統計数理(2024)
第 72 巻 第 2 号 273–303

［研究詳解］

高次元統計解析で探る銀河の分子ガスの物理状態と天文学への展望

竹内 努[1,2]・矢田 和善[3]・江頭 健斗[4]・青嶋 誠[3]・吉川 耕司[5]・石井 晶[4]・加納 龍生[1]・施 文[1]・曹 愛奈[1,8]・馬 海霞[1]・松井 瀬奈[1]・中西 康一郎[6,9]・クレ スチェータ[6,7]・河野 孝太郎[10]

(受付 2024 年 1 月 30 日；改訂 7 月 18 日；採択 7 月 19 日)

要　　旨

現代科学のデータ解析では，次元が d，標本数が n であるデータにおいて，$n \ll d$ となる場合が往々にして見られる．天文学では従来，このような状況は不適切と見なされ，データの持つ次元のほとんどの情報を破棄して $d < n$ にする以外に選択肢はないと考えられていた．$n \ll d$ となるデータは，高次元小標本(high-dimensional low sample size: HDLSS)と呼ばれる．HDLSS を含む高次元データの解析には，高次元データ特有の新しい理論と方法論が必要となる．青嶋らの研究グループは，それらを高次元統計解析と名付け，さまざまな統計理論とその方法論を構築した．本論文ではまず高次元統計解析を紹介し，その代表的な手法であるノイズ掃出し主成分分析(principal component analysis by noise-reduction method: NRPCA)と自動スパース主成分分析(automatic sparse PCA: A-SPCA)を導入する．これらの方法の実際のデータへの応用例として，アタカマ大型ミリ波/サブミリ波干渉計(Atacama Large Millimeter/Submillimeter Array: ALMA)が撮影した近傍宇宙の典型的な爆発的星形成(スターバースト)銀河 NGC 253 の分光マップに適用する．ALMA の分光マップは典型的な HDLSS データである．NGC 253 の中心部は全体として回転しており，放射される電磁波は回転速度に対応したドップラー効果によって波長が偏移している．元のデータはドップラー効果がそのまま表れており，予備解析としてまずこの元データを解析した．その結果，高次元 PCA はドップラー効果を正確に抽出し，物理的なモデルを介在させることなく回転の空間構造を正確に記述できた．次に，同じ方法を回転のドップラーシフトを補正したデータに適用し，より詳細なスペクトルの特徴を解析

[1] 名古屋大学 素粒子宇宙物理学専攻：〒 4664–8602 名古屋市千種区不老町
[2] 統計数理研究所 統計的機械学習研究センター 客員：〒 190–8562 東京都立川市緑町 10–3
[3] 筑波大学 数理物質系数学域：〒 305–8571 つくば市天王台 1–1–1
[4] 東京理科大学 創域理工学部：〒 278–8510 千葉県野田市山崎 2641
[5] 筑波大学 計算科学研究センター：〒 305–8571 つくば市天王台 1–1–1
[6] 国立天文台：〒 181–8588 東京都三鷹市大沢 2–21–1
[7] 日本学術振興会 特別研究員(PD)
[8] 学習院大学 理学部：〒 171–8588 東京都豊島区目白 1–5–1
[9] 総合研究大学院大学 物理科学研究科：〒 181–8588 東京都三鷹市大沢 2–21–1
[10] 東京大学 大学院理学系研究科附属天文学教育研究センター：〒 181–0015 東京都三鷹市大沢 2–21–1

した．これらの解析により，NRPCA と A-SPCA によって ALMA スペクトルマップの非常に複雑な特性を定量化できることが示された．特に，これらの方法はモデルを仮定することなく NGC 253 の中心からの大規模な質量流の情報を抽出でき，高次元統計学の強力さを示した．この方法は，分光サーベイデータだけでなく，HDLSS であるあらゆるタイプのデータに適用可能である．

キーワード：高次元統計解析，高次元主成分分析，星間物質，分子輝線，爆発的星形成，銀河進化．

1. はじめに

最近 10 年間で「ビッグデータ」という用語は広く認識され，天文学を含む科学研究のほぼすべての分野でみられるようになってきている．現代の天文観測装置は，従来に比べ圧倒的に大量のデータを取得でき，それらの物理状態についての定量的情報を提供してくれる．天文学で最も典型的なビッグデータとして，Sloan Digital Sky Survey（SDSS）[1]，Dark Energy Spectroscopic Instrument（DESI）Survey[2]，Gaia[3]，PanSTARRS（Panoramic Survey Telescope And Rapid Response System）[4] などが挙げられる．そのデータサイズは数十万から数百万に達し，将来計画では数億のオーダーになると期待されている．

しかし実は，天文学にはこれとは異なった種類のビッグデータが存在する．天文学では往々にして，詳細な観測には非常に時間がかかる．たとえばアタカマ大型ミリ波/サブミリ波干渉計（Atacama Large Millimeter/submillimeter Array: 以下 ALMA）によって対象天体をマッピング観測する場合，天体上の多くの点で電波強度を測定することは一般に非常に時間がかかり，容易ではない．このような状況は，より一般的に面分光[5] で頻繁に見られる（例：Very Large Telescope（VLT）での UVES[6]，William Herschel Telescope の SAURON[7]，SDSS MaNGA data[8] など）．この種の別の典型的な例としては，電波干渉計（例えば ALMA）や X 線，紫外線，または中/遠赤外線波長の衛星機器による分光マッピング調査が挙げられる．波長（または振動数）の次元を d で表し，標本数を n で表すと，分光マッピング観測の場合往々にして $n \ll d$ となる．

従来の統計解析では，標本数 n がデータの次元 d ($n \gg d$) よりもはるかに大きいことを前提としている．すでに述べたように，様々な自然科学研究においてこの前提は成り立っていないことが多い．天体物理学では，伝統的にこのような状況は不良設定問題と見なされ，上述の分光マッピングの場合では波長分散方向の情報を大幅に捨てることによって $d < n$ としていた．2010 年までこの状況は他の研究分野でもほぼ同様で，多くの研究者は単純にその方向での更なる解析を諦めていたか，ビッグデータ時代の到来以前は，そもそもそのような問題に対処する必要さえ見出していなかった．例えば，有名なデータセット集である Hand et al. (1994) には，当時の統計学で扱われた 500 を超えるデータが掲載されているが，そのほとんどは $d < n$ と従来型の前提を満たしている．しかし 1990 年代後半，情報科学の発展に伴い，突如として $d \gg n$ なる高次元データが脚光を浴びるようになった．有名な例として，Golub et al. (1999) では，白血病患者の遺伝子発現データが扱われ，次元数 d は 7129 であるのに対し，標本数 n は 72 であった．1990 年代の統計学は $d < n$ を大前提にして構成されていたため，このような場合の統計的推測に精度を保証することができなかった．

標本数がデータの次元数よりも低い ($n \ll d$) データには，以下に代表されるようなさまざまな問題が知られている．

(1) 標本共分散行列の逆行列が不安定もしくは存在しないので，古典的な統計学では扱えない．

(2) さまざまなタイプの「次元の呪い」が生じる．
(3) 往々にして計算コストが膨大になる．

これらの問題により，古典的な統計学の方法では高次元のデータ空間の特性を捉えることができず，データに内在する豊富な情報を活用できない．よって $n \ll d$ のデータを解析するには，高次元データ特有の統計理論と方法論を構築する必要があった．

このような状況は 2000 年代に劇的に変化した．まず，確率論とランダム行列理論に基づく理論物理学の分野から重要な結果がもたらされた．d と n が同じオーダーで増大する（すなわち $n/d \longrightarrow c > 0$ (c：定数)）場合に，母集団がガウス分布という仮定の下，標本共分散行列の固有値の漸近的挙動が議論された（たとえば Baik et al., 2005; Baik and Silverstein, 2006; Johnstone, 2001; Paul, 2007）．ただし，HDLSS データにおいては，$n/d \longrightarrow c > 0$，および母集団がガウス分布という仮定は現実的ではない．

Hall らは論文 Hall et al. (2005) において，高次元小標本 (high-dimensional low sample size: HDLSS) という新しい枠組みでの漸近理論を導入した．彼らの研究では，標本数 n を固定し，データ次元 d を $d \longrightarrow \infty$ と極限をとっている．これは，d を固定して $n \longrightarrow \infty$ とする従来の漸近理論からの革新的な脱却であった．

Hall et al. (2005) および Ahn et al. (2007) は HDLSS データの重要な幾何学的な表現について説明したものの，それでもガウス分布は仮定されていた．対照的に，Yata and Aoshima (2012) はこの仮定を取り払った新しい理論を展開し，ガウス分布の仮定が満たされない場合について全く異なるタイプの幾何学的表現も発見した (詳細は Yata and Aoshima, 2012 や 青嶋・矢田, 2019 を参照のこと)．これらの研究は，高次元のデータ解析に関する数理統計学の研究の基礎をなし，現在流行のビッグデータサイエンスの重要な部分として発展した．これらの新理論は，ゲノム解析，医学，神経科学，画像および形状解析などのさまざまな研究分野に新たな光をもたらしている．そして青嶋 誠，矢田 和善，石井 晶と共同研究者らは，この種の問題を扱う統計的方法論の枠組みをさらに発展させた（たとえば Aoshima and Yata, 2011, 2014, 2015, 2019; Yata and Aoshima, 2012, 2013; Ishii et al., 2016; およびその参考文献）．高次元統計解析自体の数学的背景に関心のある読者は，いくつかの総説論文を参照されたい（たとえば Aoshima and Yata, 2017; Aoshima et al., 2018）．

ところが，他の分野での大きな進歩にもかかわらず，この新しい統計的方法論は天文学においては全く知られていなかった．本論文では，天文学における HDLSS データの高次元統計解析の最初の応用例を紹介する．天文データに対する高次元統計解析手法のパフォーマンスを示すため，近傍銀河 NGC 253 の中心領域の分光マップに応用した結果を示す．NGC 253 は，近くにある典型的なスターバースト銀河[9]であり，さまざまな波長で観測されている（Rieke et al., 1980）．本研究では高次元主成分分析 (PCA) を ALMA 分光マップ（Ando et al., 2017）に適用した結果を示す．Ando et al. (2017) は，ALMA Band 7 に基づく NGC 253 の中心 $\sim$ 200 pc (パーセク) の領域における 8 pc $\times$ 5 pc の分解能での分光マップを取得した[10)11)]．以下で詳細に紹介するように，ALMA マップは通常 HDLSS データであるとみなせる．NGC 253 のスペクトルには，非常に多数の分子輝線が存在する（たとえば Ando et al., 2017; Martín et al., 2021）．本研究のもう 1 つの焦点は，輝線やバンド輝線[12]が非常に多数存在する ALMA のスペクトルを PCA によって客観的に分類することである．銀河の進化は，主として星間物質 (interstellar medium: ISM) から星への遷移である星形成によって駆動される．星形成進化には ISM のさまざまな相[13]が関連しており，ISM の進化は銀河の進化の完全な理解に向けた鍵となる．分光観測は，銀河の物質の情報を抽出して解釈するために非常に重要な手段である（たとえば Martín et al., 2021）．よって，高次元 PCA による輝線スペクトル分類が有効であるならば，これは銀

河進化研究のための強力なツールとなり得る．

本論文の構成は以下のようになっている．2 章では高次元 PCA に重点を置き，HDLSS データと高次元統計解析の概念を紹介する．ここでは詳細な数学的証明なしに，HDLSS データの驚くべき性質を特徴付けるいくつかの重要な定理を提示する．次に 3 章では天文データの実際の解析を示す．ここでは本解析に使用する NGC 253 の ALMA マップの性質について説明する．本論文のデータ解析は 2 段階構成になっている．まず 4 章では高次元統計解析が分光マップデータの解析に本当に有効かどうかを検証する．その優れたパフォーマンスを確認し，5 章では NGC 253 分子線のより詳細な解析に進む．結論は 6 章で示す．付録 A では星間物質からの分子輝線の基礎，付録 B では放射のドップラー効果とデータへの補正法を解説している．本論文は，主として Takeuchi et al. (2024)の内容をもとに著者らの研究を解説したものである．

2. 方法：高次元統計解析

2.1 高次元統計解析

本節では，高次元統計解析のいくつかの重要な統計理論とその方法論を紹介する．より詳細な解説は例えば 青嶋・矢田 (2013, 2019)を参照のこと．

2.1.1 HDLSS データにおける強不一致性

まず，強不一致性と言われる高次元データ特有の現象を 1 つ例示する．標本 $\vec{x}_i$ $(i = 1, \ldots, n)$ が平均 $\vec{\mu}$ をもつ d 次元母集団から抽出されたとし，標本平均 $\overline{\vec{x}}$

$$\overline{\vec{x}} \equiv \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i \tag{2.1}$$

を考える．

従来の大標本の枠組みである $d/n \longrightarrow 0$ といくつかの条件の下では

$$\left\|\overline{\vec{x}} - \vec{\mu}\right\| \xrightarrow{\mathrm{P}} 0 \tag{2.2}$$

なる一致性をもつ．ここで，$\xrightarrow{\mathrm{P}}$ は確率収束を表す[14]．一方，高次元統計解析の枠組みである $d/n \longrightarrow \infty$ では，

$$\left\|\overline{\vec{x}} - \vec{\mu}\right\| \xrightarrow{\mathrm{P}} \infty \tag{2.3}$$

