Statistical Challenges to Dimensionality in Astronomical Big Data

Tsutomu T. TAKEUCHI
1. Division of Particle and Astrophysical Science, Nagoya University, Japan
2. The Research Center for Statistical Machine Learning, the Institute of Statistical Mathematics

International Symposium on Recent Advances in Theories and Methodologies for Large Complex Data, Tsukuba, 7-9 Dec., 2023

Collaborators

Suchetha COORAY, Kai T. KONO (河野 海)
Division of Particle and Astrophysical Science, Nagoya University, Japan

Kazuyoshi YATA (矢田 和善), Makoto AOSHIMA(青嶋 誠)
Institute of Mathematics, University of Tsukuba, Japan

Kento EGASHIRA (江頭 健斗), Aki ISHII (石井 晶)
Department of Information Sciences, Tokyo University of Science, Japan

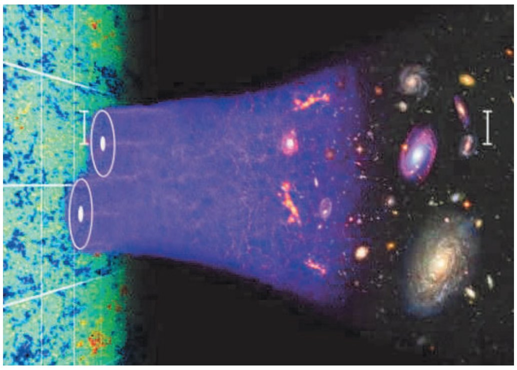
Kohji YOSHIKAWA (吉川 耕司)
Center for Computational Sciences, University of Tsukuba, Japan

Kouichiro NAKANISHI (中西 康一郎)
ALMA Project, National Astronomical Observatory of Japan

Kotaro KOHNO (河野 孝太郎)
Institute of Astronomy, The University of Tokyo, Japan

1. Introduction

1.1 Structure formation in the Universe

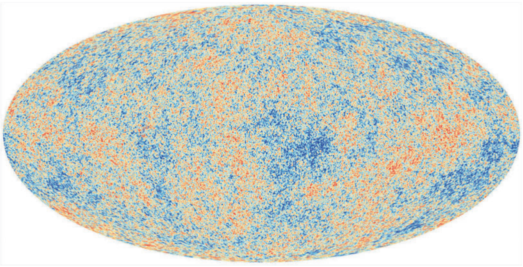


All the structures in the Universe have emerged from a tiny fluctuation at very early epoch (380,000 yr).

1.2 Galaxy formation from the cosmic initial condition

Galaxies are supposed to have formed from a tiny (order of ~ 10^{-5}) fluctuation of matter (mainly dark matter: DM) in the early Universe.

The initial condition is imprinted on the Cosmic Microwave Background (CMB) observed at radio wavelengths.



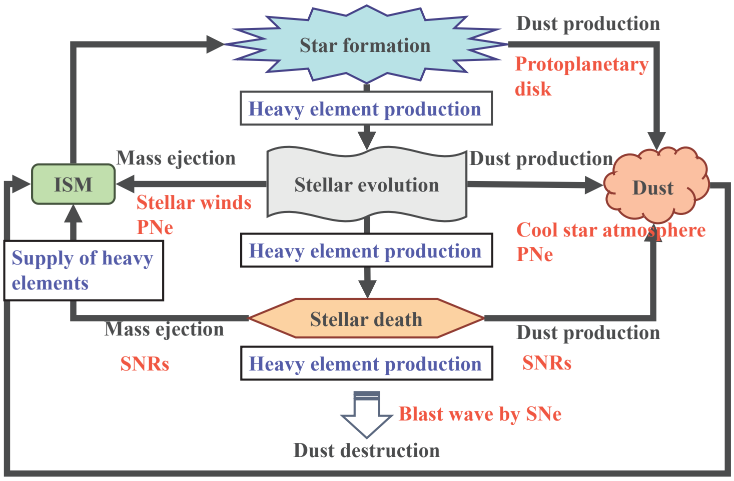
http://www.rssd.esa.int/index.php?project=Planck

1.4 Internal galaxy evolution

Star formation in galaxies

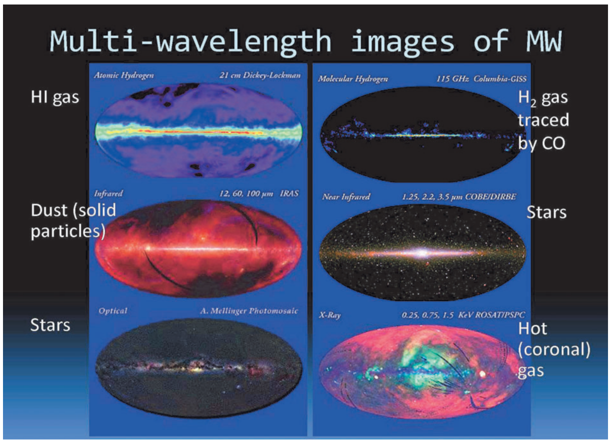
Galaxies have formed at various epochs in the Universe, merged, and grown. In parallel, gas has transformed into stars. Stars die and return back their gas into the ISM, and next generation of star formation proceeds.

Chemical evolution of galaxies: metal and dust



2. Interstellar Medium (ISM) and Spectroscopy

2.1 Matter between stars in galaxies



2.2 ISM phases and star formation

ISM has various phases

1. Plasma (ionized diffuse phase)
2. Neutral gas (mainly neutral hydrogen HI)
3. Molecular gas (mainly molecular hydrogen H_2)

Since gas must become dense enough to form stars, star formation occurs in molecular clouds. Namely,

Atomic gas ⇒ Molecular gas ⇒ Stars

Kennicutt-Schmidt (K-S) law

Stars form in molecular cores.

