統計数理(2025)
第 73 巻 第 2 号 375–399

［研究詳解］

多様体学習を用いた銀河進化の新しい定量化

竹内 努[1,2]・クレ スチェータ[3,4]・山形 大青[1]・曹 愛奈[1,6]・内田 舜也[1]・
池田 思朗[2]・福水 健次[2]・加納 龍生[1]・大森 清顕 クリストファ[1,5]・
馬 海霞[1]・施 文[1]・松井 瀬奈[1]

（受付 2024 年 2 月 13 日；改訂 2025 年 2 月 25 日；採択 2 月 25 日）

要　旨

宇宙に存在する物質は，空間的にほぼ一様の状態から始まった．その中のかすかに密度の高い領域が重力で収縮し，最終的に銀河へと成長した．宇宙物理学は，この 130 億年を超える宇宙の歴史における銀河の形成と進化を引き起こす複雑な物理現象を，物理の第一原理から解き明かそうと試みてきた．しかし，説明変数が 100 を超える現在の天文学ビッグデータに対し，このような従来の方法は限界を迎えている．そこで我々は，従来の第一原理からの理論構築とは異なる方法でこの問題へのアプローチを行っている．銀河の多波長光度および宇宙年齢が張る高次元特徴空間中での銀河分布に対し，我々はデータ科学で最近発展してきた多様体学習を適用し，銀河の進化の特徴づけを行った．これにより，我々は銀河多様体と呼ばれる，多波長光度空間内のデータ点に埋め込まれた低次元非線型構造を発見した．そして紫外線，光学，近赤外線の光度空間における銀河の進化が，銀河多様体上の星形成と星の質量進化という 2 つのパラメーターによってよく記述されることを発見した．これら銀河多様体座標を物理量に結び付ける方法についても議論する．

キーワード：銀河進化，銀河形成，星形成率，星質量，多波長光度，多様体学習．

1. はじめに

1.1 大規模銀河探査時代の銀河進化研究

銀河とは，星と星間物質(ガスとダストの混合流体)，暗黒物質からなる巨大な天体であり，観測可能な宇宙の範囲に数千億個におよぶ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．

[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, yamagata.taisei.p4@s.mail.nagoya-u.ac.jp, so.aina.t6@s.mail.nagoya-u.ac.jp, uchida.shunya.i4@s.mail.nagoya-u.ac.jp, kano.ryusei.z5@s.mail.nagoya-u.ac.jp, mhx11235@gmail.com, shiwenbaobao0223@gmail.com, matsui.sena.x7@s.mail.nagoya-u.ac.jp

[2] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3; shiro@ism.ac.jp, fukumizu@ism.ac.jp

[3] 国立天文台 科学研究部：〒 181–8588 東京都三鷹市大沢 2–21–1

[4] スタンフォード大学 カブリ素粒子宇宙物理学・宇宙論研究所：452 Lomita Mall, Stanford, CA 94305–4085; cooray@nagoya-u.jp

[5] セント・メアリー大学 天文学・物理学部：Halifax Nova Scotia, B3H 3C3 Canada; k.omori116@gmail.com

[6] 学習院大学 理学部物理学科：〒 171–8588 東京都豊島区目白 1–5–1

つまり，銀河は形成し，現在の姿に進化してきた時間的に動的に進化する存在である．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半世紀以上にわたり銀河研究の中心であり続けている．

銀河の進化を物理法則から定量的に説明する試みは 1970 年代に始まった．銀河が単一の巨大なガス雲から形成されたという仮定の下で，星の形成とそれに関連する重元素合成の歴史を扱う理論の開発が試みられた．この方向の研究は 1980 年代前半に Tinsley（1980）によって一旦理論体系としては完成されたものの，これで銀河進化の研究が終了とはならなかった．時を同じくして進められてきた宇宙論研究により，銀河は合体して成長することが明らかになってきた．これは，銀河の進化が周囲の銀河の密度とガス密度に大きく依存する非常に複雑なプロセスであることを示している．このように，銀河の進化は周囲の銀河の密度やガス密度など，銀河が置かれた環境に大きく依存する非常に複雑な過程であることが判明した．新たな銀河進化を記述する方程式は記号的には次のように表される．

$$
\begin{aligned}
&\mathrm{SFR}(t) = f_1(\mathrm{SFR}, \mathcal{M}_*, \mathcal{M}_{\mathrm{mol}}, \mathcal{M}_{\mathrm{HI}}, \mathcal{M}_{\mathrm{dust}}, \mathcal{M}_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots),\\
&\mathcal{M}_*(t) = f_2(\mathrm{SFR}, \mathcal{M}_*, \mathcal{M}_{\mathrm{mol}}, \mathcal{M}_{\mathrm{HI}}, \mathcal{M}_{\mathrm{dust}}, \mathcal{M}_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots),\\
&\mathcal{M}_{\mathrm{mol}}(t) = f_3(\mathrm{SFR}, \mathcal{M}_*, \mathcal{M}_{\mathrm{mol}}, \mathcal{M}_{\mathrm{HI}}, \mathcal{M}_{\mathrm{dust}}, \mathcal{M}_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots),\\
&\mathcal{M}_{\mathrm{HI}}(t) = f_4(\mathrm{SFR}, \mathcal{M}_*, \mathcal{M}_{\mathrm{mol}}, \mathcal{M}_{\mathrm{HI}}, \mathcal{M}_{\mathrm{dust}}, \mathcal{M}_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots),\\
&\mathcal{M}_{\mathrm{dust}}(t) = f_5(\mathrm{SFR}, \mathcal{M}_*, \mathcal{M}_{\mathrm{mol}}, \mathcal{M}_{\mathrm{HI}}, \mathcal{M}_{\mathrm{dust}}, \mathcal{M}_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots),\\
&\mathcal{M}_{\mathrm{halo}}(t) = f_6(\mathrm{SFR}, \mathcal{M}_*, \mathcal{M}_{\mathrm{mol}}, \mathcal{M}_{\mathrm{HI}}, \mathcal{M}_{\mathrm{dust}}, \mathcal{M}_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots),\\
&\delta_{\mathrm{gal}}(t) = f_7(\mathrm{SFR}, \mathcal{M}_*, \mathcal{M}_{\mathrm{mol}}, \mathcal{M}_{\mathrm{HI}}, \mathcal{M}_{\mathrm{dust}}, \mathcal{M}_{\mathrm{halo}}, \delta_{\mathrm{gal}}, \ldots),\\
&\vdots
\end{aligned}
\tag{1.1}
$$

ここで $\mathrm{SFR}(t)$, $\mathcal{M}_*(t)$, $\mathcal{M}_{\mathrm{mol}}(t)$, $\mathcal{M}_{\mathrm{HI}}(t)$, $\mathcal{M}_{\mathrm{dust}}(t)$, $\mathcal{M}_{\mathrm{halo}}(t)$, $\delta(t)$ はそれぞれ時刻 t での星形成率，星質量，分子ガス質量，水素原子ガス質量，ダスト質量，暗黒物質ハロー質量，周囲の銀河密度超過をそれぞれ表す．右辺に含まれる変数は象徴的に書かれたもので，それぞれの変数の過去の歴史全てに依存することを表している．

銀河の進化を定式化するには，このような巨大な方程式系を決定する必要がある．天体物理学者はこれまで第一原理の物理法則から支配方程式を構築してきたが，量空間が 10 次元を超えると，そのような方法はもはや現実的ではなくなる．1970 年代から 1980 年代半ばにかけて，主成分分析（PCA）などの古典的な多変量解析手法が，高次元空間で銀河の物理量を結合するために使用された．これにより，さまざまな（対数）線形関係，いわゆる銀河スケーリング関係が発見されている．スケーリング関係を統一して基本的な関係を見つけるための研究により，銀河多様体（Brosche, 1973; Djorgovski, 1992）の概念が生まれた．しかし，古典的 PCA が扱えるのは線型関係のみであり，銀河の（対数）線型関係を探索的に検証するには今でも有用であるものの，銀河多様体は極めて限定された概念に留まり，一旦ほとんど忘れられた（Hunt et al., 2012; Zhang and Zaritsky, 2016; Ginolfi et al., 2020）．時は流れ，21 世紀の銀河調査では数億個の銀河について数百の物理量が得られ，まさに質・量ともに典型的なビッグデータとなっている．解析対象となる銀河の特徴空間は 100 次元を超える．したがって，銀河進化の特徴づけは，物理的直観に頼った従来の方法では不可能で，根本的に異なった新たな発想による方法が必要である．

そこで我々は，これに代わる現代的手法による銀河進化の議論に着手した（Siudek et al., 2018）．具体的には，紫外から近赤外までの 12 の波長（波長 $\lambda = 150$ nm–2.2 μm）と各宇宙年齢ごとの光度を含む 13 次元特徴空間を構築し，教師なし機械学習の方法であるフィッシャー EM

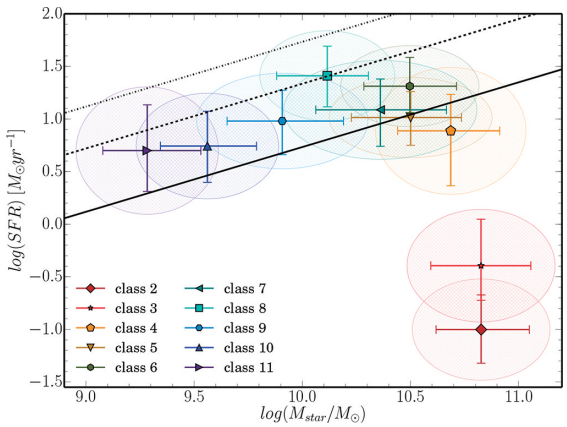

図 1. フィッシャー EM アルゴリズム(FEM)によって求められた星質量 ($\mathcal{M}_*$)–SFR 関係（Siudek et al., 2018, Fig. 7 を再掲）．FEM が分類した銀河のクラスの $\log \mathrm{SFR}$, $\log \mathcal{M}_*$ の中央値，および分散が示されている．楕円の面積は絶対偏差の中央値に対応する．黒い実線は $z = 0.7$ での星形成銀河主系列を表す（Whitaker et al., 2012）．破線と一点鎖線はそれぞれ活動的な星形成銀河とスターバースト銀河の星形成主系列に対応している（Rodighiero et al., 2011）．

アルゴリズム（FEM: Bouveyron and Brunet, 2012）を適用した．その結果，FEM は銀河の総星質量 $\mathcal{M}_*$ と星形成率 SFR の間に見られる星形成銀河主系列と呼ばれる関係を恣意的なサンプル選択なしで発見することに成功した(図 1)．また，FEM は星形成銀河主系列はある総星質量を超えると星形成を停止した銀河の系列に連続的につながっていくことも発見した．これは，銀河の星形成が急激に停止し，不連続に星形成を停止した銀河に遷移するという仮説に反する構造で，多波長光度空間の情報を全て活用することで初めて発見することができたものである．この銀河の連続体こそが，銀河進化の基本を表す銀河多様体の一つの射影になっている．多波長光度空間の銀河多様体はその非線型空間構造のため，古典的 PCA に基づく以前の研究では発見され得なかった．

しかし，宇宙物理学の研究は銀河多様体の定量的記述のみでは満足せず，その構造を完全に理解し，銀河進化の物理学を支配する(おそらくいくつかの)パラメータへの依存性を解明することが課せられる．この更なる目標のためには，より洗練された方法が必要である．