となる．この性質は，強不一致性と呼ばれる．式(2.3)は，次元 d の増加に伴って観測データのノイズが急激に増大することを意味し，データが HDLSS である場合は従来の統計手法を適用できないことが分かる．

このような現象をより具体的に捉えるために，この巨大なノイズの振る舞いを理論的に検証する．以下でその背景となる数学的概念を紹介する．

2.1.2 標本共分散行列の双対表現

簡単のため，平均ベクトル $\vec{\mu} = \vec{0}$ とし，正定値行列である共分散行列 $\tilde{\Sigma}$ をもつ母集団を考える．共分散行列 $\tilde{\Sigma}$ の固有値を

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \ (> 0) \tag{2.4}$$

とすれば，

$$\tilde{\Sigma} = \tilde{H}\tilde{\Lambda}\tilde{H}^{\top} \tag{2.5}$$

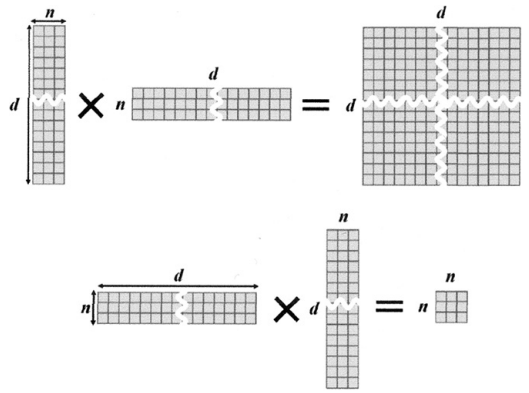

図 1. 標本共分散行列 $\tilde{S}$ とその双対行列 $\tilde{S}_{\rm D}$ の間の関係の概念図．上のパネルは標本共分散行列を，下のパネルはその双対行列を表す．

$$\tilde{\Lambda} \equiv \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_d) \tag{2.6}$$

と固有値分解することができる．ただし，$\top$ は転置を表す．ここで $\tilde{H} = [\vec{h}_1, \ldots, \vec{h}_d]$ はそれら固有値に対応する固有ベクトルで構成された直交行列である．

ここで，この母集団から n 個の標本 $(d > n)$，$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ を抽出し，$(d \times n)$ のデータ行列

$$\tilde{X} = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n) \tag{2.7}$$

で表現する．ただし，$\vec{x}_1, \ldots, \vec{x}_n$ は互いに独立に同分布に従う (i.i.d.)．標本共分散行列 $(d \times d)$ は

$$\tilde{S} = \frac{1}{n}\tilde{X}\tilde{X}^{\top} \tag{2.8}$$

で与えられる．これに対応する双対標本共分散行列 $(n \times n)$ は

$$\tilde{S}_{\rm D} \equiv \frac{1}{n}\tilde{X}^{\top}\tilde{X} \tag{2.9}$$

として定義される (図 1)．この行列 $\tilde{S}_{\rm D}$ の n 個の標本固有値

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n \geq 0 \tag{2.10}$$

は ($d > n$ のもと) $\tilde{S}$ の標本固有値のうち最初の n 個と同一であることに注意する．HDLSS データにおいて $\tilde{S}_{\rm D}$ には以下のような利点がある．

(1) 推測における計算コストの大幅な削減．
(2) 双対空間上での高次元現象の可視化．
(3) 双対空間上での漸近理論の展開．

2.1.3　高次元データの幾何学的表現

高次元データの幾何学的表現を最初に扱ったのは，Hall et al. (2005) である．彼らはその幾何学的表現を分類問題に応用した．さらに，Aoshima and Yata らの一連の研究により，HDLSS データにおける特徴的な幾何学的表現が導出され，高次元のさまざまな統計的推測に応用された[15]．本節では，高次元 PCA に関連する HDLSS の幾何学的表現を 1 つ紹介する．

共分散行列の固有値分解 $\tilde{\Sigma} = \tilde{H}\tilde{\Lambda}\tilde{H}^{\top}$ に基づき，

$$\tilde{X} \equiv \tilde{H}\tilde{\Lambda}^{\frac{1}{2}}\tilde{Z} \tag{2.11}$$

とおく．ただし，$\tilde{Z} = [\vec{z}_1, \ldots, \vec{z}_d]^\top$, $\vec{z}_i = (z_{i1}, \ldots, z_{in})^\top$ $(i = 1, \ldots, d)$ とする．ここで，$\mathbb{E}(z_{ij}z_{i'j}) = 0$ $(i \neq i')$ および $\mathbb{V}(\vec{z}_i) = \tilde{I}_n$ である（$\mathbb{E}$ および $\mathbb{V}$ は期待値および分散，$\tilde{I}_n$ は n 次元単位行列とする）．$\tilde{Z}$ のそれぞれの変数の 4 次モーメントは一様有界であると仮定する．$\tilde{X}$ がガウス分布に従うとすれば，すべての z_{ij} は i.i.d. の標準正規分布に従う確率変数となる．また，$\tilde{S}_{\mathrm{D}}$ の固有値分解を

$$\tilde{S}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\lambda}_i \vec{u}_i \vec{u}_i^\top \tag{2.12}$$

で定義する．ここで $\vec{u}_i$ は $\hat{\lambda}_i$ に対応する（長さ 1 の）固有ベクトルである．$\hat{\vec{h}}_i$ を $\tilde{S}$ の i 番目の（長さ 1 の）固有ベクトルとすると，$\hat{\vec{h}}_i$ は

$$\hat{\vec{h}}_i = (n\hat{\lambda}_i)^{-1/2}\tilde{X}\vec{u}_i \tag{2.13}$$

と計算できる．ここで以下で与えられる条件

$$\frac{\mathrm{tr}\left(\tilde{\Sigma}^2\right)}{\left(\mathrm{tr}\,\tilde{\Sigma}\right)^2} = \frac{\displaystyle\sum_{i=1}^{d}\lambda_i^2}{\left(\displaystyle\sum_{i=1}^{d}\lambda_i\right)^2} \longrightarrow 0 \quad \text{as} \quad d \longrightarrow \infty \tag{2.14}$$

を仮定する．ここで，$\mathrm{tr}(\tilde{\Sigma}^2)/(\mathrm{tr}\,\tilde{\Sigma})^2 \in [1/d, 1)$ となり，$\Sigma = c\tilde{I}_d$ $(c > 0)$ の場合にその最小値 $1/d$ を与える．すなわち，共分散行列が単位行列のような球形に近い場合はその比が小さくなる．なお，(2.14) は球形条件とも呼ばれる．$\tilde{X}$ がガウス分布に従う場合，条件 (2.14) の下で，

$$n\left(\sum_{i=1}^{d}\lambda_i\right)^{-1}\tilde{S}_{\mathrm{D}} \xrightarrow{\mathrm{P}} \tilde{I}_n \quad \text{as} \quad d \longrightarrow \infty \tag{2.15}$$

が得られる（Ahn et al., 2007; Jung and Marron, 2009; Yata and Aoshima, 2012）．ここで

$$\vec{w}_i = n\left(\sum_{i=1}^{d}\lambda_i\right)^{-1}\tilde{S}_{\mathrm{D}}\vec{u}_i = n\left(\sum_{i=1}^{d}\lambda_i\right)^{-1}\hat{\lambda}_i\vec{u}_i \tag{2.16}$$

とおく．Yata and Aoshima (2012) はガウス分布を緩めた適当な条件で，n 次元ベクトル $\vec{w}_i$ について

$$\vec{w}_i \xrightarrow{\mathrm{P}} 1 \quad \text{as} \quad d \longrightarrow \infty \ \ (i = 1, \ldots, n) \tag{2.17}$$

を示した．すなわち，データの双対空間における固有地・固有ベクトルが単位 n 次元球表面に集中する「球面集中現象」である．式 (2.17) で示される幾何学的表現により，標本固有値が一定に定まることが見て取れる．これは，$\tilde{S}_{\mathrm{D}}$（あるいは $\tilde{S}$）を用いた従来型の推定量では固有値の推定が困難であることを意味する．一方，Yata and Aoshima (2012) は，データが非ガウス分布のもとで，まったく異なる幾何学的表現（座標軸集中現象）を与えている．詳細は，Yata and Aoshima (2012) や 青嶋・矢田 (2019) を参照のこと．

これで双対空間におけるノイズの幾何学的挙動を調べる準備が整った．平均 $\vec{\mu} = \vec{0}$ と共分散行列 $\tilde{\Sigma}_d = \tilde{I}_d$ の d 次元ガウス分布から抽出された標本を考える．この分布から $n = 3$ の標本による $\vec{w}_i$ をいくつか（独立に）生成すると，図 2 に示される幾何学的表現が得られる．次元 d が

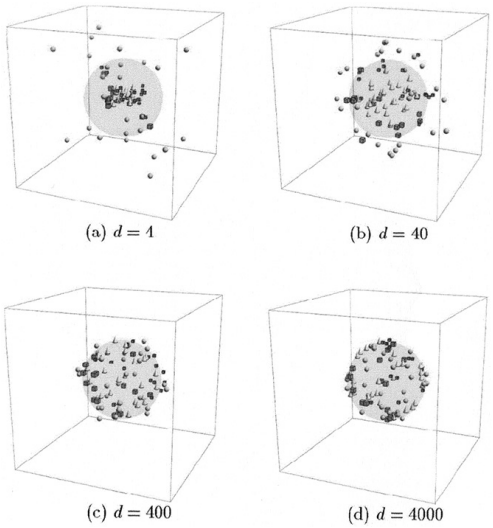

図 2. 式(2.17)で示される幾何学的表現.

低い場合，直感からそれほど外れた形状をしていない，つまり $\vec{w}_i$ は原点に集まり，ある程度散らばっている．一方より高い次元 d の場合，$\vec{w}_i$ は半径 $(d/n)^{\frac{1}{2}}$ の球に集中する．これは，高次元データになるにつれ，データの本質的な特徴が巨大なノイズ球に集まることを明示している．この幾何学的表現が，高次元データ特有の重要な性質である．

2.2 高次元主成分分析

前節と同様に，平均 $\vec{\mu} = \vec{0}$, 共分散行列 $\tilde{\Sigma}_d$ を持つ母集団を考える．$\tilde{\Sigma}_d$ の固有値について，Yata and Aoshima (2009)で導入された「一般化スパイクモデル」と呼ばれる次のモデルを仮定する[16]．

$$\lambda_i = \begin{cases} a_i d^{\alpha_i} & \text{for} \quad i = 1, \ldots, m \\ c_i & \text{for} \quad i = m+1, \ldots, d \end{cases} \tag{2.18}$$

ここで $a_i(>0)$, $c_i(>0)$, α_i $(\alpha_1 \geq \cdots \geq \alpha_m > 0)$ は未知の定数，m は未知の正の整数であり，$\lambda_1 \geq \cdots \geq \lambda_d$ を満たすとする[17]．固有値 λ_i $(i \leq m)$ は潜在的な情報を表し，λ_i $(i \geq m+1)$ はノイズを表している．ここで，

$$\frac{\sum_{i=m+1}^{d} \lambda_i^2}{\left(\sum_{i=m+1}^{d} \lambda_i\right)^2} \longrightarrow 0 \quad \text{as} \quad d \longrightarrow \infty \tag{2.19}$$

が成立する．つまりノイズの空間は球形条件 [式(2.14)] を満たす．

2.2.1 従来型 PCA の限界

次元 d に依存する標本数を $n(d) = d^{\gamma}$ $(\gamma > 0)$ の形で定義しておく．Yata and Aoshima (2009)は，HDLSS データに適用した場合における従来型 PCA を漸近的に評価し，従来型 PCA による推定が一致性を持つための条件を求め，従来型 PCA の限界を示した[18]．

彼らは，以下の条件

$$z_{ij} \quad (i = 1, \ldots, d,\ j = 1, \ldots, n) \text{ は独立} \tag{2.20}$$

を満たす場合と満たさない場合について，一般化スパイクモデル(2.18)のもと固有値に関する次の 2 つの定理を得た[19]．

定理 1. [Yata and Aoshima (2009) I] $\tilde{Z}$ の成分が条件(2.20)を満たすならば，$i \leq m$ について
(1) $\alpha_i > 1$ を満たす i について $d \longrightarrow \infty$ かつ $n \longrightarrow \infty$
(2) $\alpha_i \in (0, 1]$ を満たす i について $d \longrightarrow \infty$ かつ $d^{1-\alpha_i}/n(d) \longrightarrow 0$
の条件のもとで

$$\frac{\hat{\lambda}_i}{\lambda_i} \xrightarrow{\mathrm{P}} 1 \tag{2.21}$$

が成り立つ．

定理 2. [Yata and Aoshima (2009) II] $\tilde{Z}$ の成分が条件(2.20)を満たさないならば，$i \leq m$ について
(1) $\alpha_i > 1$ を満たす i について $d \longrightarrow \infty$ かつ $n \longrightarrow \infty$
(2) $\alpha_i \in (0, 1]$ を満たす i について $d \longrightarrow \infty$ かつ $d^{2-2\alpha_i}/n(d) \longrightarrow 0$
の条件のもとで式(2.21)が成り立つ．

