアスがかかっており，その結果として天文学的結論も必然的に偏ったものとなる．

2 バンド選択データから偏りのないサンプルを構築しようとする場合，図 6 の領域 A, B, C の和集合からから得られる最大限の情報を利用する推定量が必要である．これを達成するには，両方の手法の長所を取り入れるべきである．先の議論から明らかなように，2 変数切断データに対する van der Laan 推定量を採用し，2 変数生存関数に対して Dąbrowska 推定量

$$S_n^{\mathrm{Dab}}(x_1,x_2) = S_n^{\mathrm{Dab}}(x_1,0)S_n^{\mathrm{Dab}}(0,x_2)\prod_{\substack{0<x_1'\le x_1\\ 0<x_2'\le x_2}}[1-Q_n(\Delta x_1',\Delta x_2')] \tag{3.21}$$

を代入することでこれを実現できる．推定量の構築および関連する式は長くなるため，付録 A に提示する．式(3.21)では，打ち切り指標の集合 $(\Delta_{1,i},\Delta_{2,i})$ $(i=1,\dots,n)$ を使用する．実用上は，最終的な NPMLE の陽な表現がないため取り扱いは容易ではないが，式(3.19)および式(3.21)のセットが，2 バンド選択下での銀河の 2 変数光度関数の最終的な推定量である．

データの情報を完備にするため，式(3.21)において適切な打ち切り指標の集合 $(\Delta_{1,i},\Delta_{2,i})$ $(i=1,\dots,n)$ を割り当てる．データの元の構造は次の通りである

$$[\text{バンド 1 の等級}, \text{バンド 2 の等級}, \text{赤方偏移}] = [m_{\lambda_1,i}, m_{\lambda_2,i}, z_i]\,,$$

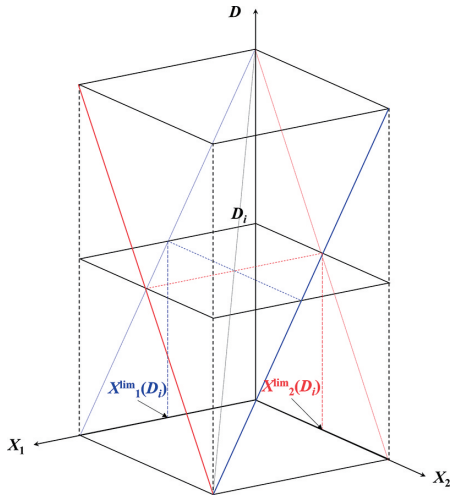


図 7. 2 変数光度選択でのデータ選択効果における距離指標 $D$ への依存性．この図では銀河の光度および検出限界は換算絶対等級 $(X_1, X_2)$, $(X^{\lim}{}_1, X^{\lim}{}_2)$ で表示している．底面で示した観測可能範囲は図 5 および図 6 における領域 A に対応し，バンド 1, 2 両方で検出され，測定値を持つ．距離 $D_1$ ではその範囲が狭まり，領域 B，C に対応する打ち切りデータが含まれる．天体の多波長探査データでは距離方向に積分されているため，異なった検出限界のデータが混在している．

対応する限界等級は $m^{\lim}{}_1$ および $m^{\lim}{}_2$ である．光度関数を構築するために，まずデータを

$$[\text{バンド 1 の光度}, \text{バンド 2 の光度}, \text{光度距離}] = [L_{\lambda_1,i}, L_{\lambda_2,i}, d_{\rm L}(z_i)] \, ,$$

あるいはまったく等価であるが

$$[\text{バンド 1 の絶対等級}, \text{バンド 2 の絶対等級}, \text{距離指数}] = [M_{\lambda_1,i}, M_{\lambda_2,i}, D_i]$$

という形式に変換する．検出限界も $(L^{\lim}{}_1, L^{\lim}{}_2)$ または $(M^{\lim}{}_1, M^{\lim}{}_2)$ に変換される．さらにこれらを簡略化のために，換算絶対等級として $[X_{1,i}, X_{2,i}, D_i]$ に変換し，関連する検出限界 $(X^{\lim}{}_1, X^{\lim}{}_2)$ を付け加えて，詳細な統計的議論の記述を行う．