1.2 多波長光度空間における銀河多様体

単位時間あたりにどのくらいの質量の星が形成されるかを星形成率とよび，[太陽質量 年$^{-1}$] ([$\mathcal{M}_\odot$ yr^{-1}] と表記)で測る．星形成率の時間発展を星形成史といい，これが銀河の進化を決める重要な要素の 1 つである．紫外線から近赤外線の波長では，銀河の放射スペクトルは星とガスの寄与が支配的である．星の温度および寿命は星の質量に強く依存しており，大質量の星ほど明るく高温で，寿命は短い．高温の星は紫外線を大量に放射するが，低温の星は紫外線では暗く，近赤外線で光る．定量的には，星が安定して定常的に核融合をする段階である主系列星でいられる時間を τ_{MS}，星の表面温度を T，星の光度を L とおくと

$$\tau_{\mathrm{MS}} \propto \mathcal{M}^{-2.5}, \tag{1.2}$$

$$L \propto \mathcal{M}^{3.5} \tag{1.3}$$

$$L \propto T^4 \tag{1.4}$$

と近似できる．この帰結として，高温の星から先に寿命が尽きるため，星形成史は銀河のスペクトルに直接反映される．つまり，星形成史は銀河の多波長(バンド)光度が張る空間において特徴が適切に表れていると期待される．

従来の天文学では，さまざまな波長での光度の比によって，多波長光度空間における進化を特徴づける方法が使われてきた．天文学ではこの比を色(color)と呼ぶ．異なる 2 つの波長 λ_1, λ_2 $(\lambda_1 < \lambda_2)$ における天体の単色光度のペア L_{λ_1} と L_{λ_2} を考える．もし $L_{\lambda_1} < L_{\lambda_2}$ であればその天体は「赤い」，$L_{\lambda_1} > L_{\lambda_2}$ の場合は「青い」と表現する．銀河の光度(絶対等級[1])と色の関係をプロットすると(色-等級図)，明らかな 2 つの系統が現れる．これは銀河の色の二峰性(bimodality)と呼ばれている．具体的には，赤い銀河のタイトな系列(レッドシークェンス: red sequence)と，より広がった青い銀河の系列(ブルークラウド: blue cloud)が普遍的に存在している．レッドシークェンスとブルークラウドの間の領域に存在す銀河は相対的に少ないため，緑の谷(green valley)と呼ばれることもある（たとえば Blanton, 2006）．ブルークラウドの銀河は星形成が活発で，短寿命かつ高温度の大質量星が存在しているが，レッドシークェンスは星形成が停止し，小質量で低温の小質量星が卓越する．銀河の進化はブルークラウドからレッドシークェンスに移行すると考えられているが，この遷移がどのように起きるかは長らく未解決問題として残っていた．最近の先行研究で，ブルークラウドとレッドシークェンスの間は不連続ではなく，色-色-絶対等級の 3 次元空間において連続的に接続する構造が存在することが示唆された（たとえば Chilingarian and Zolotukhin, 2012）．

しかし，従来の色に基づく銀河進化の評価方法には潜在的な問題がいくつか存在する．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)よりも明るい天体しか含まれないという偏りである．振動数 ν での天体の等級(magnitude)を m_ν とすると，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．前述したように，色とは 2 波長の光度比であるため，観測における選択効果は入り組んだ形で現れ，単純なコンプリートネス[2]の検証はほぼ不可能である．色-等級図上での銀河進化の研究は，この複雑な選択効果が物理的な特徴と分離できず，混乱した議論が続いていた．しかし，色は光度の比であることから，多波長の光度(絶対等級)が張る多次元空間に戻った議論も可能である．この方法ならば，選択効果は直接的な形で評価できる利点がある．たとえば色-等級図における二峰性は，元の多次元光度空間にも対応するピーク構造が存在するはずである．よって，我々は多波長光度の高次元空間内で銀河が形作る構造に注目する．

我々が Siudek et al. (2018)で発見した銀河多様体は非線型な構造を持っている．さらに驚くべきことに，この銀河多様体を構成するサンプル銀河のスペクトルは，いくつかの広帯域バンド光度[3]の情報のみで区別され，より複雑な物理量の組み合わせは必要なかった．この事実は，紫外線から可視光，近赤外線での銀河の多波長光度は，せいぜい数個の物理量で説明できることを示唆している．これは，従来の方法では決して見つけることができなかった銀河進化の新しい特徴付けである．この発見をきっかけとして，我々は銀河多様体をさらに追求し，銀河多様体から銀河進化の物理を支配するパラメータ(おそらく多くてもいくつか)の依存性を解明し，銀河進化の支配方程式を導出する研究に着手した．このため我々は，従来の天文学的方法論とはまったく異なる，データサイエンスの最新手法の 1 つである多様体学習（たとえば Ma and Fu, 2012)として知られる手法に注目し，更なる解析を進めている．本論文ではこの一連の研究の現状報告と将来展望について紹介する．結果の一部は Cooray et al. (2023)にて公表済みであるが，本稿でより一般的な解説を試みる．

本論文では，観測データに関するすべての計算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 $\Omega_{K0} = 0$ とした．これらのパラメータの意味は付録 A で説明している．

2. データ

本研究で用いたデータは Reference Catalog of Galaxy Spectral Energy Distribution（RCSED: Chilingarian et al., 2017）である．RCSED は，紫外線天文衛星 GALEX の全天探査カタログ，可視光大規模分光測光探査プロジェクト SDSS のカタログ，および近赤外線広域探査 UKIDSS カタログを統合し，最新の天文学スペクトル分析方法を用いて構築されている．RCSED は全天の約 25% をカバーし，数百万個の銀河の FUV, NUV, $u, g, r, i, z, Y, J, H, K$ の 11 バンドで k 補正[4]された測光データ，および関連する物理量の情報が記載されている．さらに，いくつかの公開データを再処理した関連情報が測光カタログに追加されている．基本となる天体リストは SDSS データリリース 7（DR7）（Abazajian et al., 2009）のうち，赤方偏移 $0.007 < z < 0.6$ の範囲にある活動銀河ではない銀河の分光サンプル[5]である．このデータには 800,299 個の銀河が含まれている．

全サンプルのうち，11 バンドすべてで測光値を持つ銀河を抽出すると，90,565 個の銀河が得られる．そして赤方偏移[6]の信頼度が ≤ 0.5 の銀河を取り除くと，銀河数は 90,460 個となる．親サンプルと比べて大きく銀河数が減ったのは，主に UKIDSS サンプルとのクロスマッチが可能な天域が狭いことによる．

本研究の主目的は銀河光度空間における普遍的な関係の発見と定量化である．等級選択効果を回避するため，上記のサンプルを SDSS g バンドの等級をもとにコンプリートなサンプルを構築した[7]．g バンドでの限界等級 $m_{\mathrm{AB},g} = 18.3$[8]から算出した限界絶対等級曲線を用い，最終サンプルの銀河数が最大となるように絶対等級を求めた．これにより，最終的に赤方偏移 $z_{\mathrm{lim}} < 0.097$, $M_{\mathrm{lim},g} \leq -20.016$ の範囲にある 27,056 個の銀河からなるサンプルが構築された．以降の解析は全てこの volume-limited サンプルに基づく．

多様体学習においては，データの値の範囲を最適化するための前処理が重要となる．本研究では，各バンドでの銀河の光度は絶対等級の単純平均で中心化し，分散を 1 に規格化した解析と，リスケールせず絶対等級の値そのものを用いた解析を行った．これら 2 つの解析の結果は（各軸を再び絶対等級にスケールしなおせば）定量的にもほとんど差がなかった．よって本稿では，サンプルの絶対等級はリスケールせずにそのままデータの特徴空間とした結果のみ示す．これは，今回のサンプルでは，どのバンドでも放射には銀河内の星からの寄与が支配的であることから，絶対等級は各バンドで互いに似た範囲に収まったことが理由である．しかし，赤方偏移を加えるなど今後の他の物理量も含めた解析のためには，適切な規格化が必要となることに注意しておく．

3. 方法：多様体学習による銀河多様体の定量化

3.1 多波長光度空間での銀河多様体

2 章で構築したサンプルの基本的検証のため，まず Siudek et al.（2018）と同様に FEM をデータに適用し，このデータからも 11 次元の多波長光度空間内で銀河が形作る低次元の構造が得られることを確認した．この銀河多様体は FEM によって抽出されたクラスタの空間配置を解析することにより，多波長光度の特徴空間において 2 次元曲面をなすことが明らかになった．RCSED の銀河多様体を図 2 に示す．

我々の銀河多様体は 11 次元光度空間内での低次元部分空間で表現されているが湾曲した構

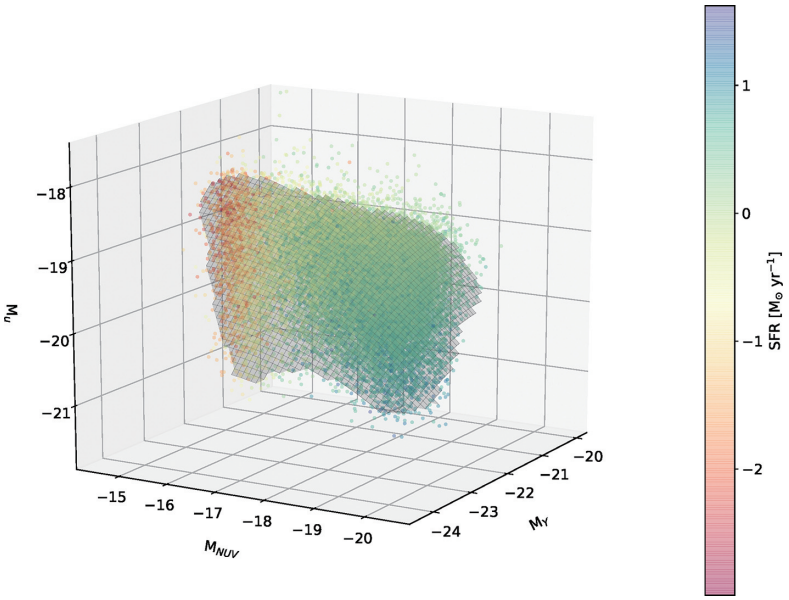

図 2. 多次元特徴空間で発見された銀河多様体．元の特徴空間は 11 次元であるが，多様体は 2 次元構造のみを持ち，基本的に紫外，可視光，および近赤外の光度の 3 次元空間に埋め込まれている．多様体は湾曲した形状をしているため，古典的 PCA のような線型関係を分析する手法では決して見つけることができない．カラーコーディングはサンプル銀河の星形成率を表している．

造を持っており，視覚的に構造を把握するのも難しく，またそれ以上の定量化も容易ではない．古典的なスケーリング法則における多峰性と分散は，銀河多様体の非線型構造を反映しない，最適ではない射影の結果であることも多い．1 章でも述べたように，銀河進化に関する観測データは巨大化する一方である．このため，これからの研究においては情報量を最大限に活かす簡潔な記述方法が必要である．