定理 1 と 2 は，従来型 PCA による固有値推定が一致性を持つための条件を与える．これらの定理において，(1)の $d \longrightarrow \infty$ と $n \longrightarrow \infty$ で記述された条件は，n は d に依存していない．すなわち，その n が d よりもはるかに小さい HDLSS の場合も含む．しかし(2)では，特に $\tilde{Z}$ の成分が条件(2.20)を満たさない場合，n は d に強く依存している．したがって，従来型 PCA を HDLSS データに適用する場合は細心の注意を払う必要がある．この従来型 PCA の高次元における問題を克服するため，Yata and Aoshima (2010, 2012)は HDLSS データのための 2 つの新しい PCA を提案した．本論文ではそのうちの 1 つであるノイズ掃出し法 (Yata and Aoshima, 2012)に焦点を当てる[20]．

2.2.2 ノイズ掃出し PCA(NRPCA)：HDLSS データのための新しい方法

ここで高次元 PCA を導入する．図 3 は高次元 PCA のイメージを説明している．HDLSS データは，図 2 に見られるような球面に集中する巨大なノイズを持つため，データが持つ潜在的に重要な情報(潜在空間)はノイズ球に完全に埋もれてしまい，隠されてしまっている．しかし，実はこの状況下でもデータの特性を推定することは不可能ではない．これは，ノイズの挙動が球面近傍で特定され，それを差し引くことができるという重要な事実による．その手法が，Yata and Aoshima (2012)によって提案されたノイズ掃出し(NR)法である．本論文ではノイズ掃出し PCA(noise-reduction PCA: NRPCA)と呼ぶ．

以下でもう少し厳密な説明を試みる．Yata and Aoshima (2012)は，幾何学的表現 [式(2.15)] から派生した NR 法を提案した．まず $\tilde{S}_{\mathrm{D}}$ を次のように分解する．

$$n\tilde{S}_{\mathrm{D}} = \sum_{i=1}^{m} \lambda_i \vec{z}_i \vec{z}_i^{\top} + \sum_{i=m+1}^{d} \lambda_i \vec{z}_i \vec{z}_i^{\top} \ . \tag{2.22}$$

第 2 項はノイズを表現しており，条件(2.20)のもとで式(2.15) と同様に，次のようなノイズの幾何学的表現が得られる．

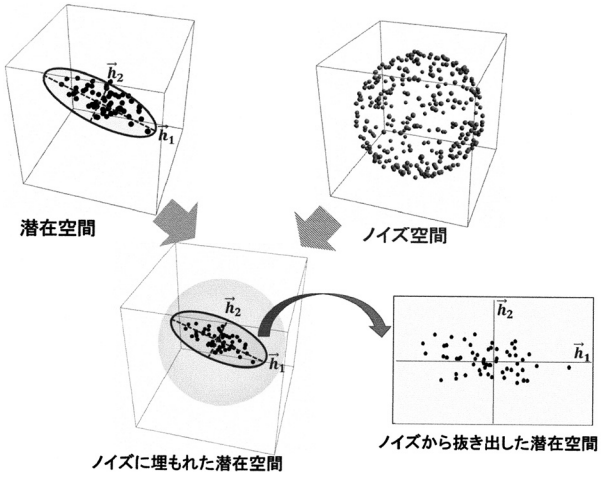

図 3. 高次元主成分分析(PCA)の概略．HDLSS データでは，巨大なノイズをもつため，潜在空間がノイズに埋もれてしまう．高次元 PCA を用いることで，その潜在空間の抽出が可能となる．

$$\frac{\sum_{i=m+1}^{d} \lambda_j \vec{z}_i \vec{z}_i^{\top}}{\sum_{i=m+1}^{d} \lambda_j} \xrightarrow{\mathrm{P}} \tilde{I}_n \quad \text{as} \quad d \longrightarrow \infty \,. \tag{2.23}$$

つまり，一般化スパイクモデルにおけるノイズ空間は，球面集中現象をもつことがわかる．この結果を用いてノイズ部分を除去した固有値の推定量

$$\check{\lambda}_i = \hat{\lambda}_i - \frac{1}{n-i}\left(\mathrm{tr}\,\tilde{S}_{\mathrm{D}} - \sum_{j=1}^{i} \hat{\lambda}_i\right) \qquad (i = 1, \ldots, n-1) \tag{2.24}$$

が得られる．ここで，以下の定理が得られている．

定理 3. [Yata and Aoshima (2012) (I)]　$\tilde{Z}$ の成分が条件(2.20)を満たすとき，$i \leq m$ について
(1)$\alpha_i > 1/2$ を満たす i に対し，$d \longrightarrow \infty$ かつ $n \longrightarrow \infty$
(2)$\alpha_i \in (0, 1/2]$ を満たす i に対し，$d \longrightarrow \infty$ かつ $d^{1-2\alpha_i}/n(d) \longrightarrow 0$
の条件のもとで

$$\frac{\check{\lambda}_i}{\lambda_i} \xrightarrow{\mathrm{P}} 1 \tag{2.25}$$

が成り立つ．

定理 3 と 1 を比較すると，(条件(2.20)の下で)$\check{\lambda}_i$ の一致性は $\hat{\lambda}$ よりも緩い条件で成立することが分かる．

次にノイズ掃出し法を主成分ベクトル推定に適用する．$\check{\vec{h}}_i \equiv (n\check{\lambda}_i)^{-1/2}\tilde{X}\vec{u}_i$ を定義し，その主成分ベクトル $\vec{h}_i$ の推定量としての性能を検証する．ここで $\check{\vec{h}}_i = (\hat{\lambda}_i/\check{\lambda}_i)^{-1/2}\hat{\vec{h}}_i$ であることに注意する．Yata and Aoshima (2012)は次の定理を示した．

定理 4. [Yata and Aoshima (2012) II] λ_i $(i \leq m)$ は単根とする．定理 3 の条件の下で，

$$\check{\vec{h}}_i^{\top} \vec{h}_i \xrightarrow{\mathrm{P}} 1 \tag{2.26}$$

が成り立つ．

また，主成分スコアについても評価する．x_i の i 番目主成分スコアは $\vec{h}_i^{\top} \vec{x}_j = \vec{z}_{ij} (\lambda_i)^{\frac{1}{2}} \equiv \vec{s}_{ij}$ で与えられる．$\vec{u}_i \equiv (\hat{u}_{i1}, \ldots, \hat{u}_{in})^{\top}$ $(i = 1, \ldots, d)$ とすると，主成分スコアは $\hat{\vec{u}}_{ij} (n \check{\lambda}_i)^{\frac{1}{2}} \equiv \check{\vec{s}}_{ij}$ と書ける．ここで，i 番目主成分スコアの平均二乗誤差(MSE)を

$$\mathrm{MSE}(\check{\vec{s}}_i) \equiv \frac{1}{n} \sum_{j=1}^{n} (\check{s}_{ij} - s_{ij})^2 \tag{2.27}$$

と定義すると，もう 1 つの重要な定理を得る．

定理 5. [Yata and Aoshima (2012) III] λ_i $(i \leq m)$ は単根とする．定理 3 の条件の下で，

$$\frac{\mathrm{MSE}(\check{s}_i)}{\lambda_i} \xrightarrow{\mathrm{P}} 0 \tag{2.28}$$

が成り立つ．

ここまでデータは平均 0 としてきたが，(中心化することにより)与えた定理はすべて平均が 0 でない場合にも適用できることに注意する．

このように，HDLSS データの固有値(寄与率)，主成分ベクトル，および主成分スコアの推定において，NRPCA が高いパフォーマンスを持つことが分かる．以降，固有値推定には NRPCA で得られた値を用いる．

2.2.3 オートマティック・スパース PCA(A-SPCA)

さらに，主成分ベクトルをより精度良く推定するために，Yata and Aoshima (2022)で提案されたオートマティック・スパース PCA (automatic sparse PCA: A-SPCA)を導入する．すべての i に対して $\hat{\vec{h}}_i = (\hat{h}_{i(1)}, \ldots, \hat{h}_{i(d)})^{\top}$ とする．閾値 $\zeta_i > 0$ が与えられたとき，

$$\hat{h}_{i*(s)} = \begin{cases} \hat{h}_{i(s)} & \text{if } |\hat{h}_{i(s)}| \geq \zeta_i, \\ 0 & \text{if } |\hat{h}_{i(s)}| < \zeta_i \end{cases} \quad \text{for } s = 1, \ldots, d \tag{2.29}$$

とおく．$\hat{\vec{h}}_{i*} = (\hat{h}_{i*(1)}, \ldots, \hat{h}_{i*(d)})^{\top}$ とすると，$\vec{h}_i$ の閾値に基づく推定量は次のように定義される．

$$\hat{\vec{h}}_{i**} = \hat{h}_{i*} / \|\hat{h}_{i*}\|. \tag{2.30}$$

しかし，この推定量は ζ_i に大きく依存している．

最近，青嶋と矢田は NR 法を拡張した A-SPCA を次のように提案した (Yata and Aoshima, 2022)．NR 法による主成分ベクトル $\check{\vec{h}}_i = (\check{h}_{i(1)}, \ldots, \tilde{h}_{i(d)})^{\top}$ について，簡単のため $|\check{h}_{i(1)}| \geq \cdots \geq |\check{h}_{i(d)}|$ と仮定する．与えられた定数 $\omega \in (0, 1]$ に対して，成分の累積寄与率が ω より大きくなる主成分ベクトルを考える．次を満たすある整数 $\check{k}_{i\omega} \in [1, d]$ が一意に存在する．

$$\sum_{s=1}^{\check{k}_{i\omega}-1} \check{h}_{i(s)}^2 < \omega \quad \text{および} \quad \sum_{s=1}^{\check{k}_{i\omega}} \check{h}_{i(s)}^2 \geq \omega. \tag{2.31}$$

さらに，

$$\check{h}_{i\omega(s)} = \begin{cases} \check{h}_{i(s)} & \text{if } |\check{h}_{i(s)}| \geq |\check{h}_{i(\check{k}_{i\omega})}|, \\ 0 & \text{otherwise} \end{cases} \quad \text{for } s = 1, \ldots, d.$$

を定義する．$\check{\tilde{h}}_{i\omega} = (\check{h}_{i\omega(1)}, \ldots, \check{h}_{i\omega(d)})^\top$ とする．Yata and Aoshima (2022)は HDLSS データで $\check{\tilde{h}}_{i\omega}$ のいくつかの一致性に関する性質を示し，実際のデータ解析でそのパフォーマンスを検証した．本論文では $\omega = 1/2$ とする[21]．A-SPCA は，NRPCA によって構築された主成分ベクトルの最も重要な成分を抽出でき，それらの成分が実質的に主成分の方向を決めているといえる．

3. データ

本研究の対象銀河 NGC 253 は，3.5 Mpc(1″ は 17 pc に対応)の距離にある近傍棒渦巻銀河であり（Rekola et al., 2005），形態は Sc とされている．この銀河の太陽中心速度[22]は 243 ± 2 km s^{-1} ($z = 0.000811$: Kamphuis et al. 2015) で，主に宇宙膨張による．この銀河は大小マゼラン銀河を除いて最も明るい遠赤外線放射源の 1 つであり，多くの波長で広く研究されてきた．NGC 253 は純粋なスターバーストとして知られており，中心の分子領域での星形成率 (SFR) は SFR $\sim 2\ M_\odot$ [yr^{-1}] と推定されている(Rieke et al., 1980; Keto et al., 1999; Bendo et al., 2015)[23]．中心には大質量，高密度で暖かい ISM の存在が確認されている（e.g. Sakamoto et al., 2011）．さらに，スーパースタークラスター[24]も可視光および近赤外線(NIR)で観測されている（Fernández-Ontiveros et al., 2009）．非常に強いアウトフローもこの領域で観測されている（Matsubayashi et al., 2009; Bolatto et al., 2013）．スーパースタークラスター，アウトフローともスターバースト活動に密接に関連している．

3.1 NGC 253 の ALMA 分光マップ

本研究では，Ando et al. (2017)によって取得された，近傍の典型的スターバースト銀河 NGC 253 の ALMA Band 7 での分光マップを用いる．最終的な画像の合成ビームサイズは $0.''45 \times 0.''3$ で，これは NGC 253 の距離において 8 pc × 5 pc に相当する．スペクトルにはさまざまな輝線が検出されている（Ando et al., 2017）．振動数方向のサンプリング数は 2248 である．

Ando et al. (2017) の元マップは，空間領域(赤経および赤緯)で 864×864 点，振動数領域で 2248 成分(単位 [Hz])で構成される．ただし，ピクセルスケールはマップの合成ビームサイズ $0.''45 \times 0.''3$ よりも小さいためオーバーサンプリングであり，各空間グリッド点は互いに独立になっていない．この問題を克服するため，$0.''45 \times 0.''3$ の楕円ビームに対応する円形ビームサイズは，$0.''37$ を使用して，イメージマップをリサンプリングした．得られた独立なグリッドからなるマップは 230×230 ピクセルで構成される．また信号の微弱な領域を除外する必要がある．このため振動数の全範囲にわたって輻射強度強度を積分し，強度 I [Jy beam^{-1}] > 2.0 の領域を選択した．図 4 は有意な信号が検出された明るい領域および「マスク」領域をあらわしている．信号のあるピクセルの数 n は 231 となった．