⇒ It is natural to suppose a relation between the star formation rate (SFR) and gas density. Schmidt (1959) proposed a relation

$$\mathrm{SFR} \propto \rho^n .$$

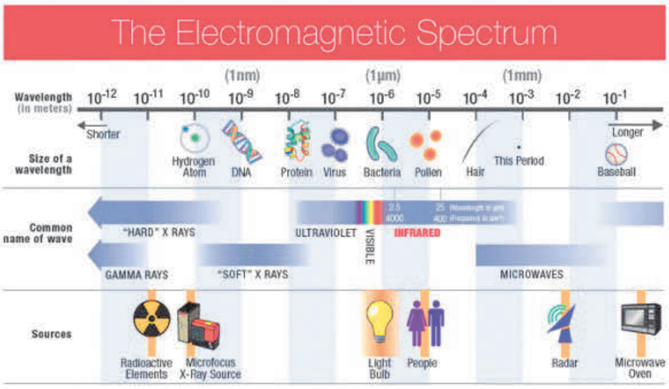
i. $n = 1$ Density controls star formation.
ii. $n = 2$ Collision-like process plays a role for star formation

⇒ **The power-law index contains substantial information on what triggers the star formation.**

It is crucial to reveal spatially resolved SF law in galaxies!

2.3 Spectroscopy of the ISM

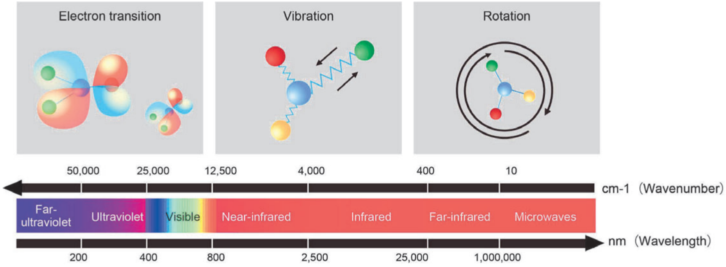
Electromagnetic spectrum and scales



https://www.americanpharmaceuticalreview.com/Featured-Articles/331616-Optical-Spectroscopy-Where-is-it-Going/

Quantum transition to spectral lines

Astronomical spectroscopy brings physical information of the objects in the remote Universe.



https://www.yokogawa.com/about/research-development/inv_center/spectroscopy/

2.4 Spectral mapping in astronomy

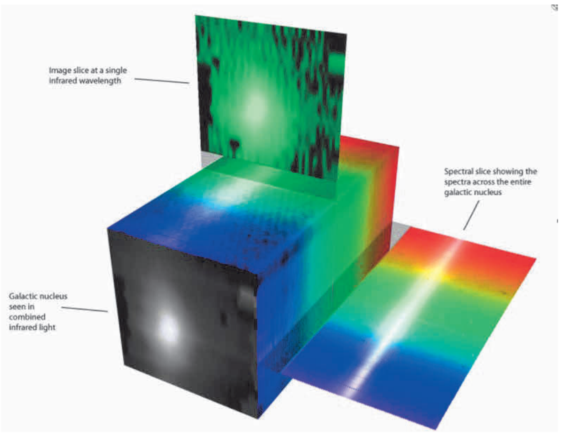


Image slice at a single infrared wavelength
Spectral slice showing the spectra across the entire galactic nucleus
Galactic nucleus seen in combined infrared light

http://ifs.wikidot.com/what-is-ifs

3. Astronomical Spectral Map as HDLSS Data

3.1 General situation in astrophysics

Classical statistical analysis

Sample size: *n*
Data dimension: *d*

The following condition is implicitly assumed

$$n >> d$$

But this is not the case for many cases in scientific researches. Astronomers and astrophysicists have ever simply given up when they face such type of problem.

3. Astronomical Spectral Map as HDLSS Data

3.1 General situation in astrophysics

High-dimensional low-sample size (HDLSS) data analysis

Sample size: *n*
Data dimension: *d*

For the HDLSS data, the condition is

$$n << d$$

This condition is often found in e.g., genomic analysis, medical analysis, etc.

In astrophysics, for example, integral field spectroscopy has this property.

3.2 Unusual behavior of high-dimensional data

For high-dimensional data, classical limit theorems do not work. If we wrongly assume them, we would be lead to a wrong conclusion.

Simplest example: for the sample mean

$$\bar{\vec{x}} = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i$$

1. as $d/n \to 0$

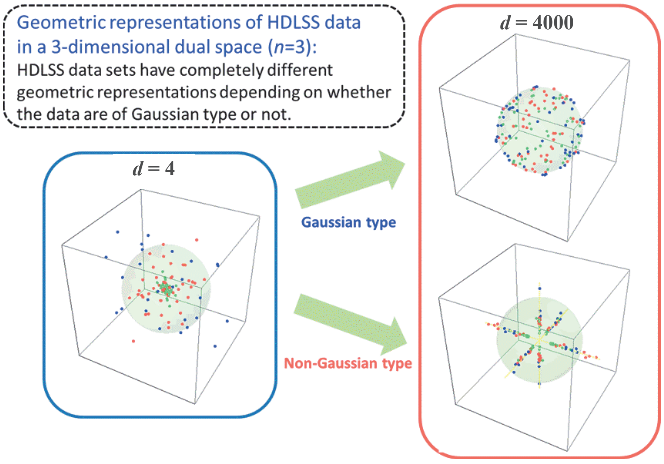
$$\| \bar{\vec{x}} - \vec{\mu} \| \xrightarrow{P} \vec{0}$$

2. as $d/n \to \infty$

$$\| \bar{\vec{x}} - \vec{\mu} \| \xrightarrow{P} \infty$$

This striking property is referred to as the strong inconsistency.