換算絶対等級単位では，切断変数 $(Y_{1,i}, Y_{2,i})$ は次のように定義される

$$Y_{1,i} = X^{\lim}{}_{1,i}(D_i) \, , \tag{3.22}$$

$$Y_{2,i} = X^{\lim}{}_{2,i}(D_i) \, . \tag{3.23}$$

$X_1 < Y_1$ または $X_2 < Y_2$（「かつ」ではない）を満たす銀河が観測され，2 バンド選択サンプルに含まれる．これは図 7 で概略的に示されている．図 6 の領域 A では，$X_1 \leq Y_1$ かつ $X_2 \leq Y_2$ を満たす銀河が含まれている．この場合，データに打ち切りは存在しない．領域 B では，バンド 1 での検出があるが，バンド 2 では検出がない．つまり，$X_1, X_2 = Y_2, \Delta_1 = 1, \Delta_2 = 0$ として打ち切りデータが存在する．同様に領域 C では，$X_1 = Y_1, X_2, \Delta_1 = 0, \Delta_2 = 1$ としてデータが存在する．したがって，次の構造

$$[X_{1,i}, X_{2,i}, \Delta_{1,i}, \Delta_{2,i}] \tag{3.24}$$

を持つデータセットを準備する．この形式で記述されたデータと式(3.19)および式(3.21)を用

いることで，本来の 2 変数光度関数を推定することができる．

## 4. まとめと議論

本研究では，天文サーベイから銀河の光度関数を推定する手法を紹介した．この種の解析は，切断データ解析として知られている．まず最初に，1 変数光度関数の場合を導入した．切断データから 1 変数光度関数を推定する手法は，天文学コミュニティにおいて Lynden-Bell (1971)によって導入されたのが最初である．Woodroofe (1985)は Lynden-Bell の推定量を再定式化した．彼はこれを統計学コミュニティに導入し，その収束特性について数学的に厳密な証明を行った．

しかし，この問題の 2 変数版である 2 変数光度関数の厳密な方法論は，天文学で適切な定式化に基づいて議論されることはほとんどなかった．天文学における 2 変数光度関数は，主選択バンドから推定された 1 変数光度関数に基づいて構築される．次に，二次選択バンドからの条件付き光度関数をカプラン-マイヤー推定量で推定し，それを組み合わせる．しかし，この手法は単一バンド選択に基づくことが原因で，2 変数関数に対して偏った推定をもたらしてしまう．一方，van der Laan (1996)は 2 変数切断データの場合の 2 変数分布関数の推定量を提案した．この手法は 2 変数切断には適しているものの，2 バンド選択の場合には避けられない打ち切りデータの情報を取りこぼしてしまう．そこで我々は，本論文で打ち切り情報も取り入れることができる統合手法を提案した．これは，切断変数と打ち切り変数を設定することで達成できる．この推定法の性能評価と実データへの応用は，今後の研究で数値的検討とともに提示する予定である (Takeuchi et al. 2025, in preparation)．

最後に，天文光度関数の推定についての注意を述べる．すべての関連する議論において，$X$ と $Y$ の独立性が仮定されている．しかし，図 2 および図 7 に示されるように，この独立性が完全に成立するかどうかは慎重に検討されるべきである．特に，最近の深いサーベイでは，サンプルにおける時間進化の影響は無視できない．この問題は一部の著者によって指摘され，検討されるようになってきている (たとえば Efron and Petrosian, 1992; Chiou et al., 2018; Chiou et al., 2019)．今後の巨大サーベイプロジェクトでは，この問題がますます深刻になるため，より数学的に厳密な議論が必要となると考えられる．

### 注．

[1] 定常ユークリッド空間とは，宇宙膨張のように空間自体の計量(metric)が時間変化することのないユークリッド空間を意味する宇宙論の用語である．定常ユークリッド空間での議論では，単位振動数当たりの量 $L_\nu$ および $S_\nu$ で考えても全く同じ関係が成り立つ．