我々は，銀河多様体をより有効に利用して定量化するため，次元削減と呼ばれる一群の方法に着目した．具体的には，銀河多様体から銀河進化の物理学を支配する（おそらくは高々数個の）物理量への依存性を解明するため多様体学習を適用した．我々はこの方法を用い，古典的な天文学における第一原理的理論構築とは全く異なる，銀河進化の新しい定量化を試みている．銀河多様体を定量的に表現するもう一つの大きな利点は，例えば観測した天体に欠けている光度などの観測量や，SFR や星の質量 $\mathcal{M}_*$ などの物理量を多様体上の位置から直接推定することが可能になることである．これは銀河多様体を光度空間に引き戻すことで実現できる．この引き戻しのためには，多様体上の距離をもとの多波長光度空間の計量でも記述しておけば便利である（Lin et al., 2017）．天文観測は一般に観測限界ぎりぎりを用いる困難なデータ解析が多く，観測量の予測や推定，補間に用いることのできる銀河多様体の定量表現は今後の天文学研究における強力な武器となる．

3.2 多様体学習

多様体学習では，データは測地線距離 $d^{\mathcal{M}}$ で決まる計量を持つ，滑らかな d 次元多様体 $\mathcal{M}$ からランダムに抽出された有限個の点集合 $\{y_i\}$ $(i = 1, \ldots, N)$ と考える．これらのデータ点は，滑

らかな写像 ψ によってユークリッド計量 $\|\cdot\|_{\mathcal{X}}$ の特徴空間（あるいは入力空間）$\mathcal{X}=\mathbb{R}^n$ $(d \ll n)$ に埋め込まれている．埋め込まれた特徴空間内でのデータ点を $\{x_i\}$ $(i=1,\ldots,N)$ とすると，埋め込み写像は $\psi:\mathcal{M}\longrightarrow\mathcal{X}$ であり，多様体上の点 $y_i\in\mathcal{M}$ は

$$y_i=\psi^{-1}(x_i),\ x_i\in\mathcal{X} \tag{3.1}$$

と表される．多様体学習の目的は，特徴空間の（入力）データ点集合 $\{x_i\}\in\mathcal{X}$ が与えられたときに，多様体 $\mathcal{M}$ および写像 ψ の具体的な形を求め，元のデータ点集合 $\{y_i\}\in\mathcal{M}$ を再構成することである．多様体学習アルゴリズムを入力データ点集合に適用すると，$\mathbb{R}^n$ のデータ集合を隣接点との関係を維持しながら低次元空間 $\mathbb{R}^d (d\ll n)$ に写像する．すなわち

$$x_i\mapsto\hat{y}_i=\left(\psi_1^{-1}(x_i),\ldots,\psi_d^{-1}(x_i)\right)^{\top}\in\mathbb{R}^{\hat{d}} \tag{3.2}$$

を通じて本来の $\{y_i\}\subset\mathbb{R}^d$ の推定値としての $\{\hat{y}_i\}\subset\mathbb{R}^{\hat{d}}$ が得られる．ここで，$\top$ はベクトルの転置を表す．これは，データ集合は高次元の特徴空間内で低次元の部分多様体上に分布するという「多様体仮説」という仮定（たとえば Goodfellow et al., 2016）に基づいてデータの次元を縮小する方法で，非線型次元削減と呼ばれる方法論の一つである．多様体の次元 d も $\hat{d}$ を通じてデータから推定できれば理想的である．しかし実際の解析では我々が $\hat{d}$ を設定し，いくつかの基準から最適な $\hat{h}$ を選ぶという手順を取る（第 4.1 章）．

多様体学習に関する研究の萌芽は散発的に発表されてきた 90 年代の研究まで遡るが，2 編の独創的な論文（Roweis and Saul, 2000; Tenenbaum et al., 2000）が出版されて以降人気が高まり，積極的に研究されるようになった．多様体学習アルゴリズムは，特徴空間内で複雑な形状を持つ多様体を「展開」し，その上に局所座標系を提供できる（Tenenbaum et al., 2000; Roweis and Saul, 2000）．重要なのは，次元削減後のデータ点どうしの繋がりが元の高次元空間でのデータ点どうしの繋がりをなるべく忠実に表現することで，そのためにはアルゴリズムがデータの形を学習する必要がある．これが多様体学習という言葉の由来である．

古典的 PCA などの線型な方法がデータ構造の大域的構造を表すのに有効なのに対し，非線型な方法は局所的な構造の表現に効力を発揮する．一方，非線型手法の多くはデータ点の近傍の位置関係の低次元表現に注目して次元削減を行うため，大域的構造が失われることがある．よって，多様体学習を用いる場合，目的によって適切なアルゴリズムを選択する必要がある．また，多様体学習が与える低次元表現の座標系は，直感的あるいは物理的意味を持つことが必ずしも保証されないことに注意が必要である（たとえば Liu et al., 2017）．この点については 4 章で議論する．

3.3 `Isomap` および `UMAP` アルゴリズム

本研究では，多様体学習のアルゴリズムとして `Isomap`（isometric feature mapping: Tenenbaum et al., 2000）および `UMAP`（uniform manifold approximation and projection: McInnes et al., 2018, 2020）を採用する．本研究の目的は銀河進化の物理量依存性の定量化であり，元の高次元特徴空間で連結な構造は多様体としても連結した構造に写像されることが必要である．2 つのアルゴリズムはともに，元のデータ点分布の持つ連結性を保つ性質を持っており，理想的な方法である．これらの計算には `scikit-learn`（Pedregosa et al., 2011）を使用した．

3.3.1 `Isomap`

`Isomap` アルゴリズムは滑らかな多様体 $\mathcal{M}$ が $\mathbb{R}^d$ $(d\ll n)$ の測地的凸（geodesically convex）領域であり，埋込み写像 $\psi:\mathcal{M}\longrightarrow\mathcal{X}$ が等長写像（isometry）であると仮定する．まず測地的凸性を以下で定義する（Tenenbaum et al., 2000）．

定義 1. 測地的凸(**geodesically convex**) $(\mathcal{M}, g)$ をリーマン多様体とする．$\mathcal{M}$ の部分集合 $\mathcal{U}$ は，$\mathcal{U}$ 内の任意の 2 点を結ぶ $\mathcal{U}$ 内の最短測地線が一意に存在するとき，測地的凸集合であるという．

測地的凸なリーマン多様体は，測地線距離に関して凸な計量空間でもある．

これより，Isomap の仮定は以下のように表現される．

凸性 $\mathcal{M}$ は $\mathbb{R}^d$ の測地的凸な部分集合である．

等長性 測地線距離は写像 ψ の下で保存される．多様体 $\mathcal{M}$ 上のあらゆる 2 点 $y, y' \in \mathcal{M}$ について，これらの間の測地線距離が対応する埋め込まれた $\mathbb{R}^n$ の 2 点 $x = \psi(y), x' = \psi(y') \in \mathcal{X}$ のユークリッド距離に等しくなる．即ち

$$d^{\mathcal{M}}(y, y') = \|x - x'\|_{\mathcal{X}} \,. \tag{3.3}$$

Isomap は $\mathcal{M}$ が測地的凸の領域，ψ が等長変換であるという仮定を用いて多次元尺度構成法(multidimensional scaling: MDS)を一般化したアルゴリズムである．MDS はデータの 2 点間のユークリッド距離を保存しつつ，データ点が分布するより低次元の部分空間を探す方法である．MDS は線型次元削減法の 1 つであり，曲がった領域ではうまく機能しない．Isomap は MDS の思想を踏襲し，データのすべてのペア間の $\mathcal{M}$ 上での測地線距離を近似することで，非線型多様体の大域的幾何構造を最大限に保存するアルゴリズムである．この意味で，Isomap は多様体学習としては局所的方法でもあり，同時に大域的方法でもある．

Isomap アルゴリズムは 5 つのステップから構成される．

(1)最近傍探索

ある整数 K あるいは $\epsilon > 0$ を取る．特徴空間 $\mathcal{X}$ 内全てのデータ点ペア $x_i, x_j \in \mathcal{X}$, $(i, j = 1, \ldots, n)$ 間の距離

$$d_{ij}^{\mathcal{X}} \equiv d^{\mathcal{X}}(x_i, x_j) = \|x - x'\|_{\mathcal{X}} \tag{3.4}$$

を計算する．距離としては一般にユークリッド距離を取る．K 番目に近い点(K-nearest neighbor)まで，あるいは半径 ϵ の球内にあるすべての点を結合することにより，$\mathcal{M}$ 上の隣接点を決める．Isomap の性能は K ないし ϵ の選択によって定まる．

Isomap は，効率的な近傍検索のため sklearn.neighbors.BallTree を使用する．

(2)グラフ距離の算出

入力データ点群 $\{x_i\}$ $(i = 1, \ldots, N)$ に対し，重み付き近傍グラフ(weighted neighborhood graph)$\mathcal{G} = \mathcal{G}(\mathcal{V}, \mathcal{E})$ を構成する．グラフの頂点集合(vertices)$\mathcal{V}$ はデータ点群 $\{x_1, \ldots, x_N\}$ で，辺集合(edges)$\mathcal{E}$ はデータ間の近傍関係を示す辺 e_{ij} である．辺 e_{ij} には 2 点間の距離 $d_{ij}^{\mathcal{X}}$ に対応する重み w_{ij} が与えられている．2 点 x_i, x_j が直接辺で接続していなければ重みは ∞ となる．

$\mathcal{M}$ 上の 2 点間の測地線距離 $\{d_{ij}^{\mathcal{X}}\}$ を，グラフ $\mathcal{G}$ によるグラフ距離(graph distance)$d_{ij}^{\mathcal{G}}$ によって推定する．グラフ距離 $d_{ij}^{\mathcal{G}}$ とは，グラフ $\mathcal{G}$ 上の 2 点間の最短の道(path)の長さで定義する．近傍にない 2 点は最近傍をつないだ最短距離の道で接続し，道の長さは重みの和で与える．この長さが離れた 2 点の測地線距離の近似を与える．

データ点が多様体 $\mathcal{M}$ 上で定義される確率分布から抽出されたとすると，多様体が平坦ならば $N \longrightarrow \infty$ でグラフ距離 $d^{\mathcal{G}}$ は測地線距離 $d^{\mathcal{M}}$ に収束する(Bernstein et al., 2001)．このための最も効率的なアルゴリズムとして Floyd–Warshall アルゴリズム(Floyd, 1962; Warshall, 1962)や Dijkstra のアルゴリズム(Dijkstra, 1959)が用いられる．前者はグラフが密な場合，後者は疎な場合に有効であることが知られている(たとえば Ma and Fu, 2012)．

(3) **MDS によるスペクトル埋め込み(spectral embedding)**

距離行列 $\mathsf{D}^{\mathcal{G}} \equiv (d_{ij}^{\mathcal{G}})$ を考える($N \times N$ 対称行列)．行列 $\mathsf{D}^{\mathcal{G}}$ に古典的 MDS を適用し，多様体 $\mathcal{M}$ 上のデータ点間の測地線距離ができるだけ保存されるように d 次元空間 $\mathcal{Y}$ を再構成する．$\mathsf{S}^{\mathcal{G}} \equiv ((d_{ij}^{\mathcal{G}})^2)$ をグラフ距離の 2 乗を成分とする $N \times N$ 対称行列とする．これを以下によって二重中心化する

$$\mathsf{K}_N^{\mathcal{G}} = -\frac{1}{2}\mathsf{H}\mathsf{S}^{\mathcal{G}}\mathsf{H}\,, \tag{3.5}$$

$$\mathsf{H} \equiv \mathbb{I}_N - \frac{1}{N}\mathbf{1}_N\,. \tag{3.6}$$

ここで $\mathbb{I}_N$ は $(N \times N)$ 単位行列，$\mathbf{1}_N$ は $(N \times N)$ の成分がすべて 1 の対称行列である．