Unusual behavior of high-dimensional data

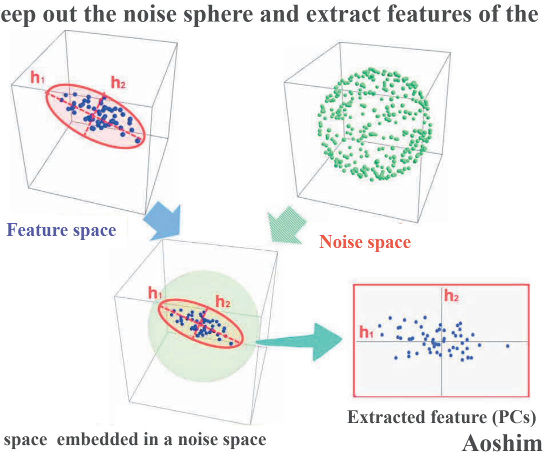


Geometric representations of HDLSS data in a 3-dimensional dual space (n=3):
HDLSS data sets have completely different geometric representations depending on whether the data are of Gaussian type or not.

https://www.math.tsukuba.ac.jp/~aoshima-lab/research.html

3.3 High-dimensional PCA

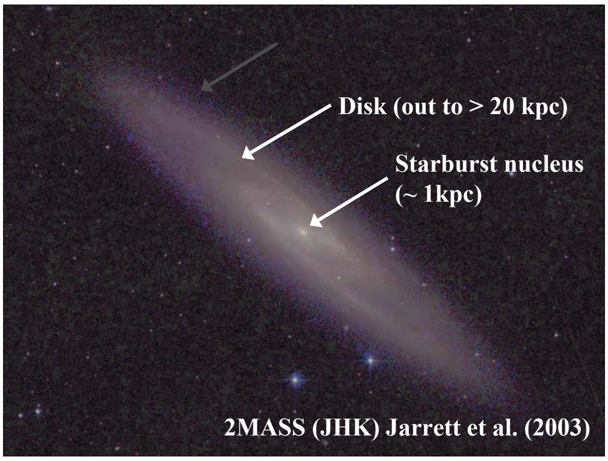
A specially designed PCA, the high-dimensional PCA, can sweep out the noise sphere and extract features of the data.



Aoshima (2012)

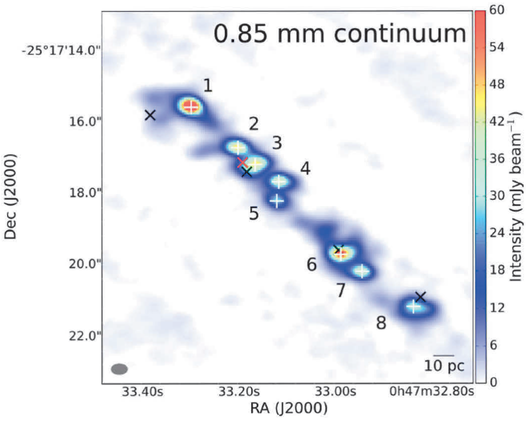
3.4 Actual data: ALMA data cube of NGC253

NGC 253: prototypal starburst



2MASS (JHK) Jarrett et al. (2003)

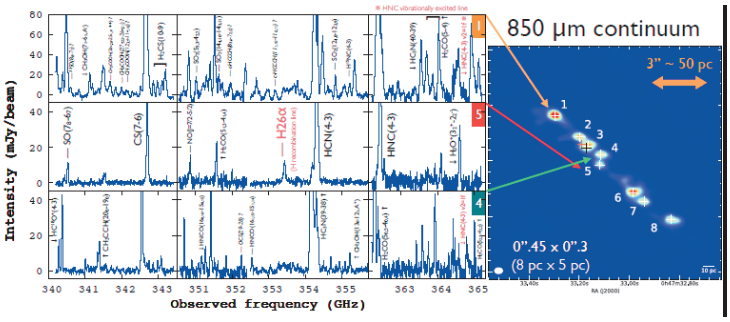
Close up of the starburst



Ando et al. (2017)

Rich in molecular lines

ALMA resolved diverse star-forming activities at ~ 10 pc scale.



ALMA Band7 spectra

Ando et al. (2017)

3.5 Structure of the Data

Data: Ando et al. (2017)

~ spatial dimension 231 × spectral dimension 2248

⇒ A case with n = 231 and d = 2248 ($n << d$)

Problems from astrophysical side

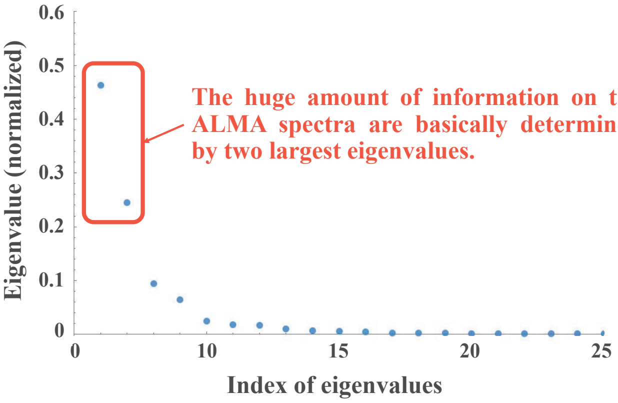
- Too much information on spectra.
- Too large variety of spectral lines compared to n.

We apply the high-dimensional statistical analysis to the ALMA spectral mapping data of NGC253.

4. Analysis of Starburst Region in NGC253

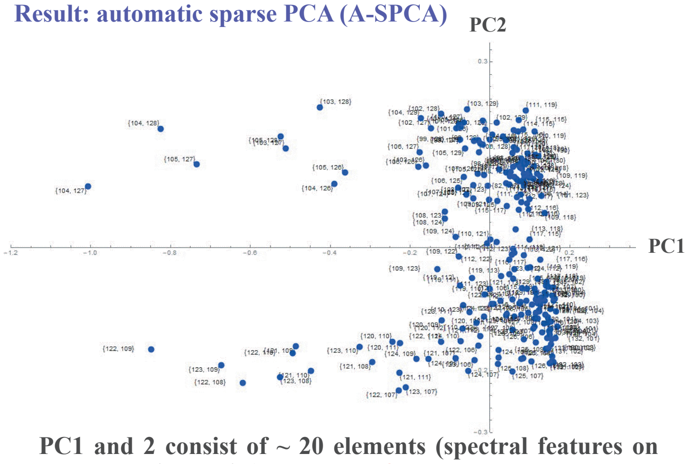
4.1 Analysis of Raw Data

Eigenvalues of the PCA (contribution)