[2] 宇宙物理学分野では，常用対数を log, 自然対数を ln で表すのが習慣である．本稿でもこれ以降は底の 10 を省略する．

[3] 銀河の集団化の度合いをクラスタリングと呼ぶ．クラスタリングを $n$ 次 factorial cumulant (宇宙論では $n$ 点相関関数と呼ぶ)を用いて定量的に評価する方法を相関関数解析という．

[4] 天文学で用いられる長さの単位 Mpc (メガパーセク)は $3.26 \times 10^{24}$ cm に対応する．

[5] 超銀河団と呼ばれる構造は長く伸びた形状の銀河の集中を指し，銀河団の集団ではない．現在はフィラメント構造と呼ぶのが一般的になっている．

[6] Dorota M. Dąbrowska は論文中で自身の名前を Dabrowska と綴っているが，本稿ではオリジナルの綴りを採用する．

謝　辞

本研究は日本学術振興会（JSPS）科学研究費補助金（21H01128 および 24H00247）の補助を受けて行っている．また本研究の一部は統計数理研究所共同研究費「機械学習的宇宙構造論: 構造形成から銀河進化へ」（2024-ISMCRP-2025）の支援も受けて行った．

本論文は天文学と統計学野両方にまたがる内容を扱っており，両分野の慣習の違いのため初稿は未整理な部分やミスが多数残っていた．辛抱強く丁寧に議論や導出を追い，極めて有用なコメントをしていただいた 2 人の査読者に深く感謝申し上げる．統計数理研究所の江村 剛志氏には本特集号にご招待いただき，また内容についても大変重要な示唆を頂いた．氏の大きな貢献に心から感謝する．

## 付録 A.　2 変数打ち切りデータに対する Dąbrowska 推定量

本章では，2 変数打ち切りデータの分布関数に対する Dąbrowska 推定量[6]を紹介する（Dabrowska, 1988）．ここでは本論で必要な最低限の議論に留めたが，Dąbrowska 推定量の入門的解説はたとえば杉本・田中（2023, 第 2.1 節）を参照されたい．また漸近的性質等の詳細は Dabrowska（1989）を参照されることをお勧めする．

2 変数打ち切りデータ $(X_1, X_2)$ と打ち切り変数 $(Y_1, Y_2)$ のセットを考える．打ち切りデータ $(X_1, X_2)$ と $(Y_1, Y_2)$ の分布関数をそれぞれ $F$ と $G$ とする．本文での議論に即し，左側打ち切り問題を扱う．すなわち，$X_1 \leq Y_1$ かつ $X_2 \leq Y_2$ のときにのみ $(X_1, X_2)$ が観測される．2 変数生存関数 $S(x_1, x_2)$ を次のように定義する．

$$S(x_1, x_2) \equiv \mathbb{P}\left(X_1 > x_1, X_2 > x_2\right).$$

これより，次が成り立つ．

$$F(x_1, x_2) = 1 - S(x_1, 0) - S(0, x_2) + S(x_1, x_2)\,.$$

同様に，分布 $G$ に対して生存関数 $R(x_1, x_2)$ を定義する．さらに次を定義する．

$$\begin{aligned}(\tilde{X_1}, \tilde{X_2}) &\equiv (X_1 \wedge Y_1, X_2 \wedge Y_2)\,,\\ \Delta_1 &\equiv \mathbb{1}(X_1 \leq Y_1)\,,\\ \Delta_2 &\equiv \mathbb{1}(X_2 \leq Y_2)\,.\end{aligned}$$

ここで $(X_1, X_2)$ と $(Y_1, Y_2)$ が独立であると仮定する．

まず以下の関数を定義する．

$$\begin{aligned}H(x_1, x_2) &\equiv \mathbb{P}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2)\,,\\ K_{11}(x_1, x_2) &\equiv \mathbb{P}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_1 = 1, \Delta_2 = 1\,,\\ K_{10}(x_1, x_2) &\equiv \mathbb{P}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_1 = 1)\,,\\ K_{01}(x_1, x_2) &\equiv \mathbb{P}(\tilde{X}_1 > x_1, \tilde{X}_2 > x_2, \Delta_2 = 1)\,.\end{aligned}$$