(4) 埋め込みベクトル $\{\hat{y}_i\}$ を $\|\mathsf{K}_N^{\mathcal{G}} - \mathsf{K}_N^{\mathcal{Y}}\|$ を最小化するように選ぶ．ここで

$$\mathsf{K}_N^{\mathcal{Y}} = -\frac{1}{2}\mathsf{H}\mathsf{S}^{\mathcal{Y}}\mathsf{H}\,. \tag{3.7}$$

ここで $\mathsf{S}^{\mathcal{Y}} = \big((d_{ij}^{\mathcal{Y}})^2\big)$, $d_{ij}^{\mathcal{Y}} = \|y_i - y_j\|$ は y_i と y_j のユークリッド距離である．$\mathsf{K}_N^{\mathcal{G}}$ を固有値行列 $\mathsf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$，固有ベクトル行列 $\mathsf{V} = (v_1, \ldots, v_N)$ によって固有値分解すると

$$\mathsf{K}_N^{\mathcal{G}} = \mathsf{V}\mathsf{\Lambda}\mathsf{V}^{\top}\,. \tag{3.8}$$

最適解は $\mathsf{K}_N^{\mathcal{G}}$ の大きい方から d 個の固有値 $\lambda_1 \geq \cdots \geq \lambda_d$ に対応する固有ベクトル $v_1, \ldots, v_d$ によって得られる．

(5) グラフ $\mathcal{G}$ は $d \times N$ 行列

$$\mathsf{Y} \equiv (\hat{y}_1, \ldots, \hat{y}_N) = (\lambda_1^{\frac{1}{2}} v_1, \ldots, \lambda_d^{\frac{1}{2}} v_d) \tag{3.9}$$

によって d 次元部分空間 $\mathcal{Y}$ に埋め込まれる．

その構造上，Isomap は 2 点間の計量を保持するため，特徴空間内のデータ点の多様体上での「表面密度」を保存する．すなわち，データ点が密に存在する領域は多様体上でも密になり，疎な領域は多様体上でも疎となる．Isomap は多様体 $\mathcal{M}$ がユークリッド空間の測地的凸な部分多様体であること，および ψ の等長性を仮定するため，曲率が大きすぎる，多様体に穴がある，非凸であるなどの場合にはうまく機能しない．実践上の問題として，ノイズがある，すなわちデータ点が必ずしも多様体上に分布していない場合，Isomap のパフォーマンスは近傍の取り方に依存する．ノイズが大きすぎなければ，Isomap は基本的にノイズの影響に対してある程度頑健である．本研究では，Isomap の近傍は $K = 5$ に取った．

3.3.2 UMAP

UMAP(uniform manifold approximation and projection)は 2018 年に提案された比較的新しい手法で，微分幾何学と代数トポロジーに基づいている．UMAP では，元の特徴空間で近いデータ点は多様体上でも近くなる．実行時間が速いことから計算コストが低減でき，4 次元以上の多様体への次元削減も可能である．UMAP は位相的データ解析とリーマン幾何学が元になっているアルゴリズムである．このアルゴリズムは以下の 3 つの仮定に基づく：1)データはリーマン多様体上に一様に分布している，2)リーマン計量は局所的に一定である(そのように近似できる)，3)多様体は局所連結性を持つ．これらの仮定から，ファジーな位相構造をもつ多様体をモデル化することが可能である．データ点が可能な限り均一に分布するように多様体を定義するため，Isomap とは違いデータ点の表面密度は保存されないことに注意する．UMAP アルゴリズムは次の 3 つの段階で構成される:

(1) リーマン多様体の推定,
(2) 距離空間のファジートポロジーによる表現,
(3) 次元削減.

UMAP の核となる概念はファジートポロジー表現であるが，圏論を用いて説明されていることもあって簡潔な記述は難しいため，ここでは概略を示すにとどめる．詳細は関連文献を参照されたい． UMAP はその構成法から，Isomap よりもノイズに対して頑健である．本研究では，UMAP の近傍は $K = 50$ に取った.

4. 結果と議論

4.1 結果: Isomap および UMAP 銀河多様体

Isomap および UMAP によって得られた銀河多様体を図 3 に示す．異なったアルゴリズムである Isomap と UMAP が定性的にかなり似通った 2 次元多様体を与えることは注目すべき点である (図 3)．2 つの方法で推定された銀河多様体の違いは図 3 からはっきり見て取れる．

Isomap はデータ点群の密度を保存するため，多様体には密度構造，つまり多様体上の密な領域と疎な領域がある．対照的に，UMAP は多様体の構築において密度を可能な限り均一にするため，UMAP 多様体はかなり一様な密度分布となっている．つまり，Isomap 多様体で密度が高い領域はそのまま現れるが，UMAP 多様体では面積が拡大する．

Isomap および UMAP によって抽出された多様体の次元を推定するため，再構成誤差および情報量基準をもとに評価した[9]．Isomap，UMAP とも再構成誤差は次元数を 2 から順に増加させても数値誤差の範囲で変化がみれらない．また Akaike information criterion (Akaike, 1974)，Bayesian information criterion (Schwarz, 1978) からはともに次元数は 2 が選ばれた．多波長光度空間での結果と合わせ，このデータから得られた銀河多様体の次元は 2 であると結論される．

銀河多様体に銀河進化に関する情報がどう表れるかを検証するため，図 4 で SFR と星の質量を多様体上の関数として比較する．2 次元銀河多様体の座標軸 1 および 2 と星形成率，星質量の各相関を Isomap および UMAP それぞれについて示したのが図 5 および図 6 である．ここで，Isomap 多様体の軸は第 1，第 2 固有値に対応する固有ベクトルであった．UMAP でも 2 次元

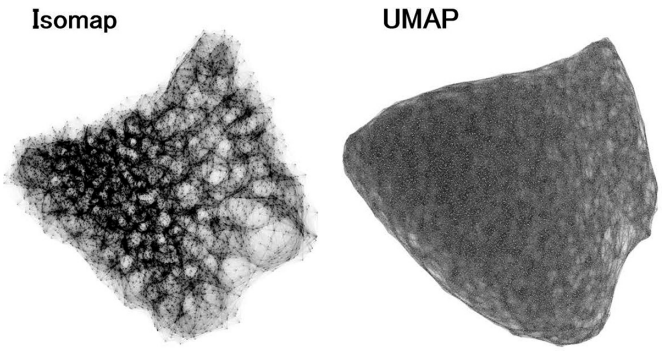

図 3. 多様体学習アルゴリズム Isomap および UMAP による「展開された」銀河多様体．左側と右側のパネルは，それぞれ Isomap と UMAP からの多様体を示している．この空間上の多様体の構造は，図 2 よりもはるかに簡単に認識できる．全体的な形状は互いにわずかに異なるが，続く解析で示すように星形成率や星質量の多様体上の分布などの特徴は共通している．

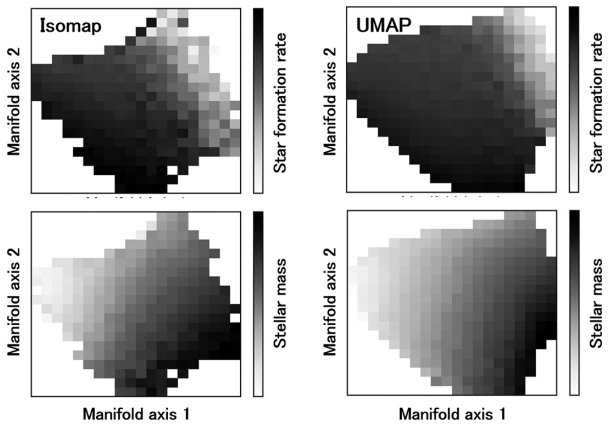

図 4. 2 つの異なる多様体学習アルゴリズム，Isomap と UMAP によって取得された銀河多様体．星形成率 SFR と星質量 M_* は多様体上の関数として表される．左パネルは，それぞれ Isomap によって得られた多様体上の星形成率と星質量を示している．右パネルは，UMAP によって得られた多様体で，カラーコーディングは左パネルと同じである．

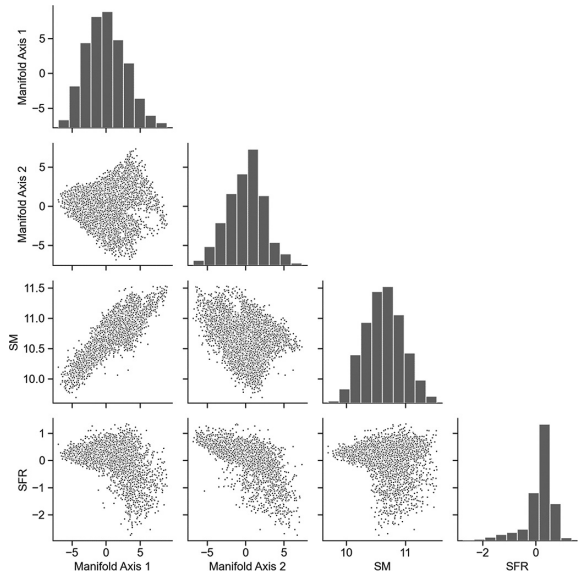

図 5. Isomap による 2 次元銀河多様体の座標軸と星形成率(SFR)，星質量(SM)の散布図．多様体座標 1 は星質量と，座標 2 は星形成率と強く相関している．

的な多様体が得られるが，その意味は明確には与えられない．Isomap 多様体との対応は，ここで物理量の多様体上での分布を評価することで明らかになる．以下，座標軸はこれによって揃えた形で議論する．星形成率 SFR と星質量 $\mathcal{M}_*$ の挙動は 2 つの図で定性的に非常に似ており，推定された多様体構造がロバストであることを示唆している．これは，多様体学習が実際に多波長光度空間における銀河進化の重要な特徴を「学習」したことを意味する．図 5，図 6 から，多様体座標 1 は星質量と，座標 2 は星形成率と強く相関していることが分かる．我々はすでに可視光光度空間における銀河多様体が基本的に 2 次元であることを見た．これはすなわち，紫

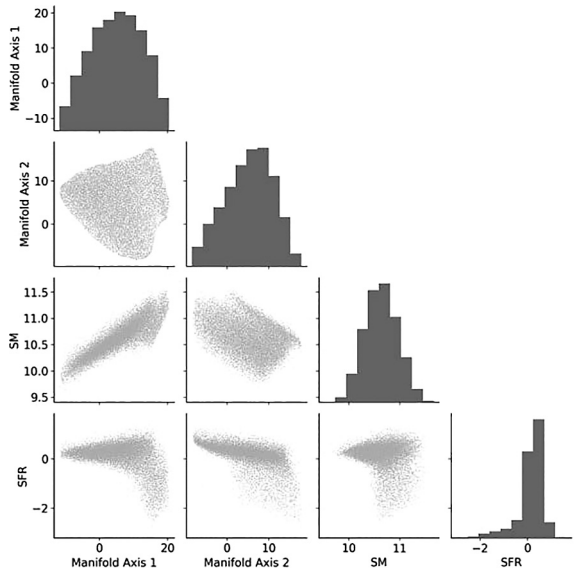

図 6. UMAP による 2 次元銀河多様体の座標軸と星形成率(SFR)，星質量(SM)の散布図．図 5 と同様，多様体座標 1 は星質量と，座標 2 は星形成率と強く相関している．