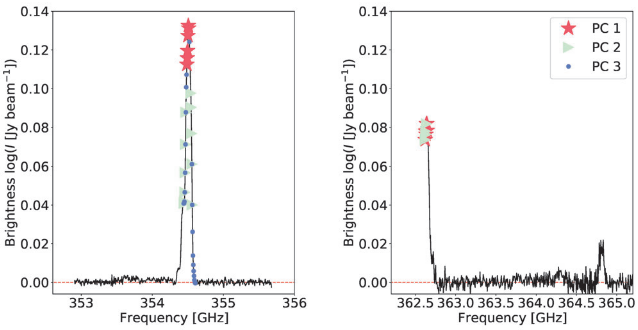
The huge amount of information on the ALMA spectra are basically determined by two largest eigenvalues.

Result: automatic sparse PCA (A-SPCA)



PC1 and 2 consist of ~ 20 elements (spectral features on the resolution units). The key features may be reduced only to a few to several lines!

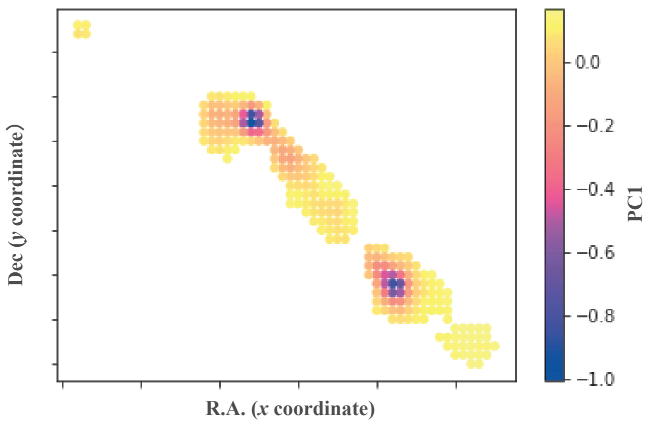
Responsible spectral features for PC1, PC2 and PC3



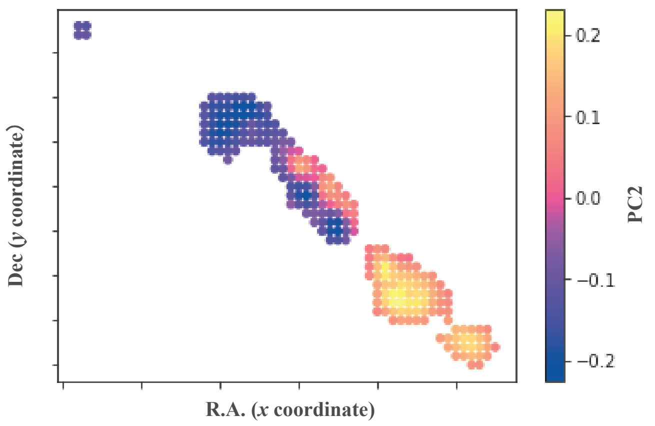
Takeuchi et al. (2023)

Now PC1 more clearly represents the total intensity, and PC2 and 3 represent smaller-scale velocity structures.

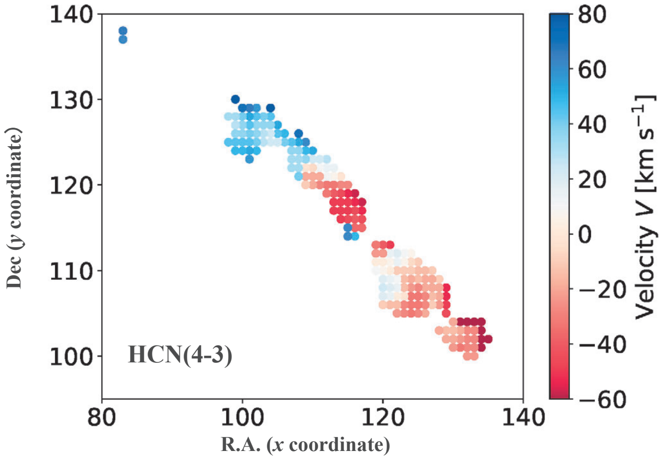
Spatial map of PC1



Spatial map of PC1 and PC2

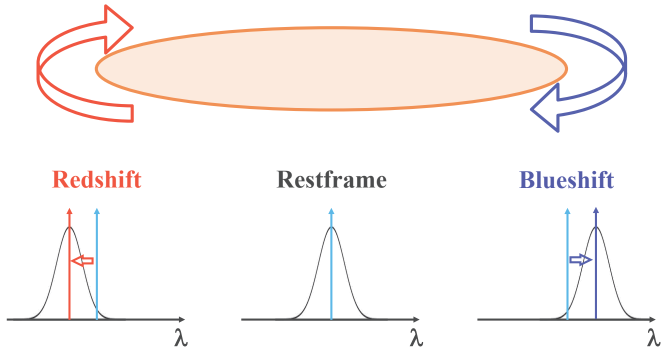


Velocity field of the systemic rotation



⇒ Doppler shift correction to remove the systemic rotation.

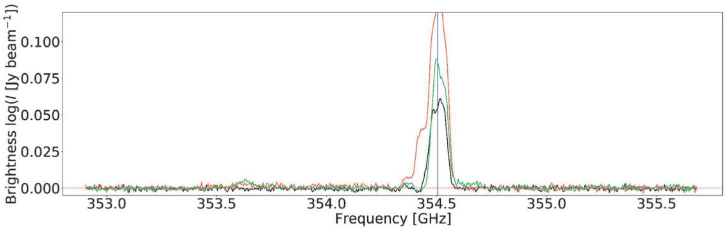
Systemic rotation and Doppler shift



If the system is rotating as a whole, the observed wavelength is affected by the Doppler shift. PC2 beautifully describes the Doppler shift!