$H(x_1, x_2) > 0$ を満たす $(x_1, x_2)$ に対して次が成り立つ．

$$H(x_1, x_2) = R(x_1, x_2) S(x_1, x_2)\,, \tag{A.1}$$

$$K_{11}(\mathrm{d}x_1, \mathrm{d}x_2) = R(x_1-, x_2-) S(\mathrm{d}x_1, \mathrm{d}x_2)\,, \tag{A.2}$$

$$K_{10}(\mathrm{d}x_1, x_2) = R(x_1-, x_2) S(\mathrm{d}x_1, x_2)\,, \tag{A.3}$$

$$K_{01}(x_1, \mathrm{d}x_2) = R(x_1, x_2-) S(x_1, \mathrm{d}x_2)\,. \tag{A.4}$$

2 変数累積ハザード関数を次のように定義する．

$$\Lambda_{11}(x_1,x_2)=\int_0^{x_1}\int_0^{x_2}\frac{K_{11}(\mathrm{d}x_1',\mathrm{d}x_2')}{H(x_1'-,x_2'-)}\,, \tag{A.5}$$

$$\Lambda_{10}(x_1,x_2)=-\int_0^{x_1}\frac{K_{10}(\mathrm{d}x_1',x_2)}{H(x_1'-,x_2)}\,, \tag{A.6}$$

$$\Lambda_{01}(x_1,x_2)=-\int_0^{x_2}\frac{K_{01}(x_1,\mathrm{d}x_2')}{H(x_1,x_2'-)}\,. \tag{A.7}$$

式(A.1)–(A.7)の標本推定量は次のように表される．

$$H_n(x_1,x_2)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\tilde{X}_{1,i}>x_1,\tilde{X}_{2,i}>x_2)\,, \tag{A.8}$$

$$K_{n11}(x_1,x_2)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\tilde{X}_1>x_1,\tilde{X}_2>x_2,\Delta_{1,i}=1,\Delta_{2,i}=1)\,, \tag{A.9}$$

$$K_{n10}(x_1,x_2)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\tilde{X}_1>x_1,\tilde{X}_2>x_2,\Delta_{1,i}=1)\,, \tag{A.10}$$

$$K_{n01}(x_1,x_2)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\tilde{X}_1>x_1,\tilde{X}_2>x_2,\Delta_{2,i}=1)\,, \tag{A.11}$$

$$\Lambda_{n11}(x_1,x_2)=\int_0^{x_1}\int_0^{x_2}\frac{K_{n11}(\mathrm{d}x_1',\mathrm{d}x_2')}{H_n(x_1'-,x_2'-)}\,, \tag{A.12}$$

$$\Lambda_{n10}(x_1,x_2)=-\int_0^{x_1}\frac{K_{n10}(\mathrm{d}x_1',x_2)}{H_n(x_1'-,x_2)}\,, \tag{A.13}$$

$$\Lambda_{n01}(x_1,x_2)=-\int_0^{x_2}\frac{K_{n01}(x_1,\mathrm{d}x_2')}{H_n(x_1,x_2'-)}\,. \tag{A.14}$$

さらに，記号を簡略化するために，任意の 2 変数関数 $f(s,t)$ に対して次の記号を定義する．

$$f(\Delta s,\Delta t)=f(s,t)-f(s,t-)-f(s-,t)+f(s-,t-)\,, \tag{A.15}$$

$$f(\Delta s,t)=f(s,t)-f(s-,t)\,, \tag{A.16}$$

$$f(s,\Delta t)=f(s,t)-f(s,t-)\,. \tag{A.17}$$

式(A.7)–(A.17)を組み合わせることで，$S(x_1,x_2)$ の自然な推定量が次の式で与えられる．

$$S_n^{\mathrm{Dab}}(x_1,x_2)=S_n^{\mathrm{Dab}}(x_1,0)S_n^{\mathrm{Dab}}(0,x_2)\prod_{\substack{0<x_1'\le x_1\\0<x_2'\le x_2}}\left[1-Q_n(\Delta x_1'\,\Delta x_2')\right]\,, \tag{A.18}$$

