Application of High-dimensional Statistical Analysis to Astrophysical Problems

Tsutomu T. TAKEUCHI

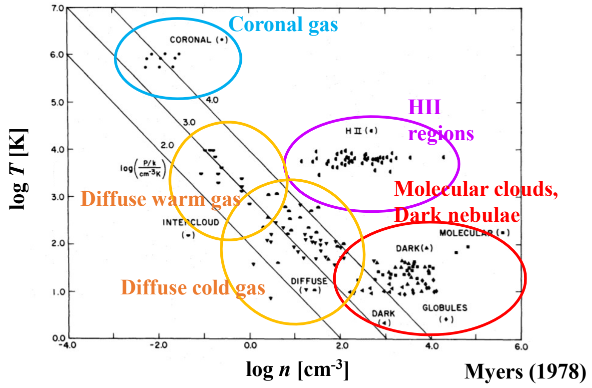
1. Division of Particle and Astrophysical Science, Nagoya University, Japan

2. The Research Center for Statistical Machine Learning, the Institute of Statistical Mathematics

Symposium on Theories, Methodologies and Applications for Large Complex Data, Tsukuba, Japan 30-31 Oct., 2025

1. Interstellar Medium (ISM)

1.1 Phase in ISM



Myers (1978)

Collaborators

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

Nanase HARADA (原田 ななせ), Kouichiro NAKANISHI (中西 康一郎)
National Astronomical Observatory of Japan

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

Yu OGANE (大金 有羽), Ryusei R. KANO (加納龍生), Hai-Xia MA (馬 海霞), Sena A. MATSUI (松井 瀬奈)
Division of Particle and Astrophysical Science, Nagoya University, Japan

Suchetha COORAY (クレ スチェータ)
Kavli Institute Particle Astrophysics and Cosmology, Stanford University, USA

Hiroma OKUBO (大久保 宏真)
School of Science and Engineering, University of Tsukuba, Japan

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

1.2 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

0. Background: What Are Galaxies?

A galaxy is a huge agglomeration of **stars, interstellar medium (ISM: gas+dust), and dark matter (DM),** a complex system with a complicated interaction between each component.

Spatial scales

Spatial scales of galaxies and star formation (SF) are some orders of magnitude different:

Galaxies ~ kpc
Star formation ~ a few pc (for molecular clouds)

However, global properties of galaxies and SF activity are mysteriously correlated in various aspects!

⇒ Meso-scale physics to connect the scales of a galaxy and SF should be explored.

Kennicutt-Schmidt (K-S) law

Stars form in molecular cores.

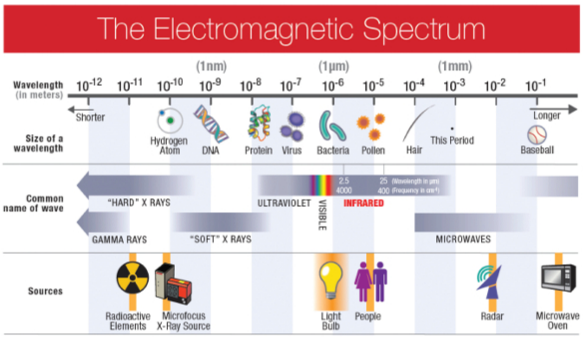
⇒ It is natural to suppose a relation between the star formation rate (SFR) and gas density. Schmidt (1959) proposed a relation

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

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

⇒ It is crucial to explore the properties of molecular clouds in star forming galaxies!

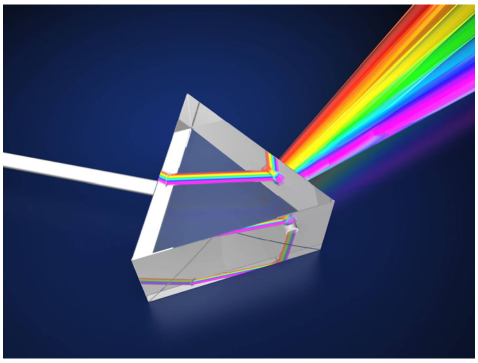
Electromagnetic waves



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

1.3 What does spectroscopy tell us?

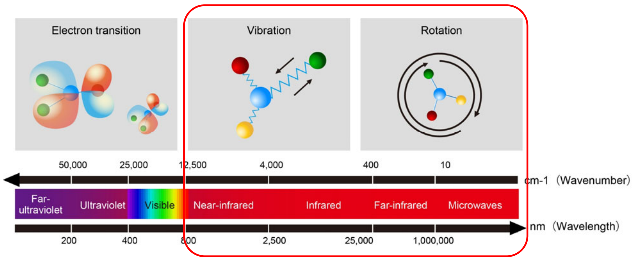
Spectroscopy



https://www.atascientific.com.au/spectrometry/

Quantum transition to spectral lines

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

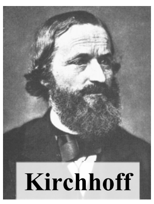


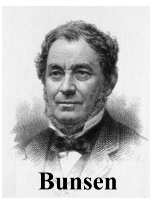
Information on molecules

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

1.3 What does spectroscopy tell us?

Kirchhoff and Bunsen showed that the emission lines corresponds to the absorption lines, which can be used for the identification of elements. Kirchhoff immediately pointed out that this leads to a very important application in astronomy.