4.3 Main analysis

Doppler shift correction

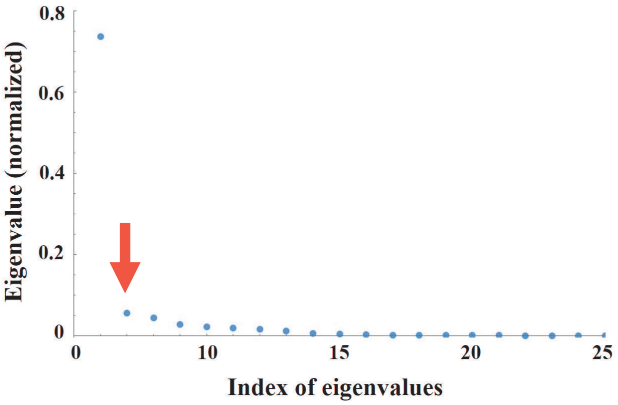


Takeuchi et al. (2021)

We estimated the peculiar velocity field (mainly due to the systemic rotation of the central region of NGC253) by averaging the results from HCN(4-3), HNC(4-3) and CS(7-6) lines, and corrected the Doppler shift.

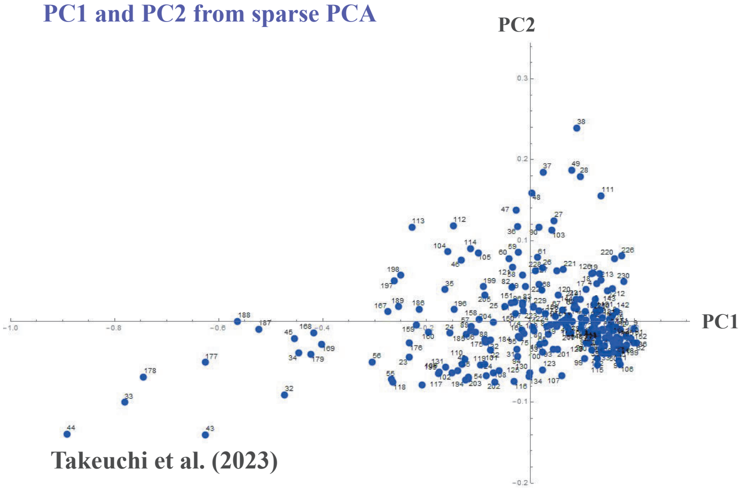
Due to this correction, the final data dimension is $d = 1971$.

Eigenvalues of the NGC253 after Doppler correction



Takeuchi et al. (2023)

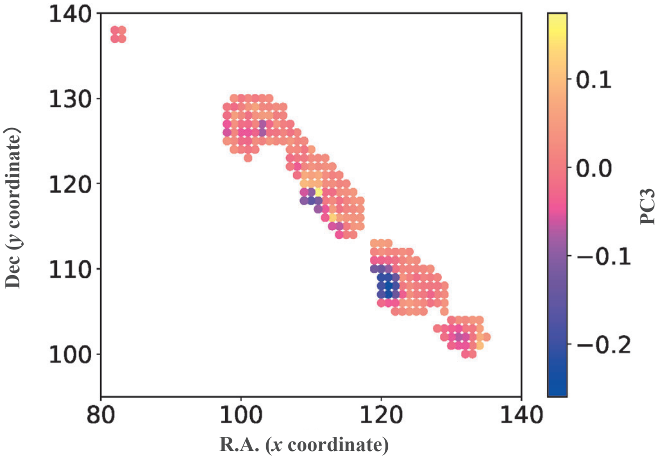
PC1 and PC2 from sparse PCA



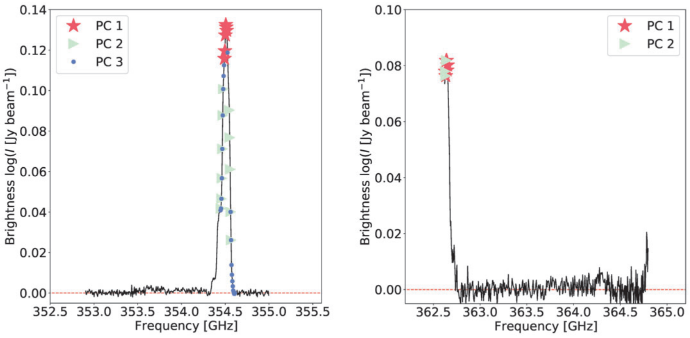
Takeuchi et al. (2023)

Butterfly-like pattern completely disappeared.

Spatial map of PC3 after Doppler correction



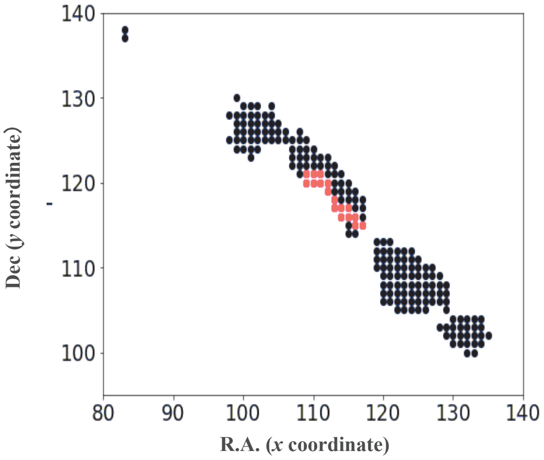
Responsible spectral features for PC1, PC2 and PC3



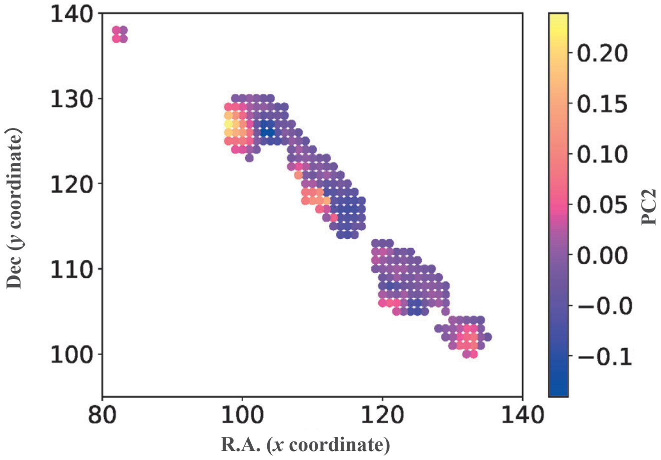
Takeuchi et al. (2023)

Now PC1 more clearly represents the total intensity, and PC2 and 3 represent smaller-scale velocity structures.