3.2 宇宙膨張による赤方偏移および系の回転によるドップラーシフト

NGC 253 の中心部は大きく傾いている ($i \sim 78°$)．これにより，ドップラーシフトによって分光マップ上にかなりコヒーレントな回転パターンが生じる．この効果を考慮するため，スペクトル上の 3 本の顕著な輝線 HCN(4–3)，HNC(4–3)，および CS(7–6) にガウス関数フィットを行い，輝線の中心波長を推定した．速度場のマップを図 5 に示す．これら 3 本の輝線から得られたドップラーシフトを平均することにより，シフトの大きさを決定した．次に，宇宙膨張に

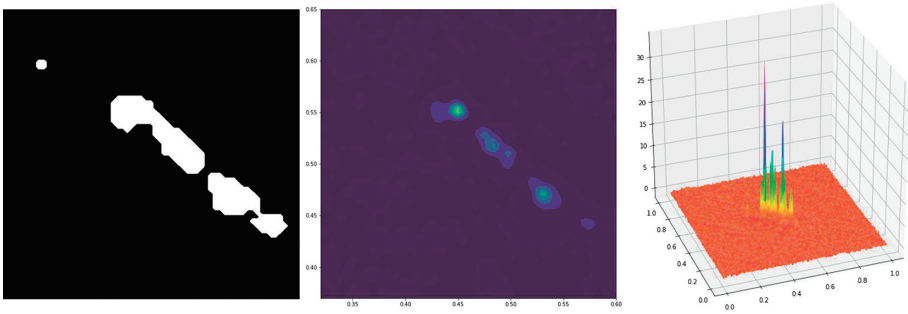

図 4. マスクで抽出した NGC 253 マップの明るい領域（Takeuchi et al., 2024）．左：マスク領域マップ．白抜きは有意な強度の信号がきている領域．中央：信号の強度マップ．右：信号の強度マップの鳥瞰図．

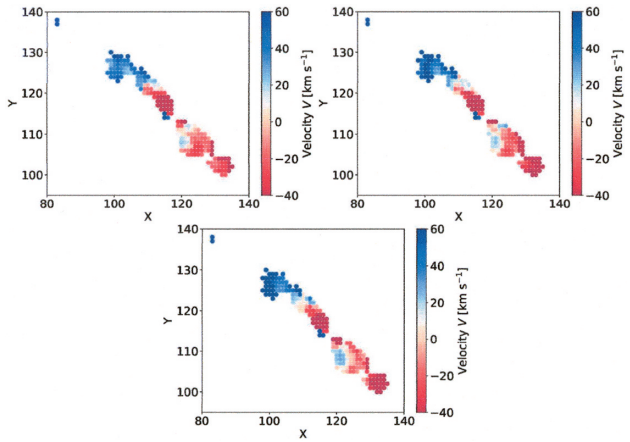

図 5. HCN(4–3)（左上），HNC(4–3)（右上），および CS(7–6)（下）輝線から推定した NGC 253 の速度マップ（Takeuchi et al., 2024）．この銀河の系全体の速度 243 [km s^{-1}] を差し引いた値を示す．

よって引き起こされる NGC 253 の系全体の速度[25]に関して振動数をシフトした．この補正によってコヒーレントな回転の系統的影響のないマップが得られ，サブミリ波輝線のより詳細な特性を解析できる．ドップラーシフトについては付録 B に詳しく述べている．

データキューブ[26]には振動数にギャップがある．速度に応じてスペクトルをシフトすると，スペクトルの境界もシフトするため，本研究では単純にすべてのピクセルに共通の振動数範囲を採用した．データの最終的なサイズは，空間位置に沿って 231, 補正後のスペクトル要素数は 1971 である．

3.3 HDLSS データとしての分光マップ

上記のように，NGC 253 のマップは 空間次元 231 × スペクトル次元 1971, つまり $n = 231$ および $d = 1971$ であり，明らかに $n \ll d$ の条件を満たしている．DNA マイクロアレイデータ（通常，$n \sim 10$–100 および $d \sim 10^4$–10^5）ほど極端ではないものの，この ALMA マップは典型的な HDLSS データである．天体物理学の観点から問題となるのは，スペクトルに非常に多く

の情報が含まれていること，そして n に比べて多種多様なスペクトル輝線が観測されることである．しかし，これは高次元統計解析においては特別な状況ではない．

ただし，分光データへの高次元統計の適用は全く単純というわけにはいかない．スペクトル線は，ISM の乱流運動や他の物理機構によって拡がったりシフトしたりする．その場合，各スペクトル分解能ユニットが持つ情報は独立にはならない[27]．この複雑な問題は，HDLSS データを扱う他の研究分野で調べられたことはない．しかし，天文学のあらゆる分光データに共通して見られる問題であるため，まず高次元統計解析が実際に輝線のシフトと広がりのある天体分光マップに対して機能するかどうかを本格的な解析の前に検証する．

4. 予備解析

すでに述べたように，ゲノム解析やその他の分野における HDLSS データの典型的な例とは異なり，天文分光データには特有の潜在的な問題が存在する．分光データをベクトルデータ $\vec{x} = (x_1, \ldots, x_d)^\top$ と考えたとき，根本的な問題は隣接する振動数(または波長)指標 i と $i+1$ $(i = 1, \ldots, d-1)$ における情報が独立ではないことである．つまり，銀河の ISM には内部運動があり，ドップラー効果によって振動数方向に広がる．また，天体の観測領域全体にわたってシステムの系統的な回転が見られる(3 章)．高次元統計解析がこのようなタイプの分光データに対して機能するかどうかを調べるため，高次元 PCA を用いて NGC 253 の元の分光マップデータの予備解析を行った．「元の」という言葉は，系全体の回転によるマップ上のドップラーシフトを補正せずにこの解析をしたことを意味する．すなわち，高次元 PCA をそのまま元データに適用し，どのような特徴が検出されるかを検証した．このため画像の 2 次元座標を赤経方向，次に赤緯方向に沿って並べ替え，データを 1 次元ベクトルと見なして指標を付けなおした．PCA を 2 次元画像に直接適用することも可能だが，本論文では解析を単純にするため 1 次元化する方針を採用した．

4.1 固有スペクトル(eigenspectra)

次の段階の解析では，主として A-SPCA の結果に注目する．ただし，NRPCA は全てのスペクトル特徴を対象として機能するので，NRPCA で求めた固有スペクトル (PCA によって観測されたスペクトルから構築された固有ベクトル)も示す．図 6 の上のパネルは NGC 253 の ALMA 分光マップから NRPCA によって構成された固有スペクトルを示している．この図では，すべての PC に対応する固有スペクトルにおいて，同じ振動数範囲で同一のスペクトルの特徴が現れていることが見て取れる．これは，一部のスペクトルの特徴が，他のスペクトルの領域よりもスペクトル全体を決定する重要な情報を含んでいることを示唆する．

オートマティック・スパース PCA(A-SPCA)は，これらの特定のスペクトルの特徴に対応する最も重要な要素を選び出し，局在化させる機能を持つ．図 6 の下のパネルは，A-SPCA によって得られた固有スペクトルを示す．図 6 からわかるように，A-SPCA はある特徴が全体のスペクトルの決定に重要となる場合のみ値を残し，他の成分を強制的に 0 にする．したがって，A-SPCA の固有スペクトルはすべての特徴を含むスペクトルの再構成には用いず，完全なスペクトルの再構成には NRPCA を用いる．A-SPCA のこの利点を利用し，次の解析でスペクトル全体を決める特徴を特定する．

4.2 PC の寄与率分布

まず NRPCA によって得られた寄与度の大きさ順にソートした PC の固有値分布を図 7 に示す．最初のいくつかの固有値は，残りの固有値よりも著しく大きい．この振舞いは，高次元統

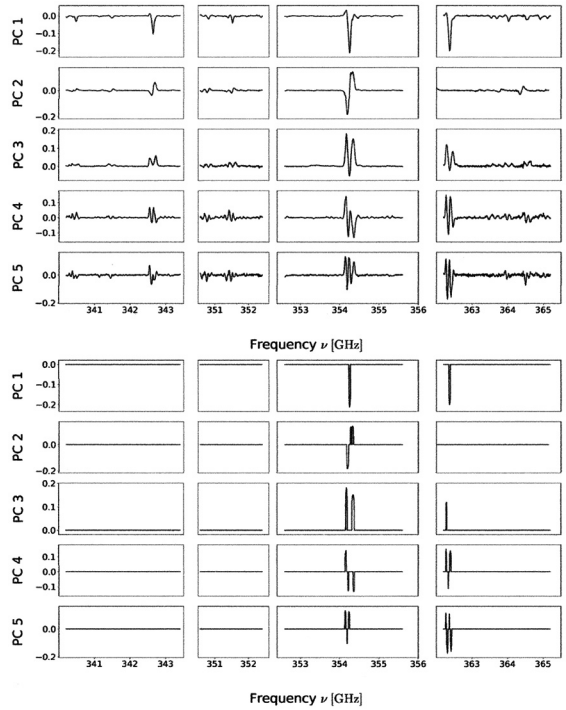

図 6. NGC 253 の ALMA 分光マップから作成された第 1 から第 5 主成分(PC)に対応する固有スペクトル (Takeuchi et al., 2024)．上段：ノイズ掃出し PCA(NRPCA)により得られた固有スペクトル．下段：オートマティック・スパース PCA(A-SPCA)によって得られた固有スペクトル．PC を決定付ける成分のみが非ゼロの値を持つ．NRPCA と通常の PCA の結果は，$\check{\vec{h}}_i = (\hat{\lambda}_i/\check{\lambda}_i)^{-1/2}\hat{\vec{h}}_i$ (p.11) より，形状は同じでスケールが違うだけの図となるためここでは示していない．

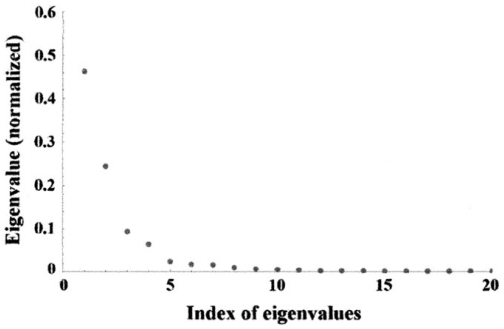

図 7. NRPCA で得られた NGC 253 の ALMA 分光マップの規格化した固有値(寄与率) (Takeuchi et al., 2024)．

計解析の文脈では「スパイク(spike)」と呼ばれる．PCA の固有値は各 PC の寄与を表しているので，PCA は共分散行列全体を各 PC からの寄与に分解することが分かる．得られた固有値分布がこのスパイク構造を示していることから，高次元 PCA が NGC 253 の ALMA 分光マップのいくつかの特徴を検出したことが理解できる．すなわち，本研究のデータが通常の HDLSS

データとみなせ，高次元 PCA の前提が成り立つことが保証されている．

天体物理学的観点からは，図 7 は分光マップデータに対する高次元統計解析の極めて有望な可能性を示している．Ando et al. (2017)は，マップ上で 30 を超える分子ないしラジカル輝線(Ando et al., 2017 の図 1 を参照)，およびいくつかの広がったスペクトル放射を特定した．たとえ興味の対象を輝線に限定したとしても，物理的直感だけで輝線を分類することは非現実的である．このような分光データから物理情報を抽出するために，伝統的には輝線比が使用されてきた．しかし，輝線の数が増えるにしたがって，輝線の組み合わせ数は爆発的に増加する(組み合わせ論的爆発)ため，力業で多数の輝線比を一度に診断することは全く現実的ではない．これと対照的に，図 7 に示されているように，複雑な分光マップの全情報は最初の数個の PC で表現できることがわかる．これは，30 本を超える輝線を含む複雑なスペクトルの特徴が，ごく少数の PC に対応するベクトル基底(固有スペクトル：eigenspectra)て決まっていることを意味する．これは PCA を含む，一般に次元削減として知られる方法を用いることのよく知られた利点である．(e.g., Galaz and de Lapparent, 1998; Ronen et al., 1999; Wang et al., 2011; Pace et al., 2019; Portillo et al., 2020)，しかし，これら先行研究はすべて，十分に大きな標本数 n についての 1 次元積分スペクトルへの応用にとどまり，HDLSS データに適用したものではなかった．本論文の解析は，これら先行研究の解析とは全く異なることを強調しておく．ここで示す方法は，第 2 章で述べたように，従来型 PCA では巨大なノイズ球の存在のため歯が立たなかった典型的 HDLSS データである分光マップに問題なく応用できる．

このように，高次元統計解析の方法が天体分光マップデータに対してうまく機能することを確認できた．これに基づき，高次元 PCA によるさらなる解析に問題なく進むことができる．

4.3 第 1 および第 2 主成分の物理的意味

得られた第 1 主成分(PC1)および第 2 主成分(PC2)が何を表すかを考察するのが次の段階である．図 8 に PC1 および PC2 の散布図を示す．最も顕著な特徴は，x 軸に関してほぼ対称な蝶の羽根のようなパターンである．これは，NGC 253 中心領域のコヒーレントな回転を連想させる．PC1 と 2 の 2 次元構造を再構築したマップを図 9 に示す．