外線・可視光線・近赤外線の波長における銀河進化は星形成率，星質量のわずか 2 つの物理量で過不足なくあらわされることを意味しており，銀河進化理論に強い制限を与える重要な発見である（Cooray et al., 2023）．

このように，多様体学習は銀河多様体と星形成率や星質量などの物理量を結び付けることができる．これをさらに押し進めることで，多様体上の銀河進化をパラメータ化することも原理的には可能である．星が形成されると，星質量，つまり蓄積された星の総質量が増加する．これは銀河進化の基本的な側面の 1 つであり，この進化を多様体上のベクトル場として可視化できる．星形成のベクトル場を図 7 および図 8 に示す．銀河進化の「速度場」はこれら 2 つの図から明らかに見て取れる．小質量の銀河は急速に進化して星形成率が減少し，星質量が大きくなる(図上左から下への向き)．大質量の銀河は進化が遅く，多様体上の同じ位置(右上)に長く滞在する．

4.2 銀河多様体と観測量

銀河多様体から入力空間である多波長光度空間における情報に引き戻して理解するため，多様体の軸と観測された光度の間のペアプロットを図 9 に示す．多様体座標 1 は，銀河体の骨格を構成する古い恒星集団の寄与が大きい可視光長波長域から近赤外線，r, i, Y, I, J, H, K バンドの光度と密接に相関している．対照的に，多様体座標 2 は紫外線 FUV, NUV から可視光短波長側 u, g バンドと密接に相関しており，現在あるいは非常に最近の星形成活動を表している．可視光での光度に注目すると，それらが互いに非常にタイトに相関している．これは，可視光の波長範囲で複数の光度を解析に含めても，銀河の特性に関する情報は本質的に追加されないことを意味する．対照的に，紫外線光度は散布図で自明ではない非線型相関を示しており，紫外線と可視光バンドの組み合わせが多様体の構造に関する基本的な情報を提供することを示している．近赤外線バンドの光度は可視光線光度と似た振舞いをするものの，この相関関係は多様体の構造に関する追加情報がまだ存在することを示唆している．したがって，11 個

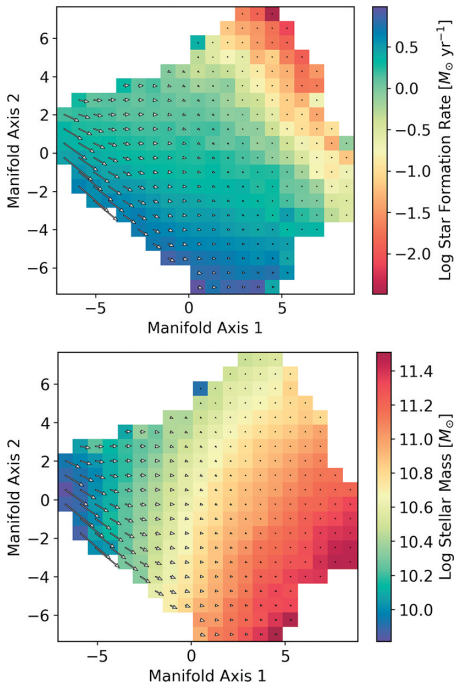

図 7. Isomap 銀河多様体上の星形成率・星質量進化のベクトル場．上パネルのカラーバーは現在の銀河の星形成率，下パネルは星質量を表している．

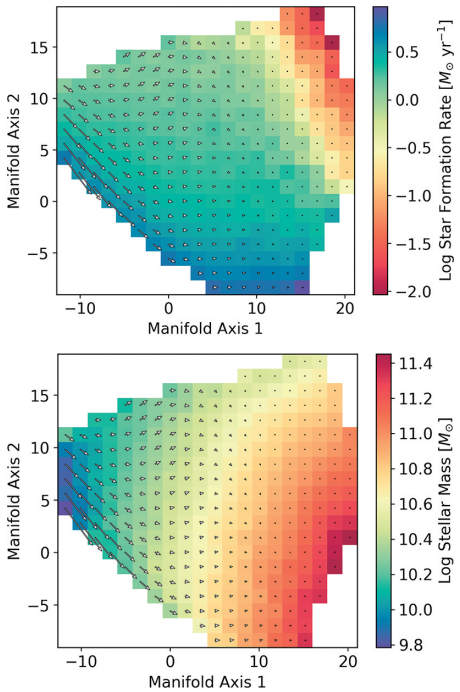

図 8. UMAP 銀河多様体上の星形成率・星質量進化のベクトル場．

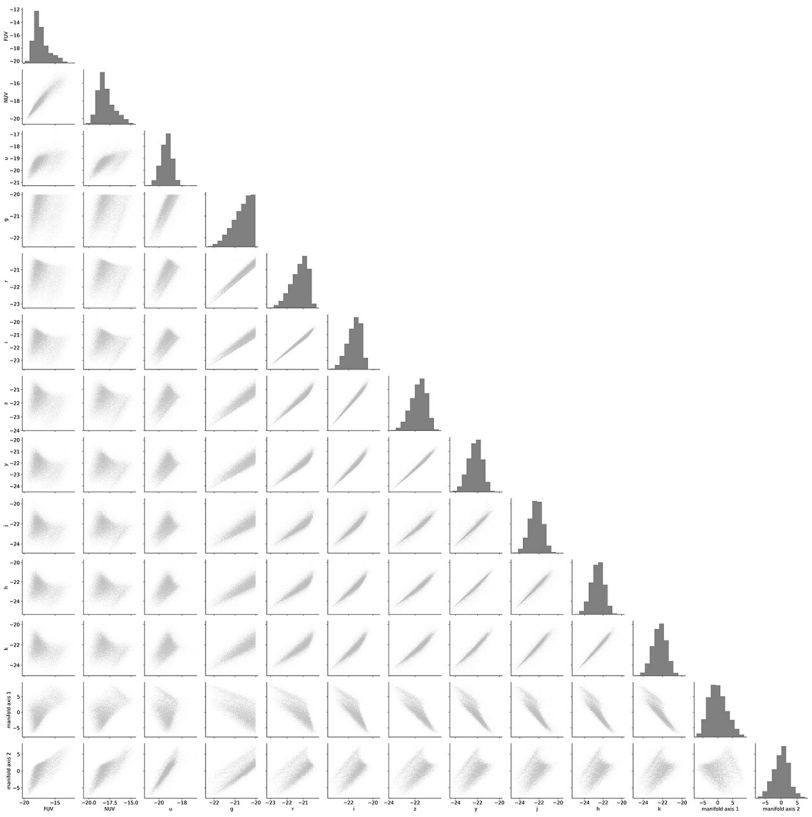

図 9. 各バンドでの銀河の光度と銀河多様体座標 1 および 2 との相関.

の多波長光度にまたがる銀河の特徴空間は，必然的により低次元の部分多様体によって表される．これが，発見された 2 次元銀河多様体の背後にある天体物理的基礎である．

4.3 定量化から定式化へ

残された課題は，多様体上の銀河進化の軌跡をどのように記述し，解釈するかである．無論これは簡単ではなく，さらなる議論が必要である．Cooray et al. (2023)において，我々は銀河の化学進化の古典的な理論モデルを適用した．化学進化は，恒星の進化理論に基づいて銀河内の元素の形成と進化を扱う銀河物理学の分野である[10]．重要な物理プロセスは，星の中心部での核融合によって生じる元素合成である．Lilly et al. (2013) によって提案された物質の流出を伴う単純なモデルを採用する．

$$\mathcal{M}_*(t_{n+1}) = \mathcal{M}_*(t_n) + (1-r)\mathrm{SFR}(t_n)\Delta t\,, \tag{4.1}$$

$$\mathcal{M}_{\mathrm{ISM}}(t_{n+1}) = \mathcal{M}_{\mathrm{ISM}}(t_n) - (1-r+\zeta)\mathrm{SFR}(t_n)\Delta t \tag{4.2}$$

ここで r はリターンドマスフラクション(returned mass fraction: ガスが星間物質に戻る割合)，ζ はマスローディングファクター(mass loading factor: 星形成率に対する質量流出の比)，Δt は時刻のタイムステップを表す (Cooray et al., 2023)．図 10 に示すように，式(4.1)と(4.2)から銀河の理論的な進化軌跡を計算できる．銀河多様体上のベクトル場の解釈は，式(4.1)および

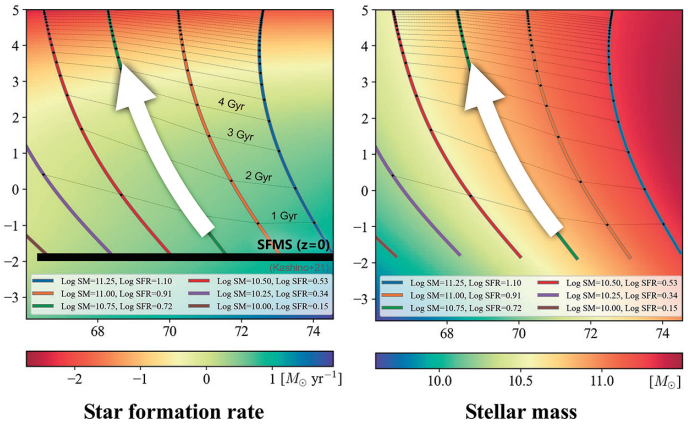

図 10. 銀河の古典的な化学進化モデルによって予測される理論的な進化の軌跡．太実線は異なった星質量を初期値としたときの銀河進化の軌跡を示す．細線は銀河年齢．SFMS は星形成銀河主系列（連続的に星形成をしている銀河の経験則）を表している．

(4.2)を図 7 あるいは図 8 と比較することで得られる．図 9 に関する第 4.2 章での議論でみたように多波長光度は互いに強く相関しており，大きく星形成率 SFR と星質量 $\mathcal{M}_*$ に関連するものに分けられる．とはいえ式(1.1)でみたように，理論のみから第一原理によって独立な物理量の数を決定することは，天体物理のような複雑な系では非常に難しい．物理学的な議論から予想される自由度の幅はかなり大きく，多様体学習などデータ科学的方法によって得られる次元の情報をもとに絞り込むのが現状では最善の手段である．

この現状の方法はしかし，銀河多様体自体から進化方程式が確定できないという点で不満が残る．より直接的な銀河多様体の解釈・定式化には本質的に洗練された方法論が必要であり，我々は現在シンボリック回帰（たとえば Cranmer, 2023）などの方法によってこの問題に取り組んでいる．

5. 結論と展望

本研究では，銀河の多波長光度が張る高次元特徴空間中での銀河分布に対し，データ科学で最近発展してきた多様体学習を適用し，銀河の進化の特徴づけを行った．これにより，我々は銀河多様体と呼ばれる，多波長光度空間内のデータ点に埋め込まれた低次元非線型構造を発見した．これが銀河多様体である．そして紫外線，光学，近赤外線の光度空間における銀河の進化が，銀河多様体上の星形成と星の質量進化という 2 つのパラメーターによってよく記述されることを発見した．これは，銀河進化研究への多様体学習の有効性を示す結果である．