http://www.hao.ucar.edu/public/education/sp/images/kirchhoff.html
https://en.wikipedia.org/wiki/Robert_Bunsen

2. High-Dimensional Statistical Analysis

2.1 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.

2. High-Dimensional Statistical Analysis

2.1 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, 2-dim spectral map such as integral field spectroscopy has this property.

2.2 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.

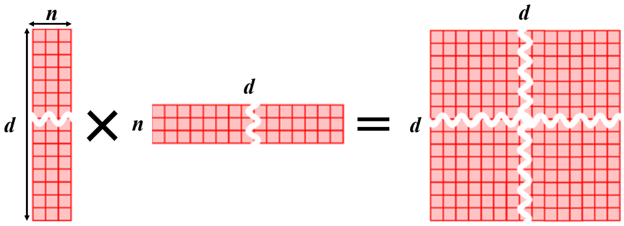
2.2 Geometric Representation

Dual representation of sample covariance matrix

When we draw a set of n samples from the parent population ($d > n$), $\vec{x}_1, \ldots, \vec{x}_n$.

The sample covariance matrix ($d \times d$) is $\tilde{S} = \frac{1}{n}\tilde{X}\tilde{X}^\top$,

$$\tilde{X} \equiv (x_1, x_2, \ldots, x_n)$$



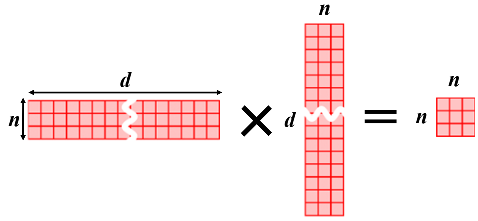
Note that this is a tremendously huge matrix!

2.2 Geometric Representation

Dual representation of sample covariance matrix

When we draw a set of n samples from the parent population ($d > n$), $\vec{x}_1, \ldots, \vec{x}_n$.

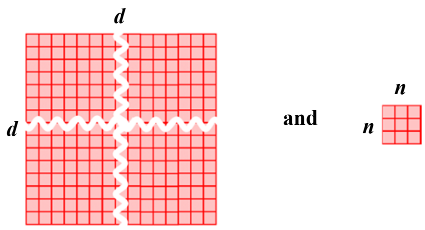
Consider a dual sample covariance matrix ($n \times n$), $\tilde{S}_{\mathrm{D}} = \frac{1}{n}\tilde{X}^\top\tilde{X}$



This can be handled much more easily!

Eigenvalues of the dual covariance matrix

When we draw a set of n samples from the parent population ($d > n$), $\vec{x}_1, \ldots, \vec{x}_n$.

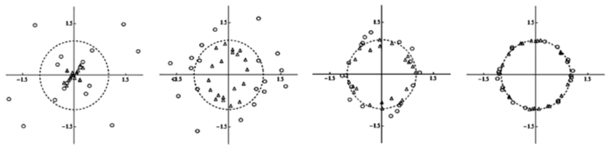


share the first n eigenvalues, i.e., the same important statistical information!

Unusual behavior of high-dimensional data: details

We can visualize the behavior of high-dimensional data vectors with dual representation. We omit all the mathematical details and jump onto the result.

1. **The population has a similar property with Gaussian**
   ⇒ The eigenvectors concentrate on a sphere!!



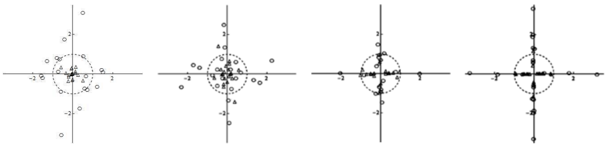
$d = 2$ $d = 20$ $d = 200$ $d = 2000$

Yata & Aoshima (2012)

Unusual behavior of high-dimensional data

We can visualize the behavior of high-dimensional data vectors with dual representation. We omit all the mathematical details and jump onto the result.