明らかに，PC1 は輝線強度を表し，PC2 は NGC 253 中心の対称でコヒーレントなパターンを表している．実際，図 9 の右パネルの PC2 マップと，同じ領域で直接測定した視線方向の速度場マップ(図 5)は非常によく一致している．視線方向の速度場は，運動する物質から放射される輝線の振動数がドップラーシフトしていることから，振動数のシフト量を特殊相対論の式を用いて速度に換算することができる．ドップラーシフトについては付録 B で詳しく述べている．

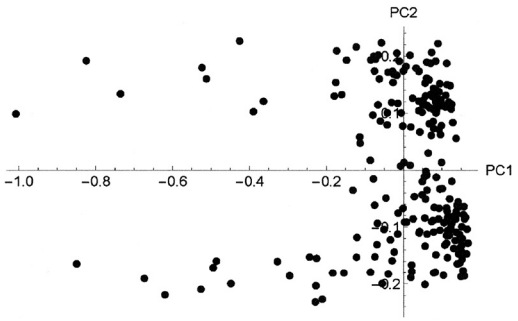

図 8. NGC 253 の ALMA 分光マップから A-SPCA によって得られた PC1 と PC2 の分布 (Takeuchi et al., 2024)．

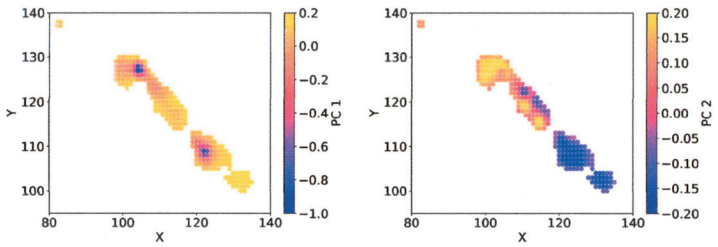

図 9. A-SPCA によって得られた PC1 と PC2 のマップ上での 2 次元分布．左パネルは第 1 主成分(PC1)，右パネルは第 2 主成分(PC2)のマップをそれぞれ示している．

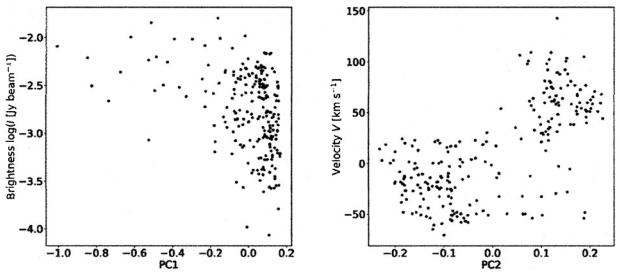

図 10. 主成分と観測された物理的特性の関係（Takeuchi et al., 2024）．左：NGC 253 の ALMA マップ上での PC1 と振動数全域での積分強度 I [Jy beam^{-1}] との相関．右：PC2 と視線速度の相関．

この関係をさらに詳しく調べるため，PC と観測された特性の相関を図 10 に示す．図 10 の左パネルは，PC1 と各ピクセルにおける振動数範囲全体にわたって積分した全強度との間の散布図である．これらの間にはよい相関が見られる．PCA はデータの線型な特徴のみを抽出するのに対し，PC と物理量との関係は必ずしも線型ではないため，この相関は必ずしもきれいな線型相関とはならない．図 10 の右パネルは固有速度と PC2 の関係を示す．すでに 2 次元マップで示したように，PC2 は図 10 で視線速度場を近似的に表している．しかし，PC2 と固有速度が一致しない特異領域が存在することに注目しよう．これらの領域は，強い質量流出など，何らかの局所的な現象の影響を受けている可能性がある．この問題については，後述の詳細な解析を用いて再検討する．

4.4 主成分に対応するスペクトルの特徴：元データの場合

各主成分に具体的にどのような特徴的スペクトルが関与しているかを検証するのは非常に興味深い解析であるが，従来の PCA はこのような目的には不向きであった．一方，A-SPCA は各主成分を決める特徴的スペクトルを特定できる．A-SPCA によって特定された，主成分を決める特徴的スペクトルを図 11 に示す．この図において，スペクトルは銀河の解析領域全域にわたって積分されている．星印は PC1 を決める特徴的スペクトル，三角形は PC2 を決める特徴的スペクトルを示している．図を見やすくするため，高次の主成分の寄与は示していない．PC1 の情報は HCN(4–3) および HNC(4–3) 分子の 2 本の輝線，特に輝線の中心部に集中している(分子輝線放射機構の詳細と表記法については付録 A を参照)．PC2 の情報は HCN(4–3) 輝線と関連するが，PC1 の場合とは対照的に輝線のウィング部に割り当てられている．この図のスペクトルは NGC 253 の解析領域全体にわたって積分されているため，各輝線のウィング部は実際にはドップラーシフトによって静止系波長からずれた輝線が重なることで形作られてい

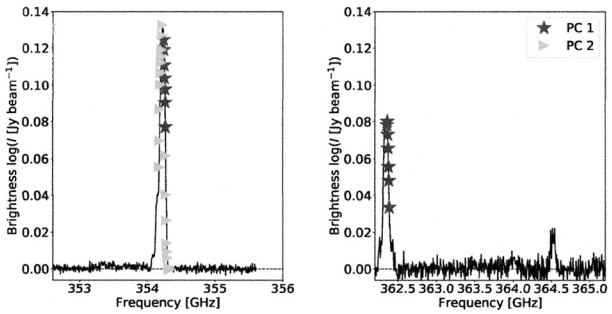

図 11. A-SPCA によって求めた NGC 253 の ALMA 分光マップの PC を特徴付ける特徴的スペクトル．星印は PC1 を決める特徴的スペクトル，三角形は PC2 を決める特徴的スペクトルである．

る．これは，PC2 のマップ(図 9)が NGC 253 のドップラー速度場マップ(図 5)とよく一致するということと整合的である．

このように，高次元 PCA は ALMA 分光マップ上の輝線の特徴抽出において高い能力を示すことが分かった．この予備解析により，HDLSS データとしての ALMA 分光マップに対する A-SPCA のパフォーマンスが保証されたので，次章では問題なく分子輝線の詳細な解析に進むことができる．

5. 本解析および議論

前章(4 章)で高次元統計解析の性能を確認した．ここでは更に詳細な物理的解析に進む．本章では，NGC 253 の分光マップの輝線の振動数から各点ごとに求めたドップラーシフトを補正したデータを用いる．これをドップラーシフト補正済み分光マップと呼ぶことにする．

5.1 NGC 253 のドップラーシフト補正済み分光マップへの高次元 PCA の応用

4 章と同様の方法で，NGC 253 のドップラーシフト補正済みマップに高次元 PCA を適用する．PC2 については 4 章と同じ表記を用いるが，ここでの PC2 はドップラーシフト補正済みマップから得られた値であり，4 章で示した値とは異なることに注意されたい．また，ドップラーシフト補正済みマップに NRPCA および A-SPCA を適用すると，PC のすべての寄与が補正の影響を受けることも注意する必要がある．

予備解析と同様に，まずドップラーシフト補正済み分光マップの固有スペクトルを図 12 に示す．図 6 と比較すると，各固有スペクトルの形状の意味がより明確になっている．スペクトル全体のドップラーシフトを既に差し引いているため，輝線の振動数はもはやコヒーレントな回転の速度を表していないことをここで強調しておく．したがって，速度構造は常に系全体の回転パターンからの 1 次のオーダーの変位であり，輝線プロファイルの広がりまたは非対称な歪みとして現れる．もう 1 つの重要な違いは，元データの解析では特定されなかった重要なスペクトル輝線が PC5 (図 12 の下のパネルの PC5 を参照)で識別できることである．これは，系の回転の圧倒的な寄与がマップから適切に取り除かれ，ドップラーシフト補正済みマップでより微妙なスペクトルの特徴が顕著になったことによる．新しい PC2 は，HCN をはじめいくつかの輝線のウィング[28]を表しており，ガスの非常に局所的な膨張を反映している．PC3 と 4 は，膨張速度よりも速いガスの動きによって生じる輝線の青方偏移を表している．PC5 はそれほど目立たないが，PC3 および 4 と同様に，輝線の赤方偏移を表している．

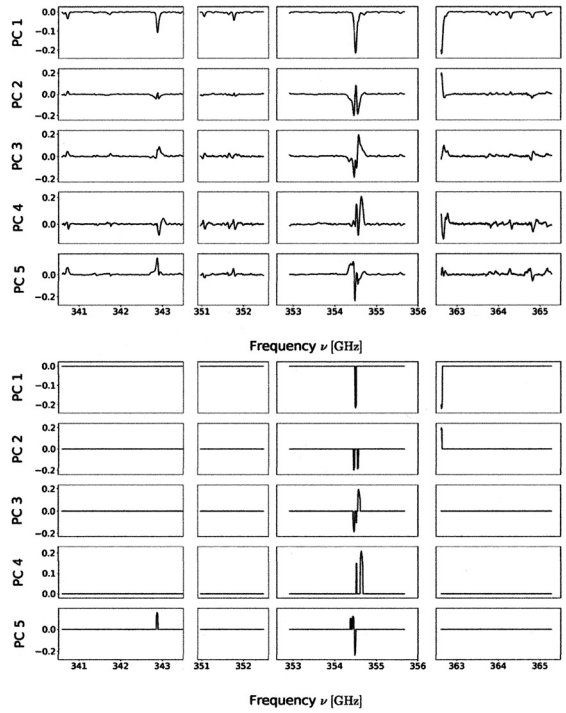

図 12. ドップラー補正済み ALMA 分光マップから高次元 PCA によって構築された第 1–第 5 主成分(PC)に対応する固有スペクトル．上 5 枚のパネルは NRPCA による固有スペクトル，下 5 枚のパネルは A-SPCA で得られた固有スペクトルを示している．

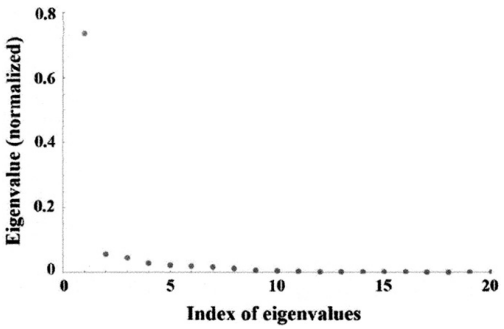

図 13. 系全体の回転によるドップラーシフト補正後の NGC 253 の ALMA 分光マップデータから NRPCA によって得られた固有値寄与率（Takeuchi et al., 2024）．

NRPCA によって得られた固有値の寄与率を図 13 に示す．ドップラーシフト補正済みマップから求められた新しい PC2 の寄与率は，図 7 の PC2 よりもはるかに小さくなっている．これは系統的回転の除去がうまくいっていることによる．そして，新しい PC2 の寄与率は小さいながらも無視できない大きさを持つ．つまり，新しい PC2 は NGC 253 の分光マップにおける，局所的で詳細な特徴を表現している．これは，固有スペクトルからの示唆とも一致している．

このことは，PC の散布図においてより顕著に表れる．各 PC 間の関係を図 14–17 に示す．

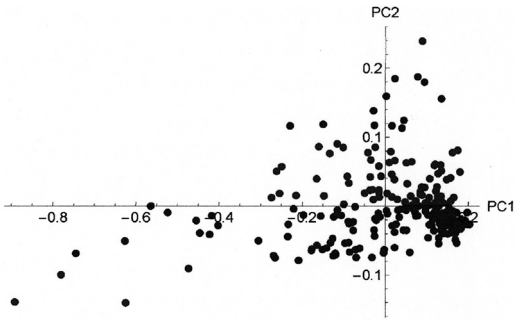

図 14. 系全体の回転によるドップラーシフト補正後の NGC 253 の分光マップから A-SPCA によって求めた PC1 と PC2 の 2 次元分布．この図における PC2 は図 9 の PC2 とは異なることに注意．

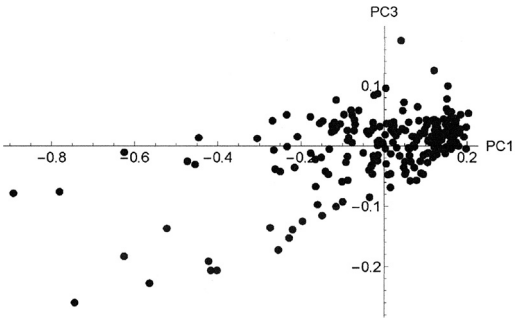

図 15. 図 14 と同様，NGC 253 のドップラーシフト補正済み分光マップから A-SPCA によって求めた PC1 と PC3 の 2 次元分布．

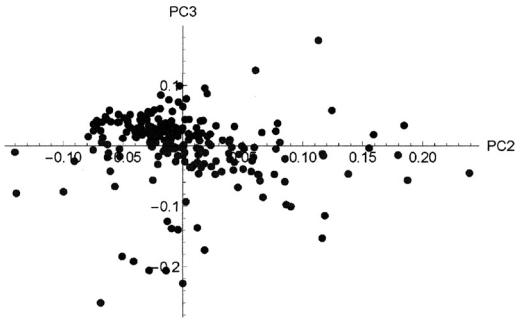

図 16. 図 14 と同様，NGC 253 のドップラーシフト補正済み分光マップから A-SPCA によって求めた PC2 と PC3 の 2 次元分布．

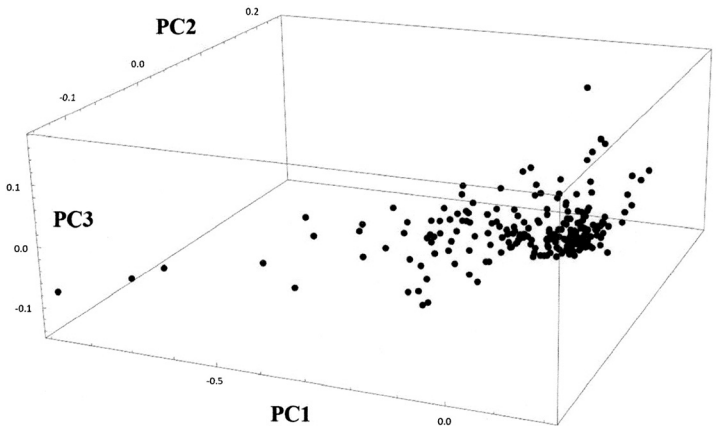

図 17. NGC 253 のドップラーシフト補正済み分光マップから求めた PC1, PC2, PC3 の分布の 3 次元構造.