Anomaly regions in the velocity field



Spatial map of PC2 after Doppler correction



What do we see from the Doppler-corrected map?

NGC253

- **Pure starburst: SFR in the central molecular zone is 2 $M_\odot$ yr^{-1} (Rieke et al. 1980; Keto et al. 1999)**
- **Intense outflow (Matsubayashi et al. 2009; Bolatto et al. 2013)**

Indeed the outflow phenomenon is mainly delineated by PC3.

5. Galaxy Evolution in Multiwavelengths

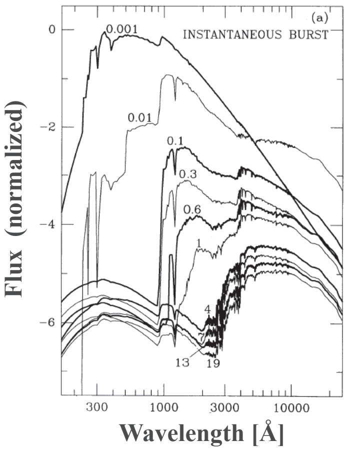
5.1 Galaxy evolution from a modern point of view

Galaxies evolve in various aspects:

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

This is the formal and ultimate goal of the studies on galaxy evolution, but clearly it is a substantially complicated problem. It is time to define the evolution of galaxies with more objective point of view.

5.2 Galaxy evolution in multiband luminosity space



Star formation history (SFH) is one of the key factors of galaxy evolution.

SFH is directly reflected to the spectral luminosity of galaxies.

⇩

Galaxy evolution related to the SFH will be well represented in the multiwavelength (band) luminosity space.

6. Galaxy Evolution in Multiwavelengths

6.1 Data: RCSED

- Reference Catalog of galaxy Spectral Energy Distributions (RCSED) (Chilingarian et al. 2016)
- Catalog of galaxies produced as join between *GALEX*, SDSS, and UKIDSS catalogs, and processed with state-of-the-art spectral analysis methods
- Covers approximately 25% of the sky and contains *k*-corrected ultraviolet-to-near-infrared photometry (11 bands of FUV, NUV, *u*, *g*, *r*, *i*, *z*, *Y*, *J*, *H*, *K*) of some 1 million galaxies, as well as some of their physical properties

RCSED

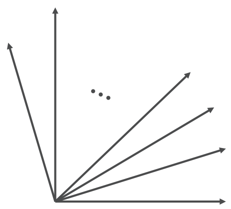
http://rcsed.sai.msu.ru

6.2 Classification in multiwavelength luminosity space

Generate a subsample with all 11 rest-frame magnitudes (FUV, NUV, *u*, *g*, *r*, *i*, *z*, *Y*, *J*, *H*, *K*) ~ 800,000 galaxies

⇒ Construct a volume limited sample that is representative of the whole galaxy sample of ~ 30,000 galaxies.

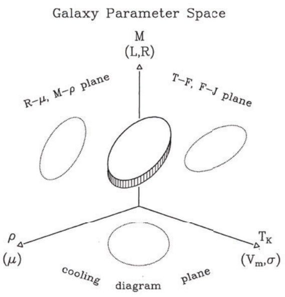
Perhaps impossible to classify by human intuition.



Galaxy manifold

Historically, in 80's, astronomers introduced a method of classical multivariate analysis such as PCA to find and unify various scaling relations (e.g., Djorgovski 1992).

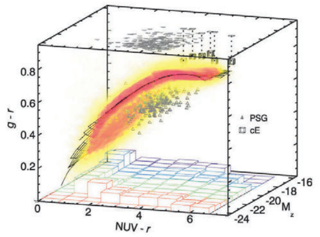
However, since classical PCA-type analysis could only find linear structure in the feature space, the idea worked only to a limited problems, and have been once forgotten.



Djorgovski (1992)

Galaxy manifold

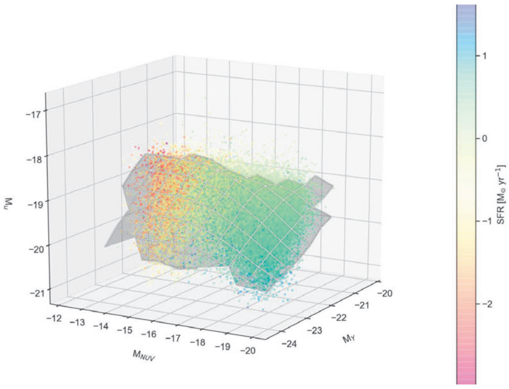
Some preceding studies have suggested the existence of a smooth relation of galaxies in the 3D color–color–magnitude space smoothly continuing from the blue cloud to the red sequence (e.g. Chilingarian et al. 2012).



Chilingarian et al. (2012)

⇒ general idea of a low dimensional submanifold existing in a higher dimensional feature space: revival of the galaxy manifold!

6.3 Galaxy manifold in the space of three bands



Takeuchi et al. (2023c)

A similar structure was found in the luminosity space of our data. We want to quantify and examine the manifold.

7. Application of Manifold Learning to Galaxies

7.1 Features of the galaxy manifold

1. Not uniformly distributed along the manifold

   Dense regions on the manifold imply that galaxies spend long time there in their evolution, and less dense regions suggest a rapid evolution there, and galaxies move on to a much more stable state.