2. The population has a similar property with non-Gaussian
   ⇒ The eigenvectors concentrate on the axes!!

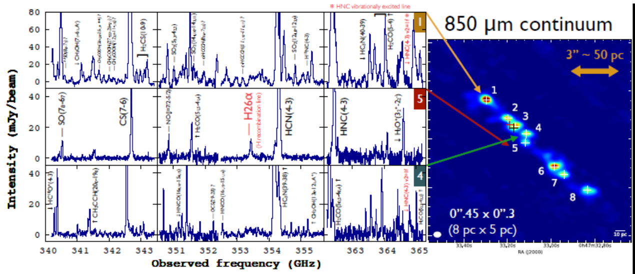


$d = 2$ $d = 20$ $d = 200$ $d = 2000$

Yata & Aoshima (2012)

Rich in molecular lines

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

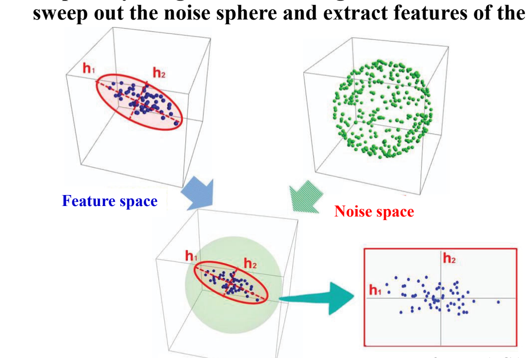


ALMA Band7 spectra

Ando et al. (2017)

High-dimensional PCA

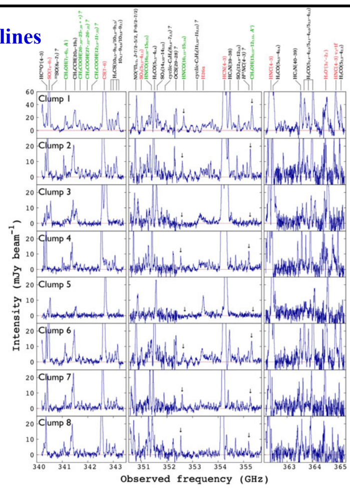
A specially designed PCA, the high-dimensional PCA, can sweep out the noise sphere and extract features of the data.



Feature space embedded in a noise space

Aoshima (2012)

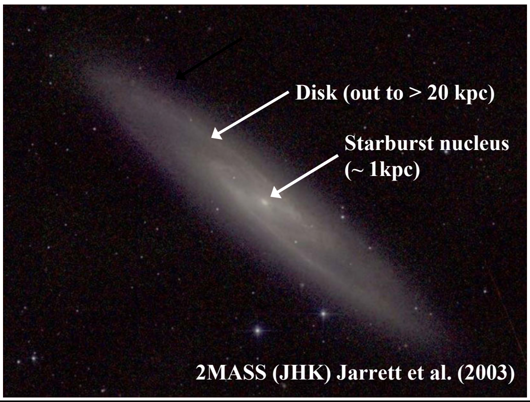
Rich in molecular lines



Ando et al. (2017)

2.3 Actual data: ALMA data cube of NGC253

NGC 253: prototypal starburst



2.4 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

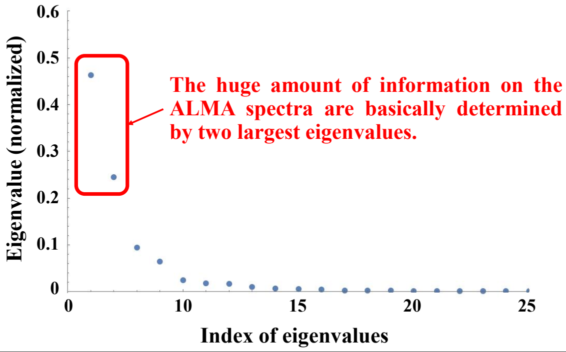
- Too much information on spectra.
- Too large variety of spectral lines compared to n.

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

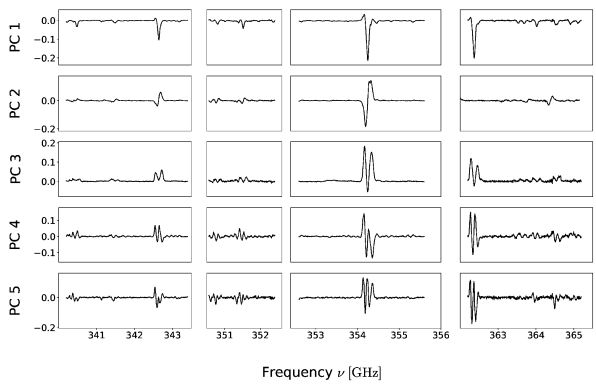
3. Analysis of Starburst Region in NGC253