$$Q_n(\Delta x_1'\,\Delta x_2')\equiv\frac{\Lambda_{n10}(\Delta x_1',x_2-)\Lambda_{n01}(x_1-,\Delta x_2)-\Lambda_{n11}(\Delta x_1'\,\Delta x_2')}{[1-\Lambda_{n10}(\Delta x_1',x_2-)]\,[1-\Lambda_{n01}(x_1'-,\Delta x_2)]}\,, \tag{A.19}$$

$$S_n^{\mathrm{Dab}}(x_1,0)\equiv\prod_{x_1'\le x_1}\left[1-\Lambda_{n10}(\Delta x_1',0)\right]\,, \tag{A.20}$$

$$S_n^{\mathrm{Dab}}(0,x_2)\equiv\prod_{x_2'\le x_2}\left[1-\Lambda_{n01}(0,\Delta x_2')\right]\,. \tag{A.21}$$

これらの推定量(式(A.18)–(A.21))を実用的な形に変形すると以下のようになる．

$$\Lambda_{n10}(\Delta x_1',0)=\frac{\sum_{i=1}^{n}\mathbb{1}(\tilde{X}_{1,i}=x_1')\mathbb{1}(\Delta_{1,i}=1)}{\sum_{i=1}^{n}\mathbb{1}(\tilde{X}_{1,i}\ge x_1')}\,, \tag{A.22}$$

$$\Lambda_{n10}(0, \Delta x_2') = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{2,i} = x_2')\mathbb{1}(\Delta_{2,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{2,i} \geq x_2')} . \tag{A.23}$$

同様に，$Q_n$ の式を次のように導出する．

$$\Lambda_{n11}(\Delta x_1', \Delta x_2') = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} = x_1', \tilde{X}_{2,i} = x_2')\mathbb{1}(\Delta_{1,i} = 1, \Delta_{2,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x_1', \tilde{X}_{2,i} \geq x_2')} , \tag{A.24}$$

$$\Lambda_{n10}(\Delta x_1', x_2'-) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} = x_1', \tilde{X}_{2,i} \geq x_2')\mathbb{1}(\Delta_{1,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x_1', \tilde{X}_{2,i} \geq x_2')} , \tag{A.25}$$

$$\Lambda_{n01}(x_1'-, \Delta x_2'-) = \frac{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x_1', \tilde{X}_{2,i} = x_2')\mathbb{1}(\Delta_{2,i} = 1)}{\sum_{i=1}^n \mathbb{1}(\tilde{X}_{1,i} \geq x_1', \tilde{X}_{2,i} \geq x_2')} . \tag{A.26}$$

以上の一連の式を使用して，Dąbrowska 推定量を計算することができる．

## 参 考 文 献


Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins Series in the Mathematical Sciences, Springer, New York.

Binggeli, B., Sandage, A. and Tammann, G. A. (1988). The luminosity function of galaxies, *Annual Review of Astronomy and Astrophysics*, **26**, 509–560, http://dx.doi.org/10.1146/annurev.aa.26.090188.002453.

Buat, V., Takeuchi, T. T., Burgarella, D., Giovannoli, E. and Murata, K. L. (2009). The infrared emission of ultraviolet-selected galaxies from z = 0 to z = 1, *Astronomy & Astrophysics*, **507**(2), 693–704, http://dx.doi.org/10.1051/0004-6361/200912024.

Caditz, D. and Petrosian, V. (1993). Smoothed nonparametric estimation of the luminosity function for flux-limited samples, *The Astrophysical Journal*, **416**, 450–457, http://dx.doi.org/10.1086/173250.

Chiou, S. H., Qian, J., Mormino, E. and Betensky, R. A. (2018). Permutation tests for general dependent truncation, *Computational Statistics & Data Analysis*, **128**(C), 308–324.

Chiou, S. H., Austin, M. D., Jing, Q. and Betensky, R. A. (2019). Transformation model estimation of survival under dependent truncation and independent censoring, *Statistical Methods in Medical Research*, **28**(12), 3785–3798.