本研究で用いたデータは紫外線から近赤外線の波長域であった．この波長の放射は基本的に星からの寄与で，ガスの放射がそれに加わっている．しかし，さらに可視光から離れた波長の放射，たとえば γ 線，X 線など短波長側，あるいはダストの放射が卓越する中間・遠赤外線，そして原子ガス，分子ガスの放射が見えてくる電波に拡張すれば，より多くの物理過程を考慮することができる．我々は現在，中間赤外線データを加えた多波長光度空間で同じ解析を行い，得られた多様体が銀河の持つダスト粒子の放射の情報を持つという示唆を得ている．短波長側の放射は高エネルギー現象と密接にかかわっており，ブラックホールへの物質降着による超高エネルギー放射天体である活動銀河核の進化も含めた理論構築も視野に入れることができる．

また，ここでは議論を多波長測光調査に限定したが，この方法は分光探査データにも用いる

ことができる．さらに放射だけでなく，力学的，構造的，環境的特性など，銀河の他の特性も含めることにより，銀河の形成や相互作用，合体など，より力学的な物理現象も含めた銀河進化の大統一理論の構築も視野に入ってくる．

本稿では銀河進化を銀河多様体上のベクトル場として記述したが，銀河多様体を宇宙年齢ごとに構成できるデータが得られれば，多様体上の進化ベクトルとしてではなく，銀河多様体自体の進化として銀河進化を記述することが可能になる．遠方宇宙まで十分なデータが取得できるような将来の赤方偏移探査が遂行された暁には，この方法はさらに強力な方法論となると期待される．

このように，多様体学習は銀河の形成と進化に関する研究に根本的に新しい洞察を提供する強力な方法である．しかしこれは，数ある物理学への応用の可能性のうち，特にシンプルな例の一つである．多様体学習は物理現象の記述にとどまらず，広く物理法則の発見と統一のための新たな方法論となるであろう．

注．

[1] 可視光天文学（紫外線，近赤外線を含む）では，天体の光度を L としたとき

$$M \equiv -2.5 \log_{10} L + \text{銀河までの距離に寄らない定数}$$

を絶対等級とよび，これが光度の代替量として広く用いられる．等級についての正確な定義は付録 B で説明する．

[2] 天体探査データにおいて，装置の検出限界まで天体が漏れなく検出されているとき，データはコンプリートであるという．

[3] ある広い波長範囲で測定された光度を広帯域バンド光度と呼ぶ．

[4] 銀河のスペクトルは複雑な波長依存性を持つため，宇宙論的赤方偏移によって波長が伸びるだけでなくスペクトルの観測する波長も変わる．この変化分の補正を k 補正と呼ぶ．詳しい定式化は付録 B を参照．

[5] 詳しくは，SDSS の分光分類指標（specclass）で GAL_EM および GALAXY とマークされた銀河．つまり，クェーサーや 1 型セイファート銀河など，ブラックホール起源の放射が支配的な銀河はサンプルから除去されている．

[6] この場合は距離の指標として用いられている．詳細は A にて解説している．

[7] 天文学ではこのようなデータを volume-limited と呼ぶ．定訳はないが，考えている体積内で，ある光度 $L_\nu^{\rm lim}$ よりも明るい天体を数え落としなく含んだサンプルを意味する．

[8] 添字 AB は物理的な定義に基づく AB 等級という等級であることを表す（付録 B）．

[9] この目的で広く用いられている Farahmand-Szepesvári-Audibert（FSA） dimension estimator（Farahmand et al., 2007）による評価も試みたが，検証してみたところ近傍点の設定によって著しく次元の推定値が変わることが判明したため，本論文では FSA の結果は議論に用いない．FSA 自体の性能の問題については更なる検証が必要である．

[10] 「化学」進化という奇妙な用語は，この理論が銀河内の星と星間物質（ISM）の化学組成の解析に用いられてきたことに由来している

謝　辞

本論文は統計数理特集号『諸科学における統計数理モデリングの拡がり II』のために執筆したが，責任著者のスケジュール上の都合により投稿が間に合わなかったため，一般論文として投稿したものである．特集号担当編集者の島谷健一郎氏の示唆に深く感謝する．本研究は日本学

術振興会(JSPS 科研費補助金(24H00247，21H01128，19K03937，および JP17H06130)の補助を受けている．また本研究の一部は，住友財団平成 30 年度基礎科学研究費補助金(180923)，統計数理研究所共同研究費「データサイエンスによる銀河進化研究の新展開」の支援も受けて行った．

本論文執筆のきっかけとなるアイディアを頂き，本稿にも貴重なコメントをして頂いた栗木哲氏に心から御礼申し上げる．有益な示唆を頂いた今泉允聡氏，矢野恵佑氏に深く感謝する．

付　録

A. 宇宙論の基礎

A.1 Friedmann-Lemaître-Robertson-Walker 計量とスケール因子

一般相対性理論は，時空がその幾何学的構造によって特徴付けられることを示し，微分幾何学が使用するべき基本的な枠組みであることを示した．微分幾何において，多様体上の局所的性質を記述する基本的な概念が線素(line element)あるいは計量(metric)である．Friedmann-Lemaître-Robertson-Walker(FLRW)計量は，均一で等方的な時空の(局所的)幾何構造を表すために提唱された計量で，宇宙論のモデルとして広く用いられている．

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x_\mu \mathrm{d}x_\nu = -c^2\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2)\right] \tag{A.1}$$

ここで $g_{\mu\nu}$ が計量テンソル，t は時刻，r, θ, ϕ は極座標の動径距離および角度方向の座標を表す．また K はガウス曲率，$a(t)$ はスケール因子である．スケール因子は宇宙膨張による空間のスケールの変化を表現しており，宇宙論の慣習上 $a_0 \equiv a(t_0) = 1$ (t_0: 現在の宇宙年齢)と規格化して用いられる．ここでは K は (長さ)$^{-2}$ の次元を持つように定義を採用する．この場合 r は長さの次元を持ち，$a(t)$ は無次元になる．

スケール因子を導入すると，宇宙の進化による天体間の距離の変化から宇宙膨張による「自明な」距離の変化を除いた正味の変化分を表現することができる．即ち，距離をあらわすベクトルを $\vec{r}$ とすると

$$\vec{r} = a(t)\vec{x} \tag{A.2}$$

と書ける．この $\vec{x}$ は宇宙膨張の影響を除いた距離の指標であり，共動座標(comoving coordinate)と呼ばれる．

A.2 宇宙論的赤方偏移

過去のある時刻 $t = t_{\mathrm{em}}$ で共動座標 $(r, \phi, \theta) = (r_{\mathrm{e}}, 0, 0)$ で放射された光が，時刻 $t = t_0$ において原点 $(r, \phi, \theta) = (0, 0, 0)$ に位置する観測者に到達する状況を考える．相対性理論では，光はヌル測地線(null geodesic)に沿って移動する．ヌル測地線とは $\mathrm{d}s^2 = 0$ を満たす経路のことで，最小作用の原理で実現する軌道と考えてよい．式(A.1)を用いれば，距離と時間を

$$\frac{c\mathrm{d}t}{a(t)} = \frac{\mathrm{d}r}{\sqrt{1-Kr^2}} \tag{A.3}$$

のように関係付けられる．宇宙膨張の光への影響を議論するため，変数 r を

$$\chi \equiv \int_0^{r_{\mathrm{em}}} \frac{\mathrm{d}r}{\sqrt{1-Kr^2}}, \tag{A.4}$$

によって座標距離 χ に，時間 t を

$$\eta \equiv \int^{t} \frac{\mathrm{d}t'}{a(t')}, \tag{A.5}$$

によって共形時間 η に変換すると，光の満たすべき方程式は

$$c(\eta_{\mathrm{obs}} - \eta_{\mathrm{em}}) = \chi, \tag{A.6}$$

となる．ここで c は光速，η_{obs} は光が観測者に到達した時刻を表す．

時刻 $\eta = \eta_0$ および $\eta = \eta_0 + \delta\eta_0$ で放出された光が，それぞれ η_1 および $\eta_1 + \delta\eta_1$ で観測者に到達したとする．式(A.6)の右辺は η によらないことから

$$\delta\eta_0 = \delta\eta_1 \tag{A.7}$$

すなわち

$$\frac{\delta t_0}{a(t_0)} = \frac{\delta t_1}{a(t_1)} \tag{A.8}$$

となる．時間間隔 δt_0 および δt_1 において，光の位相が保存することから，

$$a(t_0)\nu_0 = a(t_1)\nu_1 \Longleftrightarrow \frac{\lambda_0}{a(t_0)} = \frac{\lambda_1}{a(t_1)} \tag{A.9}$$

が得られる．ここで ν は光の振動数で，波長とは $\lambda\nu = c$ の関係がある．

膨張する宇宙では $a(t_0) > a(t_1)$ なので，観測される時刻での波長は放射のときよりも長くなる，即ち $\lambda_0 > \lambda_1$ となる．この波長の伸びが宇宙膨張による赤方偏移と呼ばれている．赤方偏移 z は

$$z \equiv \frac{\lambda_0 - \lambda_1}{\lambda_1} = \frac{a(t_0)}{a(t_1)} - 1 \tag{A.10}$$

と定義される．

ここで，上記のようにスケール因子 $a(t)$ を $a(t_0) = 1$ となるよう規格化する．このとき $a(t)$ と z との関係は

$$z = \frac{1}{a(t)} - 1\ , \tag{A.11}$$

あるいは

$$a(t) = \frac{1}{1+z} \tag{A.12}$$

となる．

A.3 フリードマン方程式と宇宙論パラメータ

ここまでの準備，およびアインシュタイン方程式(Einstein's equation)を用いれば，膨張する宇宙を記述するための「運動方程式」を導くことができる．アインシュタイン方程式とは，時空の歪みとエネルギー密度を関連付けた一般相対性理論の基本方程式であり，

$$R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu} + \Lambda g^{\mu\nu} = \frac{8\pi G}{c^4}T^{\mu\nu} \tag{A.13}$$

と書かれる．ここで $R^{\mu\nu}$ はリッチテンソル(Ricci tensor)，$R \equiv R^{\alpha}{}_{\alpha}$ はそのトレースでリッチスカラー(Ricci scalar)，G はニュートンの重力定数，Λ は宇宙定数である．詳細は相対性理論の文献に譲るが，リッチテンソルは計量 $g_{\mu\nu}$ とその微分量の関数であり，時空の歪みを表現する 2 階の共変テンソルである（たとえば Foster and Nightingale, 2010）．FLRW 計量の場合，時空の運動方程式は

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G\rho}{3} - \frac{c^2 K}{a^2} + \frac{c^2\Lambda}{3} \tag{A.14}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{c^2\Lambda}{3}\ , \tag{A.15}$$

となる．これはフリードマン方程式(Friedmann equations)と呼ばれている．ここで ρ は宇宙の密度，p は圧力を現す．式(A.14)はアインシュタイン方程式(式(A.13))の 00 成分，式(A.15)はトレースから求められる．

さらに，宇宙論パラメーターと呼ばれる以下の量を定義すると，宇宙膨張の物理量依存性がより具体的に記述できる．

- ハッブルパラメータ：

$$H(t) \equiv \frac{\dot{a}(t)}{a(t)}\ . \tag{A.16}$$

これは時間依存する量であるが，観測的には現在値 H_0 が用いられることが多い．また無次元化した $h = H_0/100$ もよく用いる．

- 密度パラメータ：

$$\Omega_{\rm M}(t) \equiv \frac{\rho(t)}{\rho_{\rm c}(t)} \equiv \frac{8\pi G\rho(t)}{3H^2(t)}\ , \tag{A.17}$$