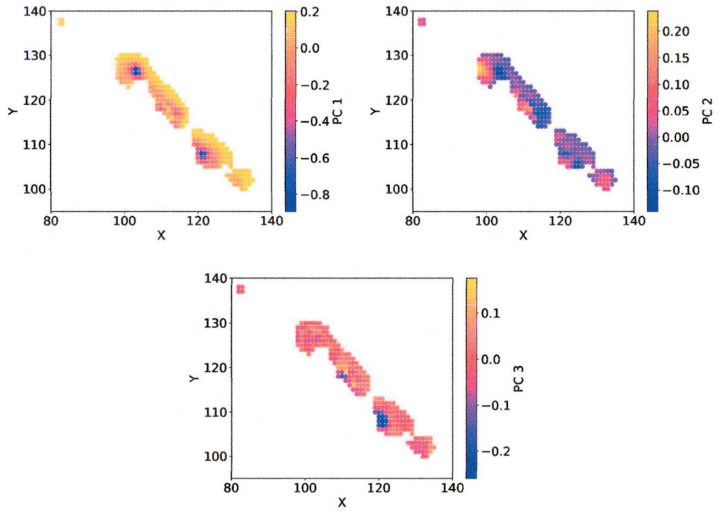

図 18. A-SPCA によって得られた NGC 253 の PC1, PC2, および PC3 の空間構造.

図 14 は，PC1 と PC2 の 2 次元分布を示している．新しい図においては，図 8 に見られる蝶の翅のようなパターンが完全に消失していることがわかる．図 15 および 16 はそれぞれ PC1 と PC3, PC2 と PC3 の散布図である．図 14 と同様，図 15 と 16 に対称なパターンはない．図 17 は PC1, 2, 3 の分布の鳥瞰図である．これらの図から，ドップラーシフト補正済みマップのスペクトルの特徴に明らかな対称パターンはなく，PC 間の散布図は比較的連続的な主要部と外れ値からなることがわかる．

5.2 NGC 253 のスペクトル主成分マップ

NGC 253 の分光マップ上での PC1, 2, 3 の特徴をさらに検証するため，図 18 で各 PC の空間構造[29]を再構成した．この図でも PC1 が輝線強度を表していることが再確認できる．一方ドップラーシフト補正済みマップの PC2 は，図 9 とは対照的に NGC 253 でより複雑で局所

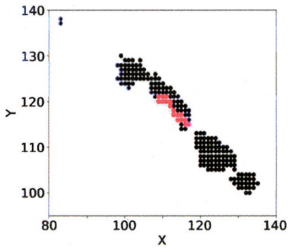

図 19．図 10 マップにおける速度異常がみられる領域．

的なパターンを表現している．PC3 の 2 次元マップには，小スケールの構造も示されている．PC2 が局所的な質量流出を示唆していることを思い出そう．Ando et al.（2017）によると，PC2 の値が最も大きい領域の空間スケール（直径）は ∼ 10 pc であり，確かに空間的に小さいスケールでの現象である．

特に興味深いのは，図 18 の下パネルに見られる PC3 が負の領域である．固有スペクトルは，PC3 および部分的に PC4 が，局所的ではあるがよりスケールの大きいアウトフロー（質量流出）を表しており，輝線の青方偏移[30]に現れていることを示している．Bolatto et al.（2013）は，電波連続波の中心[31]からの分子アウトフローを報告した．最近，Walter et al.（2017）および Krieger et al.（2019）は ALMA によるこの領域の詳細な観測を行った．彼らは詳細な解析により，この質量アウトフローを「SW ストリーマー」と名付けた．このアウトフローは私たちに向かって運動しており，SW ストリーマーの根元の位置は，図 18 に示された PC3 が負の領域と正確に一致している（Krieger et al., 2019, 図 1 を参照）．図 19 の速度異常領域も SW ストリーマーの根本とよく一致する．さらに分子輝線観測だけでなく，ファブリーペロー分光器[32]を用いて Hα[33]やその他の輝線でもアウトフローが観測されている（Matsubayashi et al., 2009）．Hα などの輝線は電離ガスのアウトフローをトレースするが，質量流の位置，形状，方向は分子のアウトフローとかなりよく一致している．これらすべての手がかりは，高次元 PCA が ALMA 分光マップからアウトフロー現象を（物理モデルに依存しないという意味で）純粋に客観的に抽出できたことを示している．

5.3 各 PC に対応するドップラー補正データの特徴的スペクトル

図 20 では，星と三角形は図 11 と同様に PC1 および PC2 に関係する特徴的スペクトルを表し，黒丸は PC3 に関係する特徴である．図 11 と比較すると，まず輝線の系統的なシフトが見られる．もちろん，これは銀河全体の宇宙論的赤方偏移の効果である．ALMA バンド間の境界は観測者の静止系に固定されているため，ドップラーシフト補正後はスペクトルの一部が境界からはみ出す．その結果，図 20 ではスペクトルのかなりの部分が失われている．ドップラー補正後の PC1 は，HCN(4–3) および HNC(4–3) 輝線の強度をより正確に表している．これは，スペクトルの関連する特徴を担う部分がこれらの輝線のピーク付近に集中しているという事実に反映されている．PC2 は HCN(4–3) の輝線中心から少し離れた振動数のプロファイルに割り当てられている．ただし，これは HNC(4–3) 輝線の低周波側に関連しているため，ドップラーシフト補正により低振動数側が切り取られてしまっている．PC3 は HCN(4–3) 輝線の高振動数側のさらに遠いプロファイルと関連するが，PC2 に関連する低振動数側のスペクトルプロファイルとも一部重複している．これらの特性は固有スペクトルから既に示唆されており（図 12），A-SPCA によって得られた詳細な情報で確かめられたといえる．したがって，スペクトルが非常に複雑に見える場合でも，NRPCA や A-SPCA など高次元 PCA を用いれば情報を解読

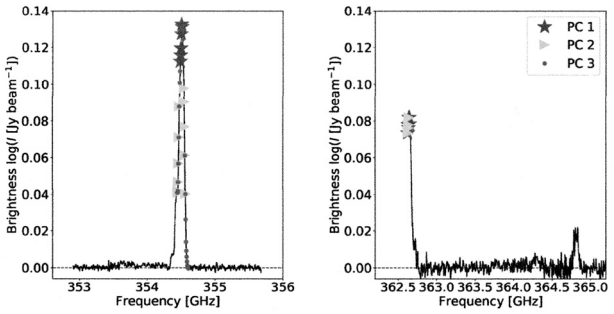

図 20. 系全体の回転によるドップラーシフトについて補正後の ALMA 分光マップから A-SPCA によって求められた，PC1, 2, 3 に対応する特徴的スペクトル．星と三角形は図 11 と同様，PC1 および PC2 に関係する特徴的スペクトルを表し，丸は PC3 に関係する．

し，分光マップを決める特徴的スペクトルを見つけ出すことができる．さらに驚くべきことに，HCN と HNC 輝線の詳細なプロファイルは，アウトフロー現象を含む NGC 253 の ALMA 分光マップのほぼすべてのスペクトル特性を支配している．NGC 253 の ALMA 分光マップによるさらに詳細な天体物理学的素過程のさらなる検証は，今後の研究で行う予定である．

6. まとめと結論

銀河進化は主に星形成によって駆動される．星形成とそれに続く元素合成によって星間物質が進化するので，これが銀河進化を理解する鍵となる．分光観測はこの過程を明らかにする重要な天文学的手法だが，詳細な分光マッピングは多大な時間がかかるため，多くの観測点で独立したデータを得るのは難しい．その結果，分光マップデータは高次元小標本(HDLSS)となる．HDLSS データには高次元特有の困難があり，その解析には古典的方法に代わる新しい統計手法が必要となる．この問題を解決するため，高次元統計解析が発展してきた．

HDLSS データの解析には「次元の呪い」に起因する多くの問題が存在し，古典的統計手法は往々にして正しくない結果を導く．高次元統計解析は双対行列とその幾何学的表現を用いてこの問題を克服し，HDLSS データの特異な振舞いを利用して解析を行う枠組みを提供する．我々は NGC 253 の ALMA 分光マップデータに対して高次元 PCA を適用し，スペクトルマップを特徴づけるスペクトルの特徴成分を特定した．高次元 PCA はモデルを仮定することなく，NGC 253 中心部の系全体の回転を抽出することに成功した．またこうして得られた回転のドップラー効果を補正したデータに高次元 PCA を用いることで，爆発的星形成に起因するガスのアウトフロー現象や速度場異常を示す固有スペクトルが得られた．

これらの結果は，物理的直観を頼りに手作業で特徴を選択するのではなく，高次元データの完全な情報を利用することによって得られたものである．このように，高次元統計解析は将来の宇宙望遠鏡の分光マップ解析に新たな方法論を提供する．また，たとえば極めて稀な天体の分光データを分類する目的においても有用である．別の例として，活動銀河核からの電波連続波を銀河間物質の水素原子が吸収することで生じる HI の森(HI forest)の SKA による観測が挙げられる．我々は希少天体のスペクトルや銀河間物質の吸収線系データに高次元統計解析を適用し，銀河形成の過程解明を目指している．このように，高次元の統計学は天文学の幅広い課題に対して全く新しい扉を開くことになるであろう．

注.

[1] https://www.sdss.org/surveys/.

[2] https://www.desi.lbl.gov/.

[3] https://www.cosmos.esa.int/web/gaia.

[4] https://panstarrs.ifa.hawaii.edu/pswww/.

[5] 天球上で広がった天体の電磁波スペクトルを，天体上の各点で取得する観測をこう呼ぶ.

[6] https://www.eso.org/sci/facilities/paranal/instruments/uves.html.

[7] https://www.ing.iac.es/PR/wht_info/whtsauron.html.

[8] https://www.sdss.org/surveys/manga/.

[9] 爆発的な勢いで星を形成している銀河.

[10] ALMA の電波分光計は，特定の振動数範囲に限ったミリ波電波を検出する．この範囲を band とよぶ.

[11] パーセクは天文学で使われる距離の単位で，1 [pc] は 3.08×10^{18} [cm] に相当する.

[12] 輝線とは，ある量子力学的エネルギー準位から別のエネルギー準位に遷移するときのエネルギー差に相当する波長あるいは振動数の放射のことで，エネルギー差が離散的な値を取ることに対応して狭い波長に集中した放射となる．バンド輝線とは，そのような輝線が比較的狭い波長域に多数存在し，お互いに重なり合って太い輝線となったものを指す.

[13] 電離ガス(プラズマ)相，原子ガス相，分子ガス相などを指す.

[14] 式(2.2)は，任意の $\epsilon > 0$ について $\lim_{d/n \to 0} \mathbb{P}(\|\overline{x} - \vec{\mu}\| > \epsilon) = 0$ を意味する．ここで $\|\cdot\|$ はユークリッドノルムを表し，$\mathbb{P}(A)$ は事象 A の生起確率である．式(2.3)は，任意の $\epsilon > 0$ について $\lim_{d/n \to \infty} \mathbb{P}(\|\overline{x} - \vec{\mu}\| < \epsilon) = 0$ を意味する.

[15] 青嶋・矢田 (2019)を参照のこと.

[16] 従来のスパイクモデルは Johnstone (2001)で与えられ，$\lambda_1, \ldots, \lambda_m$ は 1 よりも大きい定数，$\lambda_{m+1} = \cdots = \lambda_d = 1$ なる非常に限定的なモデルであった．それを高次元データの特徴に合わせて一般化したのがモデル(2.18)である．ただし，一般化スパイクモデル(2.18)は，高次元データ空間の次元数と，それに伴う固有値の大きさ，そして，それを推定するための標本数との関係をみるための簡便なモデルであり，Yata and Aoshima (2013)は，一般化スパイクモデルを拡張したパワースパイクモデルを与え，高次元 PCA の漸近理論を構築している.

[17] 4 節および 5 節で見られるように実高次元データの固有値の最初のいくつかは非常に飛びぬけており，一般化スパイクモデルはその性質を非常によく表している.