Chołoniewski, J. (1985). Bivariate luminosity function of E and S0 galaxies, *Monthly Notices of the Royal Astronomical Society*, **214**, 197–202, http://dx.doi.org/10.1093/mnras/214.2.197.

Chołoniewski, J. (1987). On Lynden-Bell's method for the determination of the luminosity function, *Monthly Notices of the Royal Astronomical Society*, **226**, 273–280, http://dx.doi.org/10.1093/mnras/226.2.273.

Cowie, L. L., Songaila, A., Hu, E. M. and Cohen, J. G. (1996). New insight on galaxy formation and evolution from Keck spectroscopy of the Hawaii deep fields, *The Astronomical Journal*, **112**, 839–864, http://dx.doi.org/10.1086/118058.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane, *The Annals of Statistics*, **16**(4), 1475–1489, http://dx.doi.org/10.1214/aos/1176351049.

Dabrowska, D. M. (1989). Kaplan-Meier estimate on the plane: Weak convergence, LIL, and the bootstrap, *Journal of Multivariate Analysis*, **29**(2), 308–325, http://dx.doi.org/https://doi.org/10.1016/0047-259X(89)90030-4.

de Lapparent, V., Geller, M. J. and Huchra, J. P. (1986). A slice of the universe, *Astrophysical Journal*, **302**, L1–L4.

de Lapparent, V., Geller, M. J. and Huchra, J. P. (1989). The luminosity function for the CfA redshift survey slices, *The Astrophysical Journal*, **343**, 1–17, http://dx.doi.org/10.1086/167679.

Eales, S. (1993). Direct construction of the galaxy luminosity function as a function of redshift, *The Astrophysical Journal*, **404**, 51–62, http://dx.doi.org/10.1086/172257.

Efron, B. and Petrosian, V. (1992). A simple test of independence for truncated data with applications to redshift surveys, *The Astrophysical Journal*, **399**, 345–352, http://dx.doi.org/10.1086/171931.

Ellis, R. S., Colless, M., Broadhurst, T., Heyl, J. and Glazebrook, K. (1996). Autofib redshift survey — I. Evolution of the galaxy luminosity function, *Monthly Notices of the Royal Astronomical Society*, **280**(1), 235–251, http://dx.doi.org/10.1093/mnras/280.1.235.

Heyl, J., Colless, M., Ellis, R. S. and Broadhurst, T. (1997). Autofib redshift survey — II. Evolution of the galaxy luminosity function by spectral type, *Monthly Notices of the Royal Astronomical Society*, **285**(3), 613–634, http://dx.doi.org/10.1093/mnras/285.3.613.

Huang, J., Vieland, V. J. and Wang, K. (2001). Nonparametric estimation of marginal distributions under bivariate truncation with application to testing for age-of-onset anticipation, *Statistica Sinica*, **11**(4), 1047–1068.

Hubble, E. P. (1936). *Realm of the Nebulae*, Yale University Press, New Haven.

Jackson, J. C. (1974). The analysis of quasar samples, *Monthly Notices of the Royal Astronomical Society*, **166**, 281–296, http://dx.doi.org/10.1093/mnras/166.2.281.

Johnston, R. (2011). Shedding light on the galaxy luminosity function, *Astronomy & Astrophysics Review*, **19**, 41, http://dx.doi.org/10.1007/s00159-011-0041-9.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**(282), 457–481.

Keiding, N. and Gill, R. D. (1990). Random truncation models and Markov processes, *The Annals of Statistics*, **18**(2), 582–602, http://dx.doi.org/10.1214/aos/1176347617.

Keres, D., Yun, M. S. and Young, J. S. (2003). CO luminosity functions for far-infrared- and B-band-selected galaxies and the first estimate for $\Omega_{\mathrm{HI+H_2}}$, *The Astrophysical Journal*, **582**(2), 659–667, http://dx.doi.org/10.1086/344820.