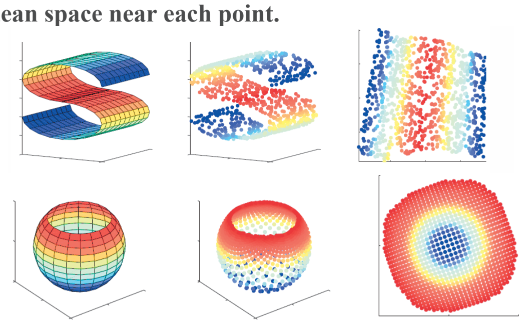
2. Noisy in feature space

   Input features depend on the spectral energy distribution (SED) modeling methods (dust needs to be considered carefully).

7.2 Manifold learning

Manifold Learning (or non-linear dimensionality reduction) embeds data that originally lies in a high dimensional space in a lower dimensional space, while preserving characteristic properties.

A manifold is a topological space that locally resembles Euclidean space near each point.



7.3 Algorithms

We applied two algorithms of manifold learning:

Isomap

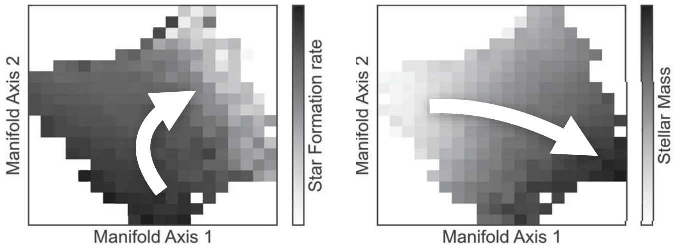
UMAP

Both methods keep connectivity of a point data cloud.

N.B. These methods do not determine the dimension of the manifold, because they reduce the dimension of the data cloud to two or three for any dataset.

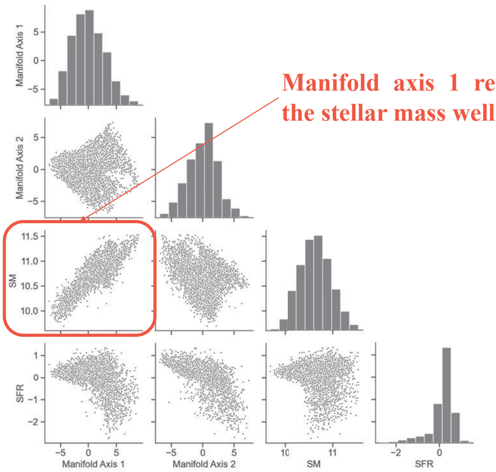
7.4 Result: galaxy manifold from Isomap

Isomap preserves the density of the data point cloud (i.e., distribution of galaxies) on the manifold (i.e. dense regions remain dense).



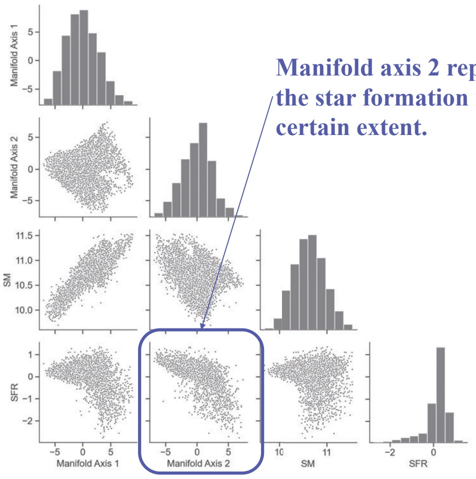
The arrows show the continuous evolution of less-massive, actively star-forming galaxies to massive, quiescent galaxies.

Isomap manifold axes and physical properties



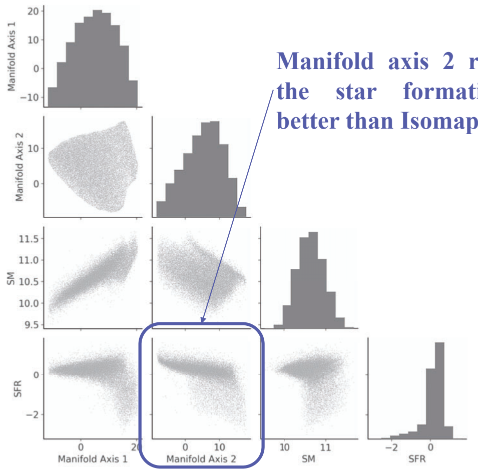
Manifold axis 1 represents the stellar mass well.

Isomap manifold axes and physical properties



Manifold axis 2 represents the star formation rate to a certain extent.

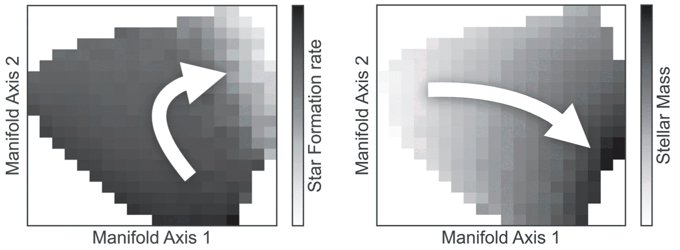
UMAP manifold axes and physical properties



Manifold axis 2 represents the star formation rate better than Isomap.

7.5 Result: galaxy manifold from UMAP

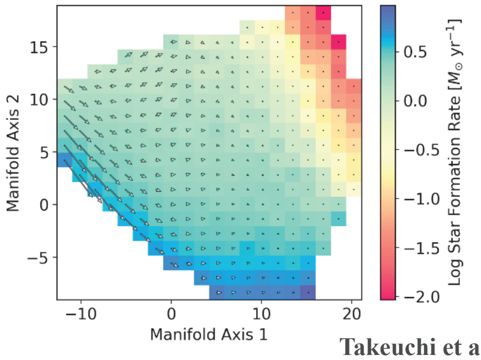
UMAP expands dense regions and contracts sparse regions, i.e., uniform representation of the observational galaxy parameters.