ここで $\rho_{\rm c}$ は臨界密度パラメータと呼ばれる量で，現在の宇宙では $\rho_{\rm c} = 1.88 \times 10^{-29} h^2\ [{\rm g\,cm^{-3}}]$ と測定されている．

- 無次元宇宙項パラメータ：

$$\Omega_\Lambda(t) \equiv \frac{c^2\Lambda}{3H^2(t)} \tag{A.18}$$

- 曲率パラメータ：

$$\Omega_K(t) \equiv -\frac{c^2 K}{a(t)^2 H^2(t)} \tag{A.19}$$

式(A.16)–(A.19)を用いると，式(A.14)は

$$\Omega_{\rm M}(t) + \Omega_\Lambda(t) + \Omega_K(t) = 1 \tag{A.20}$$

と表せる．これは一般の宇宙年齢 t で成り立つので，現在の宇宙年齢 t_0 では

$$\Omega_{\rm M0} + \Omega_{\Lambda 0} + \Omega_{K0} = 1 \tag{A.21}$$

となる．本論文で採用した宇宙論パラメータ $h = H_0/100\ [{\rm km\ s^{-1}\ Mpc^{-1}}] = 0.7$, $\Omega_{\Lambda 0} = 0.7$, $\Omega_{\rm M0} = 0.3$, 曲率パラメータ $\Omega_{K0} = 0$ は最新の精密宇宙論的観測から支持されるもので，現在の宇宙が空間的に平坦で，宇宙定数(または暗黒エネルギー)により加速膨張している宇宙モデルを示唆している．

B. 等級

検出器の単位面積を単位時間に通過するエネルギーを放射流束(flux)と呼び，さらに単位振動数当たり(あるいは単位波長当たり)の量を放射流束密度(flux density)と呼ぶ．そして天文学では，放射流束あるいは放射流束密度の代わりに，その対数の -2.5 倍に対応する数値である等級(magnitude)を用いるのが習慣となっている．これは次のように表される．

$$m_{\nu\mathrm{obs}} = -2.5\log_{10} S_{\nu_{\mathrm{obs}}} + \text{定数}\,. \tag{B.1}$$

歴史的に，等級の定義に現れる定数はベガ（こと座 α 星）を 0 となるように定義されたが，これには物理的な根拠はなく，長らく混乱の元となってきた（たとえば Bessell, 2005）．

AB 等級（AB magnitude）は，放射流束密度から等級への系統的かつ物理的な変換のために Oke and Gunn（1983）によって導入された．AB 等級は次のように表される．

$$m_{\mathrm{AB},\nu_{\mathrm{obs}}} = -2.5\log_{10} S_{\nu_{\mathrm{obs}}} - 48.60\,. \tag{B.2}$$

ここで，S_ν の単位は $[\mathrm{erg\,s^{-1}\,cm^{-2}\,Hz^{-1}}]$ である．下付きの “obs” は，その振動数が地球の静止座標系で観測される値であることを示す．また，下付き “em” はそれが光が放射されたときの銀河の静止系での振動数であることを表している．よって，宇宙論的赤方偏移 z を用いると，$\nu_{\mathrm{em}} = (1+z)\nu_{\mathrm{obs}}$ となる．また定数 -48.60 は観測される振動数 ν にも波長 λ にも依存しないことに注意してほしい．したがって，絶対等級（absolute magnitude）$M_{\mathrm{AB},\nu}$ と単色光度（monochromatic luminosity）L_ν の関係は次のようになる．

$$\begin{aligned} M_{\mathrm{AB},\nu_{\mathrm{em}}} &= m_{\mathrm{AB},\nu_{\mathrm{obs}}} - 25 - 5\log_{10} d_L(z) \\ &= -2.5\log_{10}\left[\frac{(1+z)L_{\nu_{\mathrm{em}}}}{4\pi d_{\mathrm{L}}(z)^2}\right] - 48.60 - 25 - 5\log_{10} d_{\mathrm{L}}(z) \\ &= -2.5\log_{10}[(1+z)L_{\nu_{\mathrm{em}}}] - 48.60 - 25 + 2.5\log_{10}(4\pi)\,. \end{aligned} \tag{B.3}$$

ここで

$$d_{\mathrm{L}}(z) = \frac{c}{H_0}\int_0^z \frac{\mathrm{d}z'}{\sqrt{\Omega_{M0}(1+z')^3 + \Omega_{\Lambda 0}}}\,[\mathrm{Mpc}] \tag{B.4}$$

は銀河の光度距離，z はその赤方偏移，

$$\frac{c}{H_0} = 3000h^{-1}\,[\mathrm{Mpc}]\,, \tag{B.5}$$

はハッブル長，そして

$$1\,[\mathrm{pc}] = 3.086\times 10^{18}\,[\mathrm{cm}] \tag{B.6}$$

である（例えば Peebles, 1993）．

しかし，赤方偏移の効果は単純な観測波長の伸びだけではなく，観測する波長と放射した波長が異なることによる単色光度の変化分も考慮する必要がある．単位振動数当たりの放射流束密度は単色光度を用いて

$$S_{\nu_{\mathrm{obs}}}\,\mathrm{d}\nu_{\mathrm{obs}} = \frac{L_{\nu_{\mathrm{em}}}\,\mathrm{d}\nu_{\mathrm{em}}}{4\pi d_{\mathrm{L}}(z)^2} = \frac{L_{\nu_{\mathrm{obs}}(1+z)}\,\mathrm{d}(1+z)\nu_{\mathrm{obs}}}{4\pi d_{\mathrm{L}}(z)^2} = \frac{(1+z)\,L_{\nu_{\mathrm{obs}}}\,\mathrm{d}\nu_{\mathrm{obs}}}{4\pi d_{\mathrm{L}}(z)^2} \tag{B.7}$$

となるので，

$$S_{\nu_{\mathrm{obs}}} = \frac{(1+z)\,L_{\nu_{\mathrm{obs}}}}{4\pi d_{\mathrm{L}}(z)^2} \tag{B.8}$$

が得られる．よって単位波長当たりの放射流束密度 S_λ は

$$S_{\lambda_{\mathrm{obs}}} = \frac{L_{\lambda_{\mathrm{obs}}/(1+z)}}{4\pi d_{\mathrm{L}}(z)^2(1+z)} \tag{B.9}$$

となる．正確には，観測される放射流束はある波長範囲だけの放射を透過させるフィルターを通した量である．観測するフィルターの有効波長 λ_0 における観測バンド流束 $S^{[\lambda_0]}$ $[\mathrm{erg\,s^{-1}cm^{-2}}]$ を

$$S^{[\lambda_0]} \equiv \int S_\lambda R_\lambda^{[\lambda_0]} \mathrm{d}\lambda \tag{B.10}$$

で定義する．ここで $R_\lambda^{[\lambda_0]}$ はフィルターの波長感度特性を表す関数で，応答関数と呼ばれる．観測される放射流束は式(B.9)より

$$S^{[\lambda_0]} = \frac{1}{4\pi {d_\mathrm{L}(z)}^2 (1+z)} \int L_{\frac{\lambda}{(1+z)}} R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda = \frac{\int L_\lambda R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda}{4\pi {d_\mathrm{L}(z)}^2 (1+z)} \frac{\int L_{\frac{\lambda}{(1+z)}} R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda}{\int L_\lambda R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda} \tag{B.11}$$

と表せる．ここで，次の項

$$\frac{\int L_\lambda R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda}{4\pi {d_\mathrm{L}(z)}^2 (1+z)} \equiv \tilde{S}^{[\lambda_0]} \tag{B.12}$$

は宇宙論的赤方偏移がない場合に得られる仮想的なバンド放射流束を表している．式(B.11)の残りの部分は，赤方偏移によって天体のスペクトルの異なる波長範囲が観測されることで生じる放射流束の変化を表している．これが k 補正である．等級で表せば，

$$\begin{aligned} m_{\lambda_0} - M_{\lambda_0} &= -2.5 \log S^{[\lambda_0]} + 2.5 \log \left[\frac{4\pi {d_\mathrm{L}(z)}^2 (1+z)}{4\pi (10\ [\mathrm{pc}])^2} \tilde{S}^{[\lambda_0]} \right] \\ &= 2.5 \log \left[{d_\mathrm{L}(z)}^2 (1+z) \right] - 2.5 \log \left[\frac{\int L_{\frac{\lambda}{(1+z)}} R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda}{\int L_\lambda R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda} \right] + 25 \\ &= 5 \log d_\mathrm{L}(z) + 2.5 \log (1+z) - 2.5 \log \left[\frac{\int L_{\frac{\lambda}{(1+z)}} R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda}{\int L_\lambda R_\lambda^{[\lambda_0]} \,\mathrm{d}\lambda} \right] + 25 \end{aligned} \tag{B.13}$$

となる．式(B.13)の第 2 項と第 3 項は等級で表した k 補正(K 補正と書かれる)を表す（Oke and Sandage, 1968）．第 2 項は帯域幅の伸びの影響，第 3 項はバンド放射流束の変化分を表している．

C. 銀河の星形成史と可視光スペクトル

ここでは星形成史と銀河の可視光(紫外線から近赤外線まで)スペクトルの関係について定量的に説明する．簡単のため，銀河の可視光スペクトルが星のみの寄与からなると仮定すると，ある時刻 t における銀河のスペクトル $L_\lambda(t)$ は

$$L_\lambda(t) = \int_0^t \int_{\mathcal{M}_\mathrm{low}}^{\mathcal{M}_\mathrm{up}} \mathrm{SFR}(t-\tau)\, \mathcal{F}_{\lambda, Z_{(t-\tau)}}(m, \tau) \Phi(m) \,\mathrm{d}m \,\mathrm{d}\tau \ , \tag{C.1}$$

で与えられる．ここで $\mathcal{M}_\mathrm{up}, \mathcal{M}_\mathrm{low}$ は形成される星の質量の上限と下限，$\Phi(m)$ は初期質量関数(initial mass function: IMF)，$\mathcal{F}_{\lambda, Z_{(t-\tau)}}$ は質量 m，金属量(metallicity) $Z(t-\tau)$ の星のスペクトルを表す．初期質量関数とは，集団的に形成される星の質量の分布関数(確率密度関数)に比例する量である．なお，ここで規格化は

$$\int_{\mathcal{M}_\mathrm{low}}^{\mathcal{M}_\mathrm{up}} m \Phi(m) \,\mathrm{d}m = 1 \ [\mathcal{M}_\odot] \tag{C.2}$$

としている．即ち $\Phi(m)$ はどのくらいの質量の星がどのくらいの数形成されるかを表す．金属

量とは，星の中心における核融合によって生成された重元素が全星間物質の質量に占める割合である．時刻 t で観測される銀河の中の星はそれ以前の時刻に形成されたことを考慮するため，スペクトルおよび金属量が時刻 t ではなく $t-\tau$ で評価され，積分されている．

より詳細な理論モデルではこれに加えてガスからの放射を考慮し，星などから形成されたダスト(炭素やケイ素などの重元素からなる塵の微粒子)による減光を加味する必要があるが，本研究の主要な成果を理解するためには式(C.1)を考えれば十分である．しかし，これは古典的銀河進化理論，すなわち孤立した銀河が進化していくことを念頭に置いた理論によって与えられているが，1 章で述べたように，銀河は内部での星形成だけでなく合体による成長によっても進化してゆく．つまり式(C.1)が表現できるのは合体していく銀河の断片の内的進化のみであり，銀河の統計的進化の記述のためには本研究のようなアプローチが必要になることを強調しておく．

参 考 文 献

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Allende Prieto, C., An, D., Anderson, K. S. J., Anderson, S. F., Annis, J., Bahcall, N. A., Bailer-Jones, C. A. L., Barentine, J. C. and Bassett, B. A. (2009). The seventh data release of the sloan digital sky survey, *Astrophysical Journal Supplement Series*, **182**(2), 543–558, https://dx.doi.org/10.1088/0067-0049/182/2/543.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**(6), 716–723, https://dx.doi.org/10.1109/TAC.1974.1100705.