Lilly, S. J., Le Fevre, O., Hammer, F. and Crampton, D. (1996). The Canada-France redshift survey: The luminosity density and star formation history of the universe to $z \sim 1$, *The Astrophysical Journal*, **460**, L1–L4, http://dx.doi.org/10.1086/309975.

Lin, H., Kirshner, R. P., Shectman, S. A., Landy, S. D., Oemler, A., Tucker, D. L. and Schechter, P. L. (1996). The luminosity function of galaxies in the las campanas redshift survey, *The Astrophysical Journal*, **464**, 60–78, http://dx.doi.org/10.1086/177300.

Lin, H., Yee, H. K. C., Carlberg, R. G., Morris, S. L., Sawicki, M., Patton, D. R., Wirth, G. and Shepherd, C. W. (1999). The CNOC2 field galaxy luminosity function. I. A description of luminosity function evolution, *The Astrophysical Journal*, **518**(2), 533–561, http://dx.doi.org/10.1086/307297.

Lynden-Bell, D. (1971). A method of finding distances to double galaxies, *Monthly Notices of the Royal Astronomical Society*, **155**, 95–118, http://dx.doi.org/10.1093/mnras/155.1.95.

Marshall, H. L., Tananbaum, H., Avni, Y. and Zamorani, G. (1983). Analysis of complete quasar samples to obtain parameters of luminosity and evolution functions, *The Astrophysical Journal*, **269**, 35–41, http://dx.doi.org/10.1086/161016.

Mobasher, B., Sharples, R. M. and Ellis, R. S. (1993). A complete galaxy redshift survey-V. Infrared luminosity functions for field galaxies, *Monthly Notices of the Royal Astronomical Society*, **263**, 560–574, http://dx.doi.org/10.1093/mnras/263.3.560.

Pascarelle, S. M., Lanzetta, K. M. and Fernández-Soto, A. (1998). The ultraviolet luminosity density of the universe from photometric redshifts of galaxies in the hubble deep field, *The Astrophysical Journal*, **508**(1), L1–L4, http://dx.doi.org/10.1086/311708.

Quale, C. and van der Laan, M. J. (2000). Inference with bivariate truncated data, *Lifetime Data Analysis*, **6**, 391–408, http://dx.doi.org/10.1023/A:1026513500285.

Sandage, A., Tammann, G. A. and Yahil, A. (1979). The velocity field of bright nearby galaxies. I. The

variation of mean absolute magnitude with redshift for galaxies in a quiet velocity field, *The Astrophysical Journal*, **232**, 352–364, http://dx.doi.org/10.1086/157295.

Sawicki, M. J., Lin, H. and Yee, H. K. C. (1997). Evolution of the galaxy population based on photometric redshifts in the Hubble deep field, *The Astronomical Journal*, **113**, 1–12, http://dx.doi.org/10.1086/118231.

Schechter, P. (1976). An analytic expression for the luminosity function for galaxies, *The Astrophysical Journal*, **203**, 297–306, http://dx.doi.org/10.1086/154079.

Schmidt, M. (1967). Space distribution of quasi-stellar radio sources, *Publications of the Astronomical Society of the Pacific*, **79**(470), 437–438, http://dx.doi.org/10.1086/128576.

Schmidt, M. (1968). Space distribution and luminosity functions of quasi-stellar radio sources, *The Astrophysical Journal*, **151**, 393–409, http://dx.doi.org/10.1086/149446.

Shane, C. D. and Wirtanen, C. A. (1967). *Publications of Lick Observatory*, **22**, part 1.

SubbaRao, M. U., Connolly, A. J., Szalay, A. S. and Koo, D. C. (1996). Luminosity functions from photometric redshifts I: Techniques, *The Astronomical Journal*, **112**, 929–936, http://dx.doi.org/10.1086/118066.

杉本知之, 田中健太 (2023). 2 変量生存時間モデルにおけるコピュラとその利用, 日本統計学会誌, **52**(2), 153–176, http://dx.doi.org/10.11329/jjssj.52.153.

Takeuchi, T. T. (2010). Constructing a bivariate distribution function with given marginals and correlation: Application to the galaxy luminosity function, *Monthly Notices of the Royal Astronomical Society*, **406**(3), 1830–1840, http://dx.doi.org/10.1111/j.1365-2966.2010.16778.x.