[18] この文脈での一致性とは，推定量 $t(n)$ (n: 標本数)が任意の $\epsilon > 0$ について，

$$\lim_{n \to \infty} \mathbb{P}\{|t_n - \theta| < \epsilon\} = 1\,.$$

が成り立つことを指す.

[19] 式(2.20)はガウス分布を緩めた条件である．その条件のもと，ノイズの空間は図 2 に示されるような幾何学的表現をもつ.

[20] もう一つは Yata and Aoshima (2010) で提案されたクロスデータ行列法である．これら A-SPCA のコードについては，https://github.com/Aoshima-Lab/HDLSS-Tools を参照のこと.

[21] Takeuchi et al. (2024)において，いくつかの ω の値で検証している.

[22] Heliocentric velocity: 太陽を基準として測定した視線速度．銀河系内の太陽系の運動速度ベクトルを差し引いていない.

[23] ここで $M_\odot$ は太陽質量，すなわち太陽の質量 1.99×10^{33} [g] を 1 [$M_\odot$] とする単位．
[24] 活発に星が形成される領域でみられる巨大な星の集団．
[25] 宇宙膨張による輝線のシフトは厳密には速度によるものではないが，近傍銀河については近似的に速度とみなすことができる．これをハッブル速度とよぶ．
[26] 分光マップのデータは画像の空間方向 2 次元＋振動数分散方向 1 次元の 3 次元構造を持つため，データキューブと呼ばれている．
[27] 連続放射（輝線ではなく，広い振動数域にわたって現れる放射）の存在は状況をさらに複雑にするが，本研究ではこの問題は扱わない．
[28] 分子の熱運動によるドップラー効果により輝線の振動数が変わり，静止系振動数の周りに輝線が広がる効果をドップラーブロードニングという．この効果で広がった輝線の周辺部をウィングとよぶ．
[29] ここでの空間構造とは，天球上の 2 次元マップ上での各量の位置依存性を意味する．
[30] 放射する物体が我々に向かって進むとき，ドップラー効果によって放射の波長が青い方（波長の短い側，振動数の高い側）にシフトすることを指す．
[31] 輝線ではなく，電波の波長（振動数）について連続的な放射の観測画像で最も明るい領域をこのように表現する．
[32] 電磁波の干渉を利用した面分光装置の一種．
[33] 水素が電離して正の電荷をもったイオン（この場合は 1 個の陽子）になっているとき，電子が結合すると余剰分のエネルギーを放射してより安定な状態に遷移する．水素原子内の電子のエネルギーは量子力学によって決まる離散的準位を取る．電子は高いエネルギーから低いエネルギーに遷移していき，エネルギー準位の差に応じた輝線を放射する．これが水素の再結合線で，準位 3 から 2 への遷移によって放射されるのが Hα 線と呼ばれる．Hα 線は可視光線の赤に対応する波長を持つ．

謝　辞

本論文の査読者には大変有益なコメントおよび議論をいただき，論文の可読性と議論の明快さを改善することができた．ここに心より感謝する．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〆切に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．担当編集者の島谷健一郎氏の示唆に深く感謝する．この論文では ALMA データ（ADS/JAO.ALMA#2013.1.00099.S, ADS/JAO.ALMA#2013.1.00735.S）を使用している．ALMA はチリ共和国の協力のもと ESO, NSF，NINS，NRC，MOST および ASIAA，KASI によって維持される共同研究機関である．ALMA 天文台，ESO，AUI/NRAO，および国立天文台によって運営されている．本研究は日本学術振興会科学研究費補助金（21H01128，24H00247，19K03937，および JP17H06130）のサポートを受けている．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利用研究費「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．

河野海氏には本研究の初期の解析に大きな貢献をいただいた．また内容について有意義な議論をしていただいた中山優吾氏，池田思朗氏，原田ななせ氏に感謝する．

付　録

A. 電波領域における分子スペクトル輝線の放射機構

スペクトル輝線の放射機構は基本的に量子力学的効果である．原子の大きさ程度以下のミクロなスケールでは，素粒子の運動は古典力学ではなく量子力学で記述される．古典力学との顕著な違いは，エネルギーなど物理量が離散的な値を取ることである．これを「量子化されている」と表現する．その詳細は文献に譲るが（たとえば Wilson et al., 2013），離散化されたエネルギーの値をエネルギー準位と呼び，その間で量子力学的遷移が生じるとエネルギー差 ΔE に対応する波長

$$\lambda = \frac{hc}{\Delta E} \tag{A.1}$$

の輝線が放射される（h はプランク定数，c は光速）．

分子は原子に比べて複雑な構造をしており，分子の振動や回転にともなう複数の量子状態間の遷移によって電磁波を放射する．分子の振動準位間の遷移にともなう放射は多くの場合赤外線領域となるが，回転遷移では電波領域の放射も多数あり，電波天文学の重要な観測対象である．星間空間，特に低温の領域では，H_2，CO，OH，SiO などの 2 原子分子が存在する．古典力学的では，分子の重心周りの回転エネルギー $E_{\rm rot}$ は

$$E_{\rm rot} = \frac{\omega^2}{2}\left(m_1 r_1^2 + m_2 r_2^2\right) \tag{A.2}$$

と表される．ここで，m_1，m_2 は 2 つの原子の質量，r_1，r_2 は分子の重心からそれぞれの原子までの距離，ω は回転角速度である．一方，2 原子間の距離を r とすると

$$r_1 = \frac{m_2}{m_1+m_2}\, r,\ r_2 = \frac{m_1}{m_1+m_2}\, r \tag{A.3}$$

であることを用いると

$$E_{\rm rot} = \frac{\omega^2}{2}\left(\frac{m_1 m_2}{m_1+m_2}\right) r^2 \equiv \frac{m_{\rm red}}{2} r^2 \omega^2 \tag{A.4}$$

とできる．ここで $m_{\rm red}$ は換算質量と呼ばれる量

$$m_{\rm red} = \frac{m_1 m_2}{m_1+m_2} \tag{A.5}$$

である．さらに慣性モーメント I を

$$I \equiv m_{\rm red} r^2 \tag{A.6}$$

で定義すると，

$$E_{\rm rot} = \frac{I\omega^2}{2} \tag{A.7}$$

と表される．また回転の角運動量 $L_{\rm rot}$ は

$$L_{\rm rot} = I\omega \tag{A.8}$$

なので，

$$E_{\rm rot} = \frac{L_{\rm rot}^2}{2I} \tag{A.9}$$

が得られる．

一方，量子力学では $L^2_{\rm rot}$ を角運動量演算子

$$\hat{L}^2 \equiv -\hbar^2 \left\{ \frac{1}{\sin\theta}\left(\frac{\partial}{\partial\theta}\right)\left[\sin\theta\left(\frac{\partial}{\partial\theta}\right)\right] + \frac{1}{\sin^2\theta}\left(\frac{\partial^2}{\partial\varphi^2}\right)\right\} \tag{A.10}$$

で置き換えることにより，エネルギーの演算子であるハミルトニアン $\hat{H}$ を作り，シュレーディンガー方程式

$$E_{\rm rot}\psi = \hat{H}\psi = \frac{\hat{L}^2}{2I}\psi \tag{A.11}$$

を構成する ($\hbar \equiv h/2\pi$)．変数 (θ, φ) は 3 次元球座標系における角度変数であり，ψ は系の状態を記述する波動関数と呼ばれる関数である．

量子的な状態は確率で表現されるので，その力学的考察のためには式(A.11)を解いて ψ を求めればよい．演算子 $\hat{L}^2$ の固有値は $J = 0, 1, 2, \ldots$ として $\hbar^2 J(J+1)$ であるから，回転運動のエネルギー固有値 E_J は

$$E_J = \frac{\hbar^2}{2I} J(J+1) \tag{A.12}$$

となる．これが分子の回転のエネルギー準位の具体的表式である．本文で分子輝線の名前を HCN(4–3) のように表記しているが，これは分子が HCN で，回転準位 $J = 4 \longrightarrow 3$ の遷移にともなって放出された輝線であることを意味している．

B. ドップラーシフト(Doppler shift)

本章では物体の運動によって生じる光の波長の偏移，いわゆるドップラーシフト(Doppler shift)について説明する．この効果を起こす物理を説明するのは特殊相対性理論である．天体が観測者に対して速度 v(ここでは遠ざかる方向を正と定義する)で等速直線運動しているとし，θ は視線方向と運動方向のなす角を表すとする．

天体が観測者に対して静止しているとき，天体から観測者へ伝播する波長 λ の電磁波を考える．天体と観測者間の距離を d とすると，時間間隔 $t = d/c$ の間に $n = d/\lambda$ 個の波が放出される．一方，天体が観測者に対して速度 v で運動している場合，天体と観測者間の距離は時間間隔 t ののち

$$d' = d + (v\cos\theta)t \tag{B.1}$$

に増加している．ここで波の数 n が保存することに注意すれば，

$$\frac{\lambda'}{\lambda} = \frac{d + (v\cos\theta)t}{d} = 1 + \beta\cos\theta\,, \tag{B.2}$$

$$\beta \equiv \frac{v}{c} \tag{B.3}$$

が得られる．波長と振動数の間には $\lambda\nu = c$ の関係があるので，式(B.2)を振動数の関係式に換算すると

$$\frac{\nu'}{\nu} = \frac{1}{1 + \beta\cos\theta} \tag{B.4}$$

となる．これは，視線方向の運動によって電磁波の波長が引き伸ばされる効果である．

特殊相対論では，観測者から見ると運動している天体の時間は遅れる．つまり，時計の進み方は $\sqrt{1-\beta^2}$ 倍になり，振動数もこれに比例して小さくなる．すなわち，

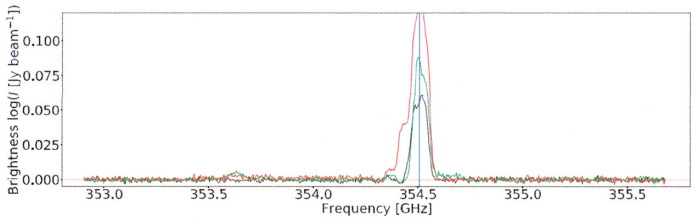

図 21. ドップラーシフト補正された分子輝線 HCN(4–3) の例．赤，緑，青は，線の中心がその静止系振動数と一致するように補正されたスペクトルである（Takeuchi et al., 2024）．

(B.5) $$\frac{\nu''}{\nu} = \sqrt{1-\beta^2}$$

である．この効果は天体の運動速度の大きさのみによるので，天体が視線に直交する方向に運動している場合 ($\theta = 2\pi$ [rad]) でも振動数が変化する．これを横ドップラー効果（transversal Doppler effect）とよぶ．

これら 2 つの効果により，最終的に特殊相対論的なドップラー効果による振動数変化は

(B.6) $$\nu_{\rm obs} = \frac{\sqrt{1-\beta^2}}{1+\beta\cos\theta}\,\nu_{\rm em}$$

と書ける．ここで，$\nu_{\rm em}$ は静止系で観測される振動数であり，$\nu_{\rm obs}$ は観測される振動数である．視線方向のみの効果を考えるならば（$\theta = 0$），

(B.7) $$\nu_{\rm obs} = \frac{\sqrt{1-\beta^2}}{1+\beta}\,\nu_{\rm em} = \sqrt{\frac{1-\beta}{1+\beta}}\,\nu_{\rm em}$$

となり，波長では

(B.8) $$\lambda_{\rm obs} = \sqrt{\frac{1+\beta}{1-\beta}}\,\lambda_{\rm em}$$

が求める関係である．下付き添え字 obs，em については式（B.7）と同様である．

本文 5 章で議論しているように，詳細な物理過程を考察するためには，特殊相対論的ドップラーシフトを補正して静止系での振動数に戻すことが必要である．これには式（B.7）を用いればよい．具体例として，HCN(4–3) 輝線のドップラー補正を図 21 に示す．赤，緑，青は NGC 253 のマップ上の異なる位置から補正されたスペクトルであり，輝線の中心がその静止系振動数と一致する．各輝線は複雑なプロファイルを持っているため，輝線の中心はガウス関数フィッティングによって決定している．青い縦の実線は，HCH(4–3) 輝線の静止系振動数 354.5 GHz を示している．

参 考 文 献

Ahn, J., Marron, J. S., Muller, K. M. and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions, *Biometrika*, **94**(3), 760–766.

Ando, R., Nakanishi, K., Kohno, K., Izumi, T., Martín, S., Harada, N., Takano, S., Kuno, N., Nakai, N., Sugai, H., Sorai, K., Tosaki, T., Matsubayashi, K., Nakajima, T., Nishimura, Y. and Tamura, Y. (2017). Diverse nuclear star-forming activities in the heart of NGC 253 resolved with 10-pc-scale

ALMA images, *The Astrophysical Journal*, **849**(2), 81, http://dx.doi.org/10.3847/1538-4357/aa8fd4.

Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data, *Sequential Analysis*, **30**(4), 356–399, http://dx.doi.org/10.1080/07474946.2011.619088.

青嶋 誠, 矢田和善 (2013). 論説：高次元小標本における統計的推測, 数学, **48**, 225–247.

Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data, *Annals of the Institute of Statistical Mathematics*, **66**, 983–1010, http://dx.doi.org/10.1007/s10463-013-0435-8.

Aoshima, M. and Yata, K. (2015). Geometric classifier for multiclass, high-dimensional data, *Sequential Analysis*, **34**(3), 279–294, http://dx.doi.org/10.1080/07474946.2015.1063256.

Aoshima, M. and Yata, K. (2017). Statistical inference for high-dimension, low-sample-size data, *Sugaku Expositions*, **30**(2), 137–158, http://dx.doi.org/10.1090/suga/421.

Aoshima, M. and Yata, K. (2019). High-dimensional quadratic classifiers in non-sparse settings, *Methodology and Computing in Applied Probability*, **21**, 663–682, http://dx.doi.org/10.1007/s11009-018-9646-z.

青嶋 誠, 矢田和善 (2019).『高次元の統計学』, 共立出版, 東京.

Aoshima, M., Shen, D., Shen, H., Yata, K., Zhou, Y.-H. and Marron, J. S. (2018). A survey of high dimension low sample size asymptotics, *Australian and New Zealand Journal of Statistics*, **60**, 4–19, http://dx.doi.org/10.1111/anzs.12212.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models, *Journal of Multivariate Analysis*, **97**(6), 1382–1408, http://dx.doi.org/10.1016/j.jmva.2005.08.003.

Baik, J., Ben Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, *Annals of Probability*, **33**(5), 1643–1697, http://dx.doi.org/10.1214/009117905000000233.

Bendo, G. J., Beswick, R. J., D'Cruze, M. J., Dickinson, C., Fuller, G. A. and Muxlow, T. W. B. (2015). ALMA observations of 99 GHz free-free and H40α line emission from star formation in the centre of NGC 253, *Monthly Notices of the Royal Astronomical Society*, **450**(1), L80–L84, http://dx.doi.org/10.1093/mnrasl/slv053.

Bolatto, A. D., Warren, S. R., Leroy, A. K., Walter, F., Veilleux, S., Ostriker, E. C., Ott, J., Zwaan, M., Fisher, D. B., Weiss, A., Rosolowsky, E. and Hodge, J. (2013). Suppression of star formation in the galaxy NGC 253 by a starburst-driven molecular wind, *Nature*, **499**(7459), 450–453, http://dx.doi.org/10.1038/nature12351.

Fernández-Ontiveros, J. A., Prieto, M. A. and Acosta-Pulido, J. A. (2009). The nucleus of NGC 253 and its massive stellar clusters at parsec scales, *Monthly Notices of the Royal Astronomical Society*, **392**(1), L16–L20, http://dx.doi.org/10.1111/j.1745-3933.2008.00575.x.

Galaz, G. and de Lapparent, V. (1998). The ESO-Sculptor Survey: Spectral classification of galaxies with $Z < 0.5$, *Astronomy & Astrophysics*, **332**, 459–478.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286**(5439), 531–537, http://dx.doi.org/10.1126/science.286.5439.531.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **67**(3), 427–444, http://dx.doi.org/10.1111/j.1467-9868.2005.00510.x.

Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. (1994). *A Handbook of Small Data Sets*, 1st edition, Chapman and Hall, London.

Ishii, A., Yata, K. and Aoshima, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context, *Journal of*

Statistical Planning and Inference, **170**, 186–199, http://dx.doi.org/10.1016/j.jspi.2015.10.007.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis, *Annals of Statistics*, **29**(2), 295–327, http://dx.doi.org/10.1214/aos/1009210544.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context, *Annals of Statistics*, **37**(6B), 4104–4130, http://dx.doi.org/10.1214/09-AOS709.

Kamphuis, P., Józsa, G. I. G., Oh, S. H., Spekkens, K., Urbancic, N., Serra, P., Koribalski, B. S. and Dettmar, R. J. (2015). Automated kinematic modelling of warped galaxy discs in large H I surveys: 3D tilted-ring fitting of H I emission cubes, *Monthly Notices of the Royal Astronomical Society*, **452**(3), 3139–3158, http://dx.doi.org/10.1093/mnras/stv1480.

Keto, E., Hora, J. L., Fazio, G. G., Hoffmann, W. and Deutsch, L. (1999). A super-star cluster in NGC 253: Mid-infrared properties, *The Astrophysical Journal*, **518**(1), 183–189, http://dx.doi.org/10.1086/307246.

Krieger, N., Bolatto, A. D., Walter, F., Leroy, A. K., Zschaechner, L. K., Meier, D. S., Ott, J., Weiss, A., Mills, E. A. C., Levy, R. C., Veilleux, S. and Gorski, M. (2019). The molecular outflow in NGC 253 at a resolution of two parsecs, *The Astrophysical Journal*, **881**(1), 43, http://dx.doi.org/10.3847/1538-4357/ab2d9c.

Martín, S., Mangum, J. G., Harada, N., Costagliola, F., Sakamoto, K., Muller, S., Aladro, R., Tanaka, K., Yoshimura, Y., Nakanishi, K., Herrero-Illana, R., Mühle, S., Aalto, S., Behrens, E., Colzi, L., Emig, K. L., Fuller, G. A., García-Burillo, S., Greve, T. R., Henkel, C., Holdship, J., Humire, P., Hunt, L., Izumi, T., Kohno, K., König, S., Meier, D. S., Nakajima, T., Nishimura, Y., Padovani, M., Rivilla, V. M., Takano, S., van der Werf, P. P., Viti, S. and Yan, Y. T. (2021). ALCHEMI, an ALMA comprehensive high-resolution extragalactic molecular inventory. Survey presentation and first results from the ACA array, *Astronomy & Astrophysics*, **656**, A46, http://dx.doi.org/10.1051/0004-6361/202141567.

Matsubayashi, K., Sugai, H., Hattori, T., Kawai, A., Ozaki, S., Kosugi, G., Ishigaki, T. and Shimono, A. (2009). Galactic wind in the nearby starburst galaxy NGC 253 observed with the Kyoto3DII Fabry-Perot mode, *The Astrophysical Journal*, **701**(2), 1636–1643, http://dx.doi.org/10.1088/0004-637X/701/2/1636.

Pace, Z. J., Tremonti, C., Chen, Y., Schaefer, A. L., Bershady, M. A., Westfall, K. B., Boquien, M., Rowlands, K., Andrews, B., Brownstein, J. R., Drory, N. and Wake, D. (2019). Resolved and integrated stellar masses in the SDSS-IV/MaNGA survey. I. PCA spectral fitting and stellar mass-to-light ratio estimates, *The Astrophysical Journal*, **883**(1), 82, http://dx.doi.org/10.3847/1538-4357/ab3723.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica*, **17**(4), 1617–1642.

Portillo, S. K. N., Parejko, J. K., Vergara, J. R. and Connolly, A. J. (2020). Dimensionality reduction of SDSS spectra with variational autoencoders, *The Astrophysical Journal*, **160**(1), 45, http://dx.doi.org/10.3847/1538-3881/ab9644.

Rekola, R., Richer, M. G., McCall, M. L., Valtonen, M. J., Kotilainen, J. K. and Flynn, C. (2005). Distance to NGC 253 based on the planetary nebula luminosity function, *Monthly Notices of the Royal Astronomical Society*, **361**(1), 330–336, http://dx.doi.org/10.1111/j.1365-2966.2005.09166.x.

Rieke, G. H., Lebofsky, M. J., Thompson, R. I., Low, F. J. and Tokunaga, A. T. (1980). The nature of the nuclear sources in M82 and NGC 253, *The Astrophysical Journal*, **238**, 24–40, http://dx.doi.org/10.1086/157954.

Ronen, S., Aragon-Salamanca, A. and Lahav, O. (1999). Principal component analysis of synthetic galaxy spectra, *Monthly Notices of the Royal Astronomical Society*, **303**(2), 284–296, http://dx.doi.org/10.1046/j.1365-8711.1999.02222.x.

Sakamoto, K., Mao, R.-Q., Matsushita, S., Peck, A. B., Sawada, T. and Wiedner, M. C. (2011). Star-

forming cloud complexes in the central molecular zone of NGC 253, *The Astrophysical Journal*, **735**(1), 19, http://dx.doi.org/10.1088/0004-637X/735/1/19.

Takeuchi, T. T., Yata, K., Egashira, K., Aoshima, M., Ishii, A., Cooray, S., Nakanishi, K., Kohno, K. and Kono, K. T. (2024). High-dimensional statistical analysis and its application to an ALMA map of NGC 253, *The Astrophysical Journal Supplement Series*, **271**(2), 44, http://dx.doi.org/10.3847/1538-4365/ad2517.

Walter, F., Bolatto, A. D., Leroy, A. K., Veilleux, S., Warren, S. R., Hodge, J., Levy, R. C., Meier, D. S., Ostriker, E. C., Ott, J., Rosolowsky, E., Scoville, N., Weiss, A., Zschaechner, L. and Zwaan, M. (2017). Dense molecular gas tracers in the outflow of the starburst galaxy NGC 253, *The Astrophysical Journal*, **835**(2), 265, http://dx.doi.org/10.3847/1538-4357/835/2/265.

Wang, L., Farrah, D., Connolly, B., Connolly, N., Lebouteiller, V., Oliver, S. and Spoon, H. (2011). Principal component analysis of the Spitzer IRS spectra of ultraluminous infrared galaxies, *Monthly Notices of the Royal Astronomical Society*, **411**(3), 1809–1818, http://dx.doi.org/10.1111/j.1365-2966.2010.17811.x.

Wilson, T. L., Rohlfs, K. and Hüttemeister, S. (2013). *Tools of Radio Astronomy*, Springer Verlag, Heidelberg, http://dx.doi.org/10.1007/978-3-642-39950-3.

Yata, K. and Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context, *Communications in Statistics — Theory and Methods*, **38**(16-17), 2634–2652, http://dx.doi.org/10.1080/03610910902936083.

Yata, K. and Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, *Journal of Multivariate Analysis*, **101**(9), 2060–2077, http://dx.doi.org/10.1016/j.jmva.2010.04.006.

Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, *Journal of Multivariate Analysis*, **105**(1), 193–215, http://dx.doi.org/https://doi.org/10.1016/j.jmva.2011.09.002.

Yata, K. and Aoshima, M. (2013). PCA consistency for the power spiked model in high-dimensional settings, *Journal of Multivariate Analysis*, **122**, 334–354, http://dx.doi.org/10.1016/j.jmva.2013.08.003.

Yata, K. and Aoshima, M. (2022). Automatic sparse PCA for high-dimensional data, *Statistica Sinica* (in press), http://dx.doi.org/10.5705/ss.202022.0319.

Proceedings of the Institute of Statistical Mathematics Vol. 72, No. 2, 273–303 (2024) 303

Exploration of the Physical Properties of Molecular Gas in a Galaxy with High-dimensional Statistical Analysis and Future Prospects to Astronomy

Tsutomu T. Takeuchi[1,2], Kazuyoshi Yata[3], Kento Egashira[4], Makoto Aoshima[3], Kohji Yoshikawa[5], Aki Ishii[4], Ryusei R. Kano[1], Wen E. Shi[1], Aina May So[1,8], Hai-Xia Ma[1], Sena A. Matsui[1], Koichiro Nakanishi[6,9], Sucheta Cooray[6,7] and Kotaro Kohno[10]

[1]Division of Particle and Astrophysical Science, Nagoya University
[2]Research Center for Statistical Machine Learning, The Institute of Statistical Mathematics
[3]Institute of Mathematics, University of Tsukuba
[4]Department of Information Sciences, Tokyo University of Science
[5]Center for Computational Sciences, University of Tsukuba
[6]National Astronomical Observatory of Japan
[7]Research Fellow of the JSPS (PD)
[8]Department of Physics, Gakushuin University
[9]Department of Astronomy, School of Science, Graduate University for Advanced Studies (SOKENDAI)
[10]Institute of Astronomy, Graduate School of Science, The University of Tokyo

If we denote the dimension of data as d and the number of samples as n, we often meet a case with $n \ll d$. Traditionally in astronomy, such a situation is regarded as ill-posed, and they thought that there was no choice but to throw away most of the information in data dimension to let $d < n$. The data with $n \ll d$ is referred to as high-dimensional low sample size (HDLSS). To deal with HDLSS problems, a method called high-dimensional statistics has been developed rapidly in the last decade. In this work, we first introduce the high-dimensional statistical analysis. We apply two representative methods in the high-dimensional statistical analysis methods, the noise-reduction principal component analysis (NRPCA) and automatic sparse principal component analysis (A-SPCA), to a spectroscopic map of a nearby archetype starburst galaxy NGC 253 taken by the Atacama Large Millimeter/Submillimeter Array (ALMA). The ALMA map is a typical HDLSS dataset. First we analyzed the original data including the Doppler shift due to the systemic rotation. The high-dimensional PCA could describe the spatial structure of the rotation precisely. We then applied to the Doppler-shift corrected data to analyze more subtle spectral features. The NRPCA and A-SPCA could quantify the very complicated characteristics of the ALMA spectra. Particularly, we could extract the information of the global outflow from the center of NGC 253. This method can also be applied not only to spectroscopic survey data, but also any type of HDLSS data.

Key words: High-dimensional statistical analysis, principal component analysis, interstellar medium, molecular emission line, starburst, galaxy evolution.