Bernstein, M., Silva, V. D., Langford, J. C. and Tenenbaum, J. B. (2001). Graph approximations to geodesics on embedded manifolds, Unpublished Technical Report, Stanford University, California.

Bessell, M. S. (2005). Standard photometric systems, *Annual Review of Astronomy and Astrophysics*, **43**(1), 293–336, https://dx.doi.org/10.1146/annurev.astro.41.082801.100251.

Blanton, M. R. (2006). Galaxies in SDSS and DEEP2: A quiet life on the blue sequence?, *Astrophysical Journal*, **648**(1), 268–280, https://dx.doi.org/10.1086/505628.

Bouveyron, C. and Brunet, C. (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace, *Statistics and Computing*, **22**(1), 301–324, https://dx.doi.org/10.1007/s11222-011-9249-9.

Brosche, P. (1973). The manifold of galaxies. galaxies with known dynamical parameters, *Astronomy and Astrophysics*, **23**, 259–268.

Chilingarian, I. V. and Zolotukhin, I. Y. (2012). A universal ultraviolet-optical colour-colour-magnitude relation of galaxies, *Monthly Notices of the Royal Astronomical Society*, **419**(2), 1727–1739, https://dx.doi.org/10.1111/j.1365-2966.2011.19837.x.

Chilingarian, I. V., Zolotukhin, I. Y., Katkov, I. Y., Melchior, A.-L., Rubtsov, E. V. and Grishin, K. A. (2017). RCSED — A value-added reference catalog of spectral energy distributions of 800,299 galaxies in 11 ultraviolet, optical, and near-infrared bands: Morphologies, colors, ionized gas, and stellar population properties, *Astrophysical Journal Supplement Series*, **228**(2), https://dx.doi.org/10.3847/1538-4365/228/2/14.

Cooray, S., Takeuchi, T. T., Kashino, D., Yoshida, S. A., Ma, H.-X. and Kono, K. T. (2023). Characterizing and understanding galaxies with two parameters, *Monthly Notices of the Royal Astronomical Society*, **524**(4), 4976–4995, https://dx.doi.org/10.1093/mnras/stad2129.

Cranmer, M. (2023). Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, arXiv, https://doi.org/10.48550/arXiv.2305.01582.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs, *Numerische Mathematik*, **1**(1), 269–271.

Djorgovski, S. (1992). *Morphological and Physical Classification of Galaxies* (eds. G. Longo, M.

Capaccioli and G. Busarello), 337–356, Springer Netherlands, Dordrecht.

Farahmand, A. m., Szepesvári, C. and Audibert, J.-Y. (2007). Manifold-adaptive dimension estimation, *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, 265–272, Association for Computing Machinery, New York, https://dx.doi.org/10.1145/1273496.1273530.

Floyd, R. W. (1962). Algorithm 97: Shortest path, *Communications of the ACM*, **5**(6), 345.

Foster, J. and Nightingale, J. (2010). *A Short Course in General Relativity*, Springer, New York.

Ginolfi, M., Hunt, L. K., Tortora, C., Schneider, R. and Cresci, G. (2020). Scaling relations and baryonic cycling in local star-forming galaxies, I. The sample, *Astronomy and Astrophysics*, **638**, https://dx.doi.org/10.1051/0004-6361/201936304.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*, MIT Press, Cambridge, Massachusetts.

Hunt, L., Magrini, L., Galli, D., Schneider, R., Bianchi, S., Maiolino, R., Romano, D., Tosi, M. and Valiante, R. (2012). Scaling relations of metallicity, stellar mass and star formation rate in metal-poor starbursts — I. A Fundamental Plane, *Monthly Notices of the Royal Astronomical Society*, **427**(2), 906–918, https://dx.doi.org/10.1111/j.1365-2966.2012.21761.x.

Lilly, S. J., Carollo, C. M., Pipino, A., Renzini, A. and Peng, Y. (2013). Gas regulation of galaxies: The evolution of the cosmic specific star formation rate, the metallicity-mass-star-formation rate relation, and the stellar content of halos, *The Astrophysical Journal*, **772**(2), https://dx.doi.org/10.1088/0004-637X/772/2/119.

Lin, L., St. Thomas, B., Zhu, H. and Dunson, D. B. (2017). Extrinsic local regression on manifold-valued data, *Journal of the American Statistical Association*, **112**(519), 1261–1273, https://dx.doi.org/10.1080/01621459.2016.1208615.

Liu, S., Maljovec, D., Wang, B., Bremer, P.-T. and Pascucci, V. (2017). Visualizing high-dimensional data: Advances in the past decade, *IEEE Transactions on Visualization and Computer Graphics*, **23**(3), 1249–1268, https://dx.doi.org/10.1109/TVCG.2016.2640960.

Ma, Y. and Fu, Y. (2012). *Manifold Learning Theory and Applications*, CRC Press, Boca Raton.

McInnes, L., Healy, J., Saul, N. and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection, *Journal of Open Source Software*, **3**(29), https://dx.doi.org/10.21105/joss.00861.

McInnes, L., Healy, J. and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for dimension reduction, arXiv, https://dx.doi.org/https://doi.org/10.48550/arXiv.1802.03426.

Oke, J. B. and Gunn, J. E. (1983). Secondary standard stars for absolute spectrophotometry, *Astrophysical Journal*, **266**, 713–717, https://dx.doi.org/10.1086/160817.

Oke, J. B. and Sandage, A. (1968). Energy distributions, K corrections, and the Stebbins-Whitford effect for giant elliptical galaxies, *The Astrophysical Journal*, **154**, https://dx.doi.org/10.1086/149737.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine learning in python, *Journal of Machine Learning Research*, **12**, 2825–2830.

Peebles, P. J. E. (1993). *Principles of Physical Cosmology*, Princeton University Press, Princeton, New Jersey, https://dx.doi.org/10.1515/9780691206721.

Rodighiero, G., Daddi, E., Baronchelli, I., Cimatti, A., Renzini, A., Aussel, H., Popesso, P., Lutz, D., Andreani, P., Berta, S., Cava, A., Elbaz, D., Feltre, A., Fontana, A., Förster Schreiber, N. M., Franceschini, A., Genzel, R., Grazian, A., Gruppioni, C., Ilbert, O., Le Floch, E., Magdis, G., Magliocchetti, M., Magnelli, B., Maiolino, R., McCracken, H., Nordon, R., Poglitsch, A., Santini, P., Pozzi, F., Riguccini, L., Tacconi, L. J., Wuyts, S. and Zamorani, G. (2011). The lesser role of starbursts in star formation at z = 2, *The Astrophysical Journal*, **739**(2), https://dx.doi.org/10.1088/2041-8205/739/2/L40.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding,

Science, **290**(5500), 2323–2326, https://dx.doi.org/10.1126/science.290.5500.2323.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464, https://dx.doi.org/10.1214/aos/1176344136.

Siudek, M., Małek, K., Pollo, A., Krakowski, T., Iovino, A., Scodeggio, M., Moutard, T., Zamorani, G., Guzzo, L., Garilli, B., Granett, B. R., Bolzonella, M., de la Torre, S., Abbas, U., Adami, C., Bottini, D., Cappi, A., Cucciati, O., Davidzon, I., Franzetti, P., Fritz, A., Krywult, J., Le Brun, V., Le Fèvre, O., Maccagni, D., Marulli, F., Polletta, M., Tasca, L. A. M., Tojeiro, R., Vergani, D., Zanichelli, A., Arnouts, S., Bel, J., Branchini, E., Coupon, J., De Lucia, G., Ilbert, O., Haines, C. P., Moscardini, L. and Takeuchi, T. T. (2018). The VIMOS Public Extragalactic Redshift Survey (VIPERS), The complexity of galaxy populations at $0.4 < z < 1.3$ revealed with unsupervised machine-learning algorithms, *Astronomy & Astrophysics*, **617**, https://dx.doi.org/10.1051/0004-6361/201832784.

Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**(5500), 2319–2323, https://dx.doi.org/10.1126/science.290.5500.2319.

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

Warshall, S. (1962). A theorem on boolean matrices, *Journal of the ACM*, **9**(1), 11–12.

Whitaker, K. E., van Dokkum, P. G., Brammer, G. and Franx, M. (2012). The star formation mass sequence out to $z = 2.5$, *The Astrophysical Journal*, **754**(2), https://dx.doi.org/10.1088/2041-8205/754/2/L29.

Zhang, H. and Zaritsky, D. (2016). Examining early-type galaxy scaling relations using simple dynamical models, *Monthly Notices of the Royal Astronomical Society*, **455**(2), 1364–1374, https://dx.doi.org/10.1093/mnras/stv2413.

New Quantification of Galaxy Evolution with Manifold Learning

Tsutomu T. Takeuchi[1,2], Sucheta Cooray[3,4], Taisei D. Yamagata[1], Aina May So[1,6], Shun-Ya S. Uchida[1], Shiro Ikeda[2], Kenji Fukumizu[2], Ryusei R. Kano[1], Kiyoaki Christopher Omori[1,5], Hai-Xia Ma[1], Wen Shi[1] and Sena A. Matsui[1]

[1]Division of Particle and Astrophysical Science, Nagoya University
[2]The Institute of Statistical Mathematics
[3]Division of Science, National Astronomical Observatory of Japan
[4]Kavli Institute Particle Astrophysics and Cosmology, Stanford University
[5]Department of Astronomy and Physics, Saint Mary's University
[6]Department of Physics, Gakushuin University

Matter in the early Universe was almost uniform, and a slightly dense region grew by gravity, finally into a galaxy. Astrophysics has been trying to unveil the complex physical phenomena that caused the formation and evolution of galaxies through the history of the Universe of 13 billion years from the first principles of physics. However, since present-day astrophysical big data contain more than 100 explanatory variables, such a conventional methodology faces a limit to deal with such data. We, instead, elucidate the physics of galaxy evolution by applying manifold learning, one of the latest methods of data science, to a feature space spanned by galaxy luminosities and cosmic time. We discovered a low-dimensional nonlinear structure of data points in this space, referred to as the galaxy manifold. We found that the galaxy evolution in the ultraviolet–optical–near infrared luminosity space is well described by two parameters, star formation and stellar mass evolution, on the manifold. We also discuss a possible way to connect the manifold coordinates to physical quantities.

Key words: Galaxy evolution, galaxy formation, star formation rate, stellar mass, multiwavelength luminosity, manifold learning.