Takeuchi, T. T. (2025). in preparation.

Takeuchi, T. T. and Kono, K. T. (2020). Constructing a multivariate distribution function with a vine copula: Towards multivariate luminosity and mass functions, *Monthly Notices of the Royal Astronomical Society*, **498**(3), 4365–4378, http://dx.doi.org/10.1093/mnras/staa2558.

Takeuchi, T. T., Yoshikawa, K. and Ishii, T. T. (2000). Tests of statistical methods for estimating galaxy luminosity function and applications to the Hubble deep field, *The Astrophysical Journal Supplement Series*, **129**(1), 1–31, http://dx.doi.org/10.1086/313409.

Takeuchi, T. T., Ishii, T. T., Dole, H., Dennefeld, M., Lagache, G. and Puget, J. L. (2006). The ISO 170 $\mu$m luminosity function of galaxies, *Astronomy & Astrophysics*, **448**(2), 525–534, http://dx.doi.org/10.1051/0004-6361:20054272.

Takeuchi, T. T., Sakurai, A., Yuan, F.-T., Buat, V. and Burgarella, D. (2013). Far-ultraviolet and far-infrared bivariate luminosity function of galaxies: Complex relation between stellar and dust emission, *Earth, Planets, and Space*, **65**(3), 281–290, http://dx.doi.org/10.5047/eps.2012.06.008.

Tinsley, B. M. (1980). Evolution of the stars and gas in galaxies, *Fundamentals of Cosmic Physics*, **5**, 287–388.

Tinsley, B. M. and Danly, L. (1980). On the density of star formation in the universe, *The Astrophysical Journal*, **242**, 435–442, http://dx.doi.org/10.1086/158477.

van der Laan, M. J. (1996). Nonparametric estimation of the bivariate survival function with truncated data, *Journal of Multivariate Analysis*, **58**(1), 107–131, http://dx.doi.org/https://doi.org/10.1006/jmva.1996.0042.

van der Vaart, A. (1991). On differentiable functionals, *The Annals of Statistics*, **19**(1), 178–204, http://dx.doi.org/10.1214/aos/1176347976.

Wall, J. and Jenkins, C. (2003). *Practical Statistics for Astronomers*, Cambridge Observing Handbooks for Research Astronomers, Cambridge University Press, Cambridge.

Wang, M.-C., Jewell, N. P. and Tsai, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation, *The Annals of Statistics*, **14**(4), 1597–1605, http://dx.doi.org/10.1214/aos/1176350180.

Woodroofe, M. (1985). Estimating a distribution function with truncated data, *The Annals of Statistics*, **13**(1), 163–177.

# Survival Analysis in Astrophysics: Estimation of Galaxy Luminosity Function

Tsutomu T. Takeuchi[1,2]

[1]Division of Particle and Astrophysical Science, Nagoya University
[2]The Institute of Statistical Mathematics

Observational data in astronomy are invariably subject to truncation due to the detection limits of observation instruments. When estimating distribution functions of astronomical statistical quantities from such data, it is appropriate to apply survival analysis of truncated data. However, survival analysis, which has been developed in the field of statistics, was practically unknown in the field of astronomy until the 1980s, and astronomers have developed estimation methods independently of the framework in mathematical statistics. Although many of the methods devised by astronomers finally converge with survival analysis, they have been mathematically unorganized, and statistical discussions became systematic only after the 21st century. In this paper, we discuss the estimation of galaxy luminosity functions as an application of survival analysis of truncated data in astronomy. First, we introduce the characteristics of astronomical data in the univariate case and explain the correspondence with survival analysis. Next, we introduce the bivariate case. The estimation of a bivariate luminosity function can become more complex, since it involves both truncation and censoring, depending on the sampling method of astronomical objects. Here, we make attempt to construct the estimation method of bivariate luminosity function as generally as possible.

Key words: Multiwavelength luminosity, galaxy evolution, luminosity function, survival analysis, truncation.