3.1 Analysis of Raw Data

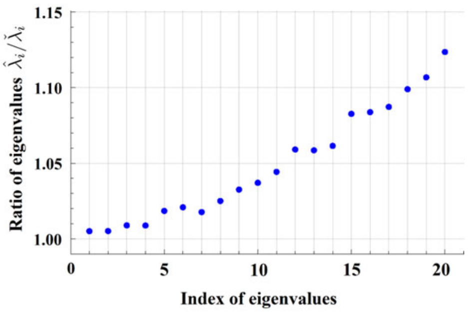
Eigenvalues of the PCA (contribution)



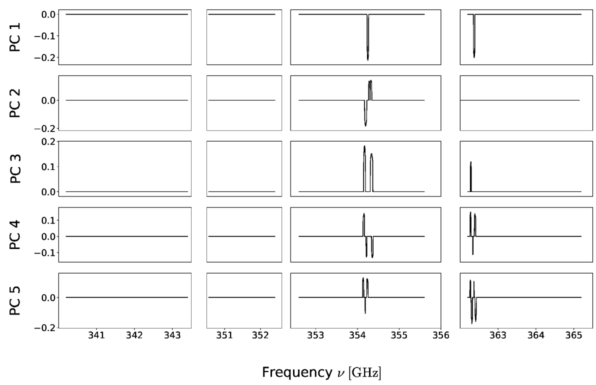
Eigenspectra for PC1-5 from NRPCA



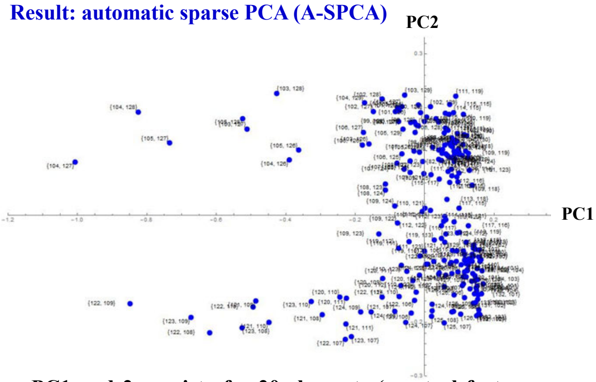
Ratio of eigenvalues obtained by traditional and high-dimensional PCAs (raw data)



Eigenspectra for PC1-5 from A-SPCA

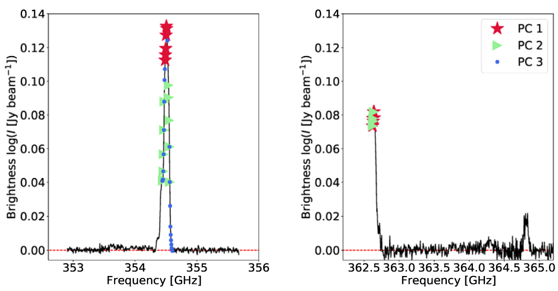


Result: automatic sparse PCA (A-SPCA)



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



Now PC1 more clearly represents the total intensity, and PC2 and 3 represent smaller-scale velocity structures.
The responsible features are extracted by the A-SPCA (Yata & Aoshima 2024).

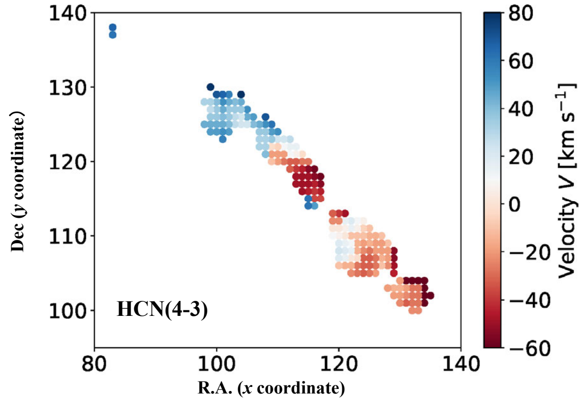
Spectral features corresponding to PC1 and PC2

https://en.wikipedia.org/wiki/Hydrogen_cyanide

https://en.wikipedia.org/wiki/Hydrogen_isocyanide

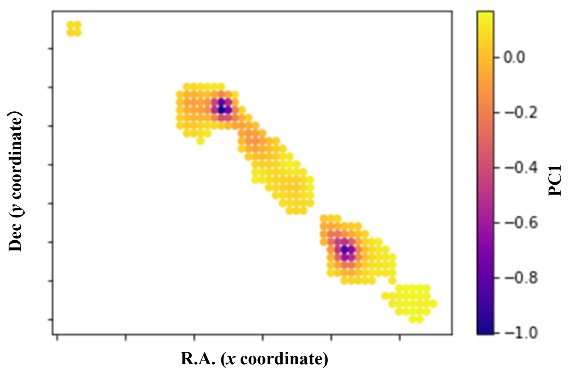
HCN (hydrogen cyanide, as known as the hydrocyanic acid) and HNC (hydrogen isocyanide) are linear molecules, which have a quantum mechanical transition corresponding to the rotation states.

Velocity field of the systemic rotation

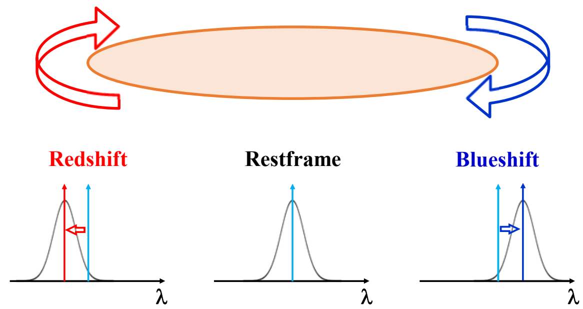


⇒ Doppler shift correction to remove the systemic rotation.

Spatial map of PC1



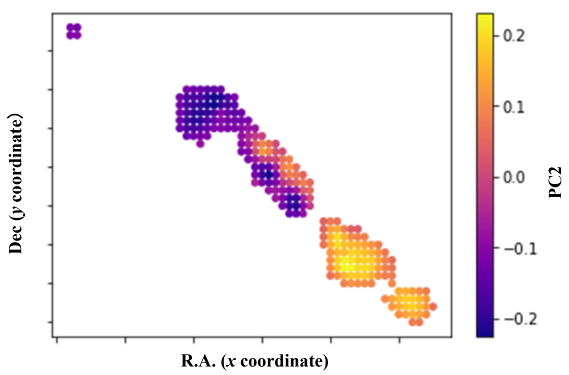
Systemic rotation and Doppler shift



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

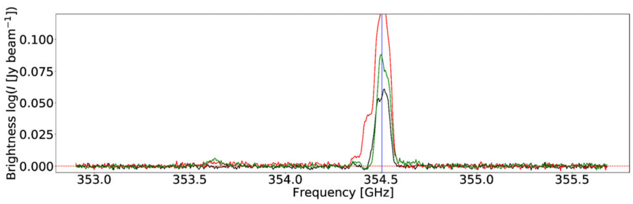
PC2 beautifully describes the Doppler shift!

Spatial map of PC1 and PC2



3.3 Main analysis

Doppler shift correction

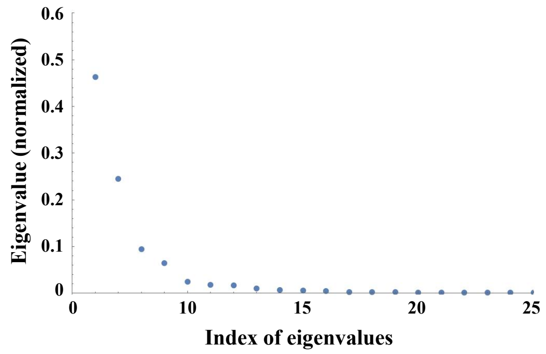


Takeuchi et al. (2021)

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.

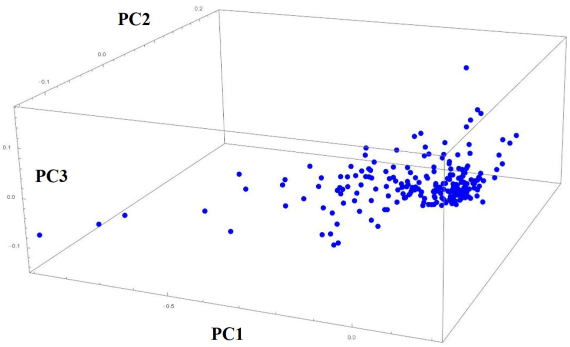
Due to this correction, the final data dimension is $d = 1971$.

Eigenvalues of the NGC253 before Doppler correction



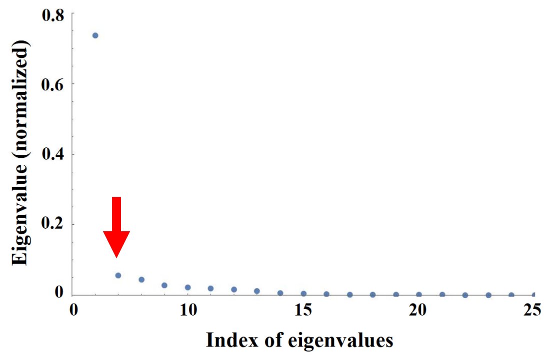
Takeuchi et al. (2024)

PC1, PC2, and PC3 from sparse PCA



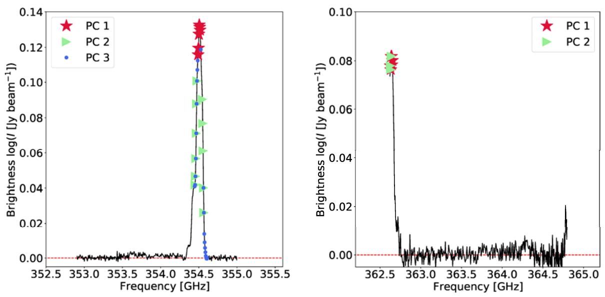
Takeuchi et al. (2024)

Eigenvalues of the NGC253 after Doppler correction



Takeuchi et al. (2024)

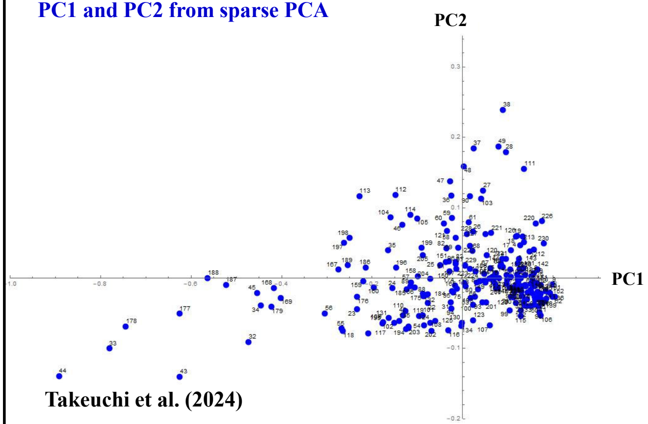
Responsible spectral features for PC1, PC2 and PC3



Takeuchi et al. (2024)

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

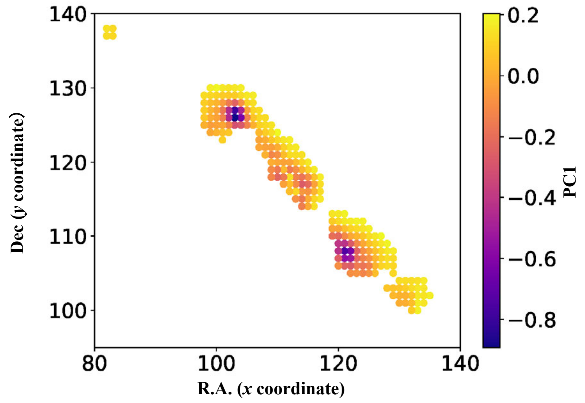
PC1 and PC2 from sparse PCA



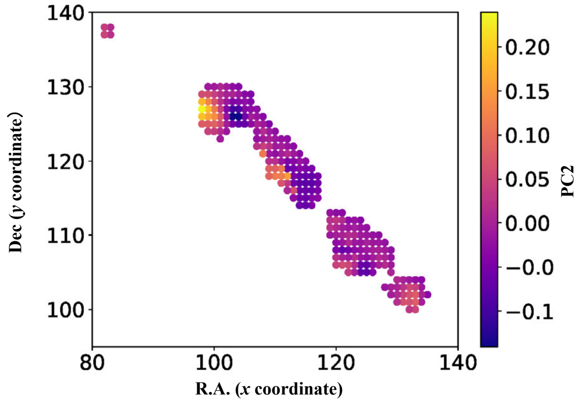
Takeuchi et al. (2024)

Butterfly-like pattern completely disappeared.

Spatial map of PC1 after Doppler correction



Spatial map of PC2 after Doppler correction



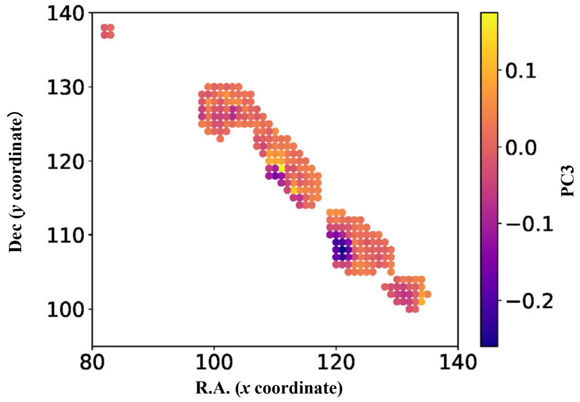
What do we see from the Doppler-corrected map?

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.

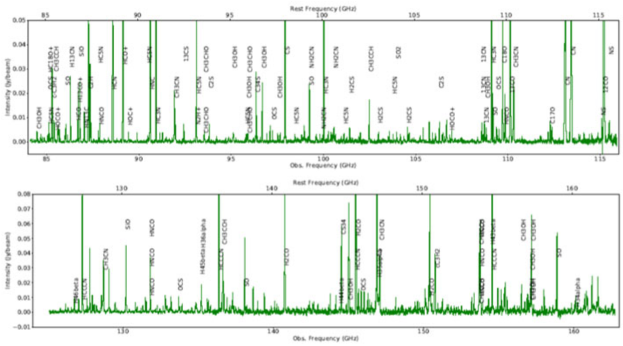
Spatial map of PC3 after Doppler correction



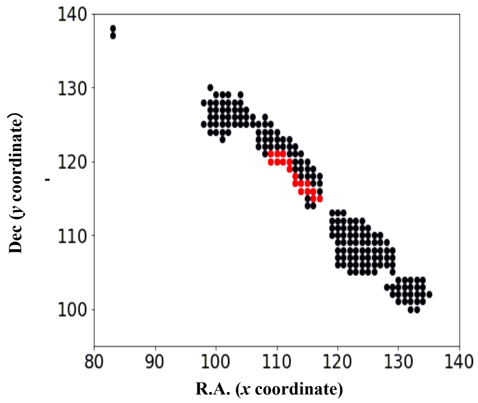
4 Further Analysis with New Data

4.1 New data: ARCHEMI

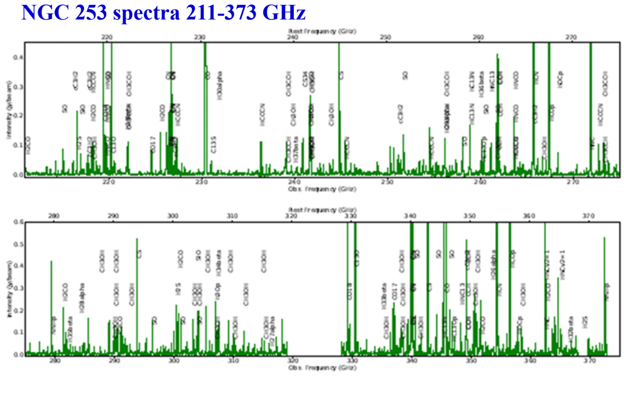
NGC 253 spectra 85-163 GHz



Anomaly regions in the velocity field



NGC 253 spectra 211-373 GHz



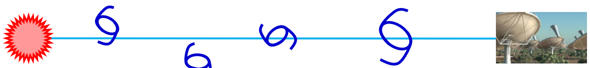
Now, much higher-quality data have been obtained. Extracted characteristics would not be necessarily linear.

4.2 Early result: finding erroneous data range

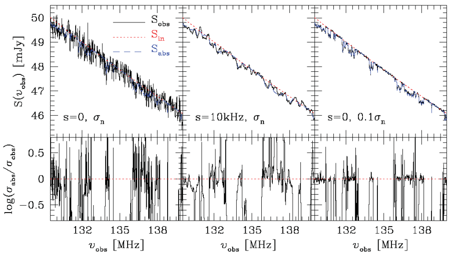
(Presented only on site)

5. Analysis of the HI Forest

5.1 What is the HI forest?



Quasar Observer



Ciardi et al. (2013)

5.2 Prospect and difficulty in the analysis of HI forest

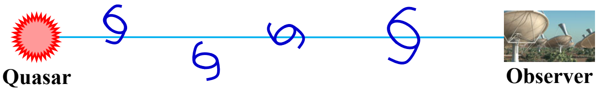
Prospect

Quasar Observer

- The HI forest carry the information on the spatial distribution of primordial galaxies. This provides a very important clue to the formation of first galaxies.
- The HI forest absorption line systems have evolved into galaxies at later epochs of the Universe. Their evolution might be reflected to the absorption lines.

5.2 Prospect and difficulty in the analysis of HI forest

Difficulty



Quasar Observer

The background quasars are very rare.

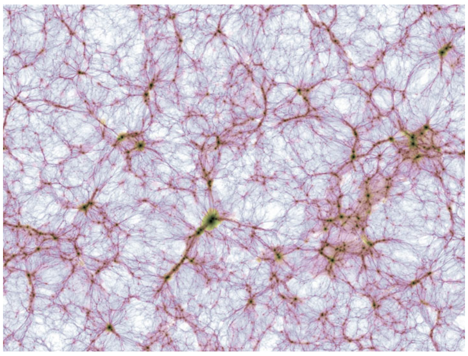
Even by the next-generation radio observational facility Square Kilometre Array (SKA), only a few tens of quasars are expected., while the absorption lines are numerous.

⇒ HDLSS data!

We constructed a new analysis method based on the high-dimensional statistical analysis.

5.3 Basic analysis with cosmological simulation

Illustris TNG simulation



https://www.tng-project.org/media/

5.3 Basic analysis with cosmological simulation

Quantification of the spatial distribution of HI gas

Absorption line frequency is described as

$$\nu_{abs} = \frac{\nu_{21\,\mathrm{cm}}}{1 + z_{abs}}$$

i.e., the restframe frequency is shifted by cosmological redshift.

After some conversion of the observable, we have a data matrix of cosmic density fluctuation $\delta^{(j)}(z_i)$ ($d \times n$) as

$$\vec{X} = \begin{pmatrix} \delta^{(1)}(z_1) & \delta^{(2)}(z_1) & \dots & \delta^{(n)}(z_1) \\ \delta^{(1)}(z_2) & \delta^{(2)}(z_2) & \dots & \delta^{(n)}(z_2) \\ \vdots & & \ddots & \\ \delta^{(1)}(z_d) & \delta^{(2)}(z_d) & \dots & \delta^{(n)}(z_d) \end{pmatrix}$$

5.3 Basic analysis with cosmological simulation

Quantification of the spatial distribution of HI gas

Two-point correlation function

$$\xi(z_1, z_2) \equiv \langle \delta(z_1)\delta(z_2) \rangle = \xi(|z_1 - z_2|)$$

From the data, we have a set of correlation functions in the form of the covariance matrix Ξ

$$\Xi = \begin{pmatrix} \xi^{(1)}(0) & \xi^{(2)}(|z_1 - z_2|) & \cdots & \xi^{(n)}(|z_1 - z_d|) \\ \xi^{(1)}(|z_2 - z_1|) & \xi^{(2)}(0) & \cdots & \xi^{(n)}(|z_2 - z_d|) \\ \vdots & & \ddots & \\ \xi^{(1)}(|z_d - z_1|) & \xi^{(2)}(|z_d - z_2|) & \cdots & \xi^{(n)}(0) \end{pmatrix}$$

This contains fundamental information on the cosmic matter density field.

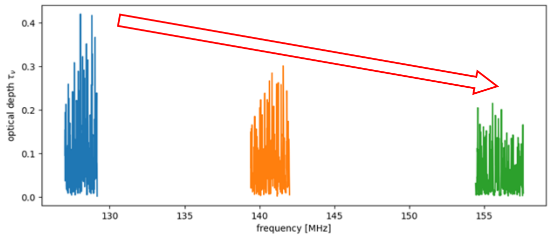
6. Summary

1. **Spectroscopic mapping and similar methods are fundamentally important to reveal the ISM physics, but the data are high-dimensional low sample size.**
2. **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 $d = 2228$.**
3. **NRPCA and A-SPCA can work very well to extract physical information from the HDLSS spectral map.**
4. **High-dimensional PCA also works as a powerful tool to find a part of the data with errors.**

5.3 Basic analysis with cosmological simulation

Quantification of the spatial distribution of HI gas

High-dimensional PCA of the HI forest absorption lines



Cang et al. (2025)

We indeed observe the effect of the cosmic evolution of the absorption line system.

6. Summary

5. **High-dimensional methods are also very useful for the analysis of neutral hydrogen absorption line systems on distant QSO spectra. This is a promising method for future cosmological studies.**

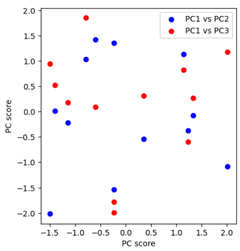
If you are interested in details, see

Takeuchi, T. T., et al. 2024a, ApJS, 271, 44
Takeuchi, T. T., et al. 2024b, Proceedings of the Institute of Statistical Mathematics (統計数理), 72(2), 273
Cang, J.-S., et al. 2025, Advancing Astrophysics: Preparing for Science with the SKAO, in press

5.3 Basic analysis with cosmological simulation

Quantification of the spatial distribution of HI gas

High-dimensional PCA of the HI forest absorption lines



Cang et al. (2025)

Plot of the PC scores (PC1 vs. PC2 and PC1 vs. PC3).
We try to quantify the evolution and specify the key factor to characterize the evolution.

Advertisement