The arrows show the continuous evolution of less-massive, actively star-forming galaxies to massive, quiescent galaxies.

8. Interpretation and modeling

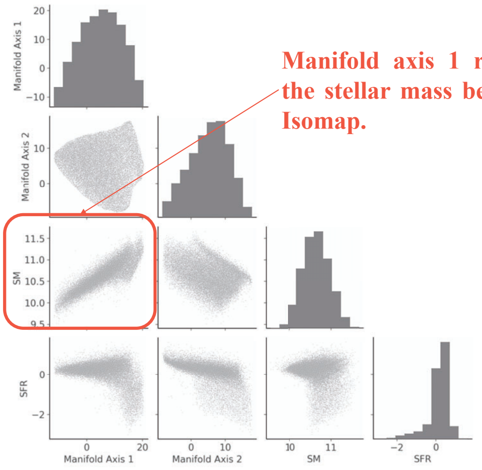
8.1 Physical interpretation of the manifold



Takeuchi et al. (2023d)

The galaxy manifold obtained with UMAP. The vector field represents the direction of galaxy evolution by star formation.

UMAP manifold axes and physical properties



Manifold axis 1 represents the stellar mass better than Isomap.

8.2 Interpretation by a physical model

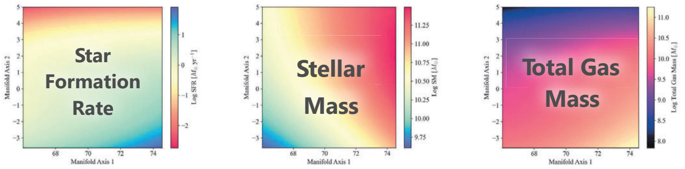
Assuming that galaxies move on the manifold with time, we model galaxies as a walk process.

The next location is decided from the following equations:

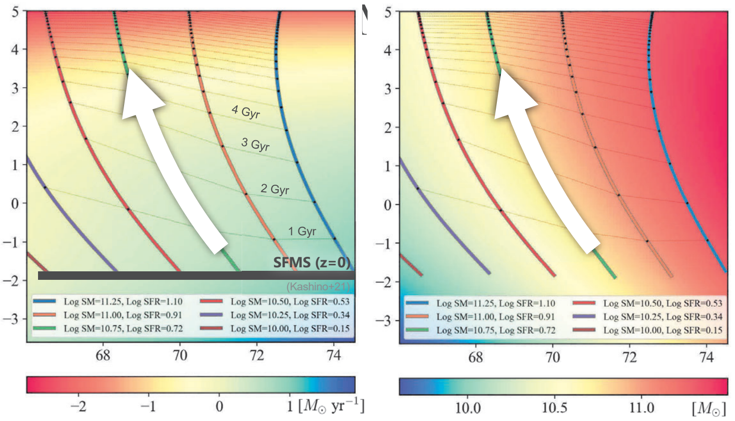
$$M_*(t_{n+1}) = M_*(t_n) + (1 - r)\mathrm{SFR}(t_n)\Delta t$$

$$M_{\mathrm{gas}}(t_{n+1}) = M_{\mathrm{gas}}(t_n) - (1 - r + \eta)\mathrm{SFR}(t_n)\Delta t$$

Δt = 10 [Myr], r (return fraction) = 0.35, η (mass loading factor) ~ 3. A manifold location that satisfies the above is chosen as the next position.



Physical interpretation of the star formation history



Star formation rate

Stellar mass

cf. Cooray et al. (2023)

9. Summary

8. The high-dimensional sparse PCA successfully chose two PCs that reproduce the general properties of the ALMA spectroscopic map of NGC253.

9. The controlling feature was HCN(4-3) rotational lines. PC1 describes the total intensity of the lines, and PC2 represents the Doppler shift caused by the systemic rotation.

10. After correcting the Doppler shift due to the systemic rotation, we could obtain information on the smaller-scale velocity field described by PC2 (new) and PC3. These may be caused by outflow phenomena of starburst regions.

9. Summary

1. Galaxies ubiquitously exist in the present-day Universe, but they have been formed from a tiny fluctuation of matter in the early Universe.

2. Evolution of galaxies is mainly driven by the star formation, a transition from ISM to stars.

3. Various phases of the ISM are related, and the evolution of the ISM is a key to complete the understanding of the galaxy evolution.

4. Spectroscopic and multiwavelength photometric observations are of vital importance to extract and interpret the information of matter in galaxies.

9. Summary

Part III
We demonstrated that manifold learning is a powerful method to extract physical laws embedded in a high-dimensional feature space of galaxies.

11. Main difficulty of the analysis of multiwavelength data was that the dimension of the feature space is high and classical methods do not give a good intuition.

12. The manifold learning is useful to find a hidden relation in a high-dimensional feature space.

13. Algorithms that keep the connectivity are desirable for the purpose of the analysis of galaxy evolution. Isomap and UMAP are promising for this purpose.

9. Summary

Part II
Spectroscopic mapping and similar methods are fundamentally important to reveal the ISM physics, but the data are high-dimensional low sample size.

6. We applied the high-dimensional PCA on the NGC253 spectral map. ALMA mapping data are typically HDLSS in general, and in this case $n = 231$ and $d = 2227$.

7. Very large variety in the molecular line spectra of NGC253 map can be described only by two PCs! Each PC consists of ~ 20 elements, much fewer than d. Because these elements may be a part of same features, the key features may be reduced to several.

9. Summary

14. Isomap and UMAP give a consistent result, and we found that the galaxy evolution is well described only by two physical quantities, the stellar mass and the star formation rate (SFR).

15. It is crucial to describe the discovered nonlinear relation by interpretable equations. This is not, however, an easy task, and at this moment a comparison with classical theory works fairly well.

Dimensionality reduction in astrophysics plays an important role for new discoveries. Further investigation is needed.

Stay tuned!