High-Dimensional Statistics in Astrophysics and its Perspective

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 Theories, Methodologies and Applications for Large Complex Data, Tsukuba, Japan 4-6 Dec., 2024

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

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

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

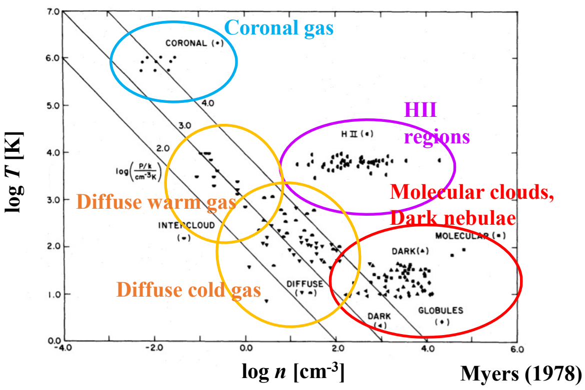
Suchetha COORAY (クレ スチェータ)
Kavli Institute Particle Astrophysics and Cosmology, Stanford University, USA

Aina May SO (曹 愛奈), Wen SHI (施 文), 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

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

1. Interstellar Medium (ISM)

1.1 Phase in ISM



Myers (1978)

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

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.

Star formation in the ISM

Hydrogen is overwhelmingly dominant among others.
⇒ Molecular clouds consist of hydrogen molecules (H_2).

Molecules are not only formed but also dissociated and turn back into atoms by an ultraviolet (UV) radiation.

The layer on which the formation and dissociation of H_2 balance forms the surface boundary of a molecular cloud.

⇒ Since UV is shielded by H_2, the center of a molecular cloud can become cooler and cooler, finally to form a very dense molecular core, where stars form.

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

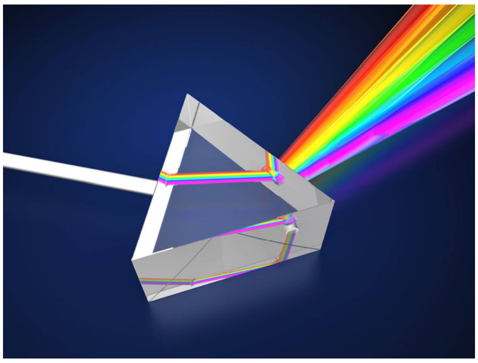
$$\mathrm{SFR} \propto \rho^n.$$

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!

1.3 What does spectroscopy tell us?

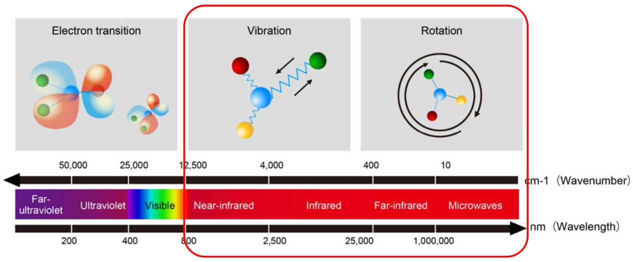
Spectroscopy



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/

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.

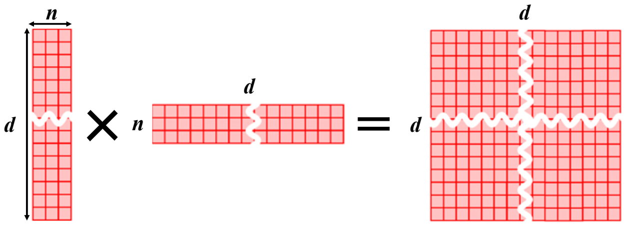
2.2 Geometric Representation

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

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

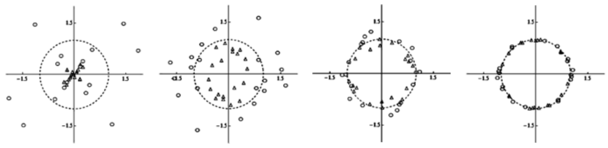


Note that this is a tremendously huge matrix!

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 data converge on a sphere!!



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

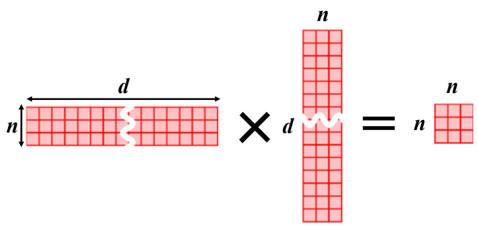
Yata & Aoshima (2012)

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 ($n \times n$), $\tilde{S}_{\mathrm{D}} = \frac{1}{n}\tilde{X}^\top\tilde{X}$

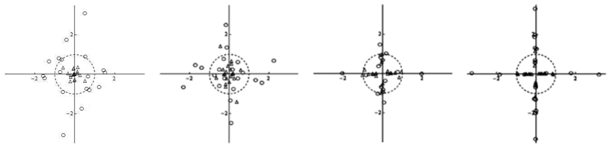


This can be handled much more easily!

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 data converge on the axes!!

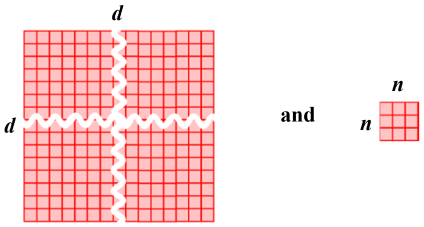


$d = 2$ $d = 20$ $d = 200$ $d = 2000$

Yata & Aoshima (2012)

Eigenvalues of the dual covariance matrix

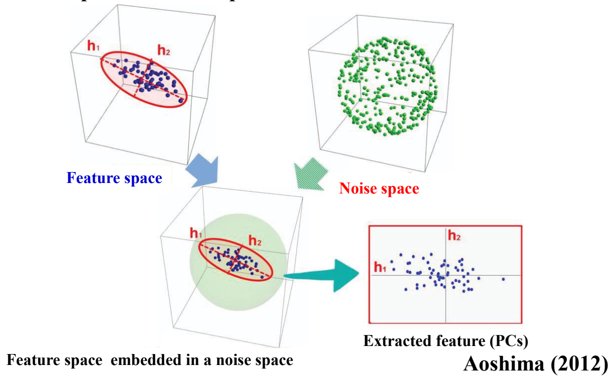
When we draw a set of n samples from the parent population ($d > n$), $\vec{x}_1, \ldots, \vec{x}_n$.



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

High-dimensional PCA

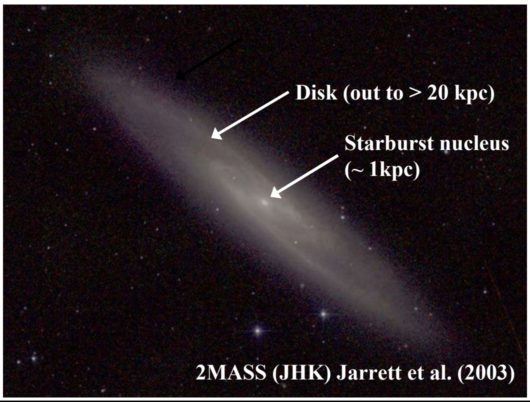
A specially designed PCA, the high-dimensional PCA, can sweep out the noise sphere and extract features of the data.



Aoshima (2012)

2.3 Actual data: ALMA data cube of NGC253

NGC 253: prototypal starburst



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

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)

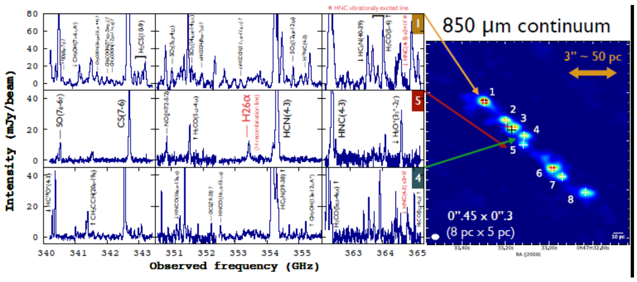
Problems from astrophysical side

- 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.

Rich in molecular lines

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



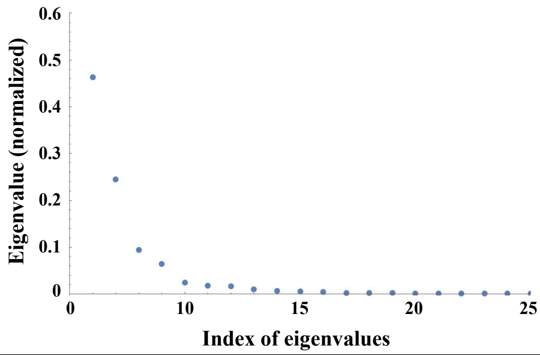
ALMA Band7 spectra

Ando et al. (2017)

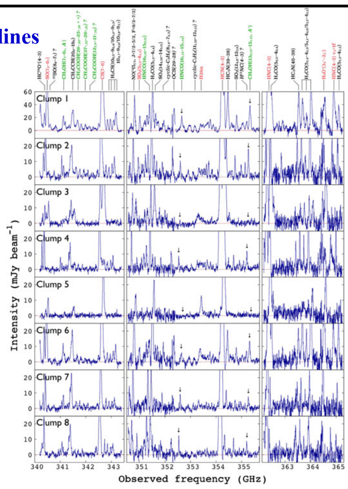
3. Analysis of Starburst Region in NGC253

3.1 Analysis of Raw Data

Eigenvalues of the PCA (contribution)



Rich in molecular lines

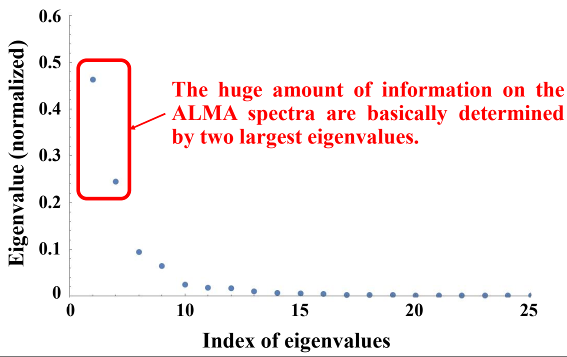


Ando et al. (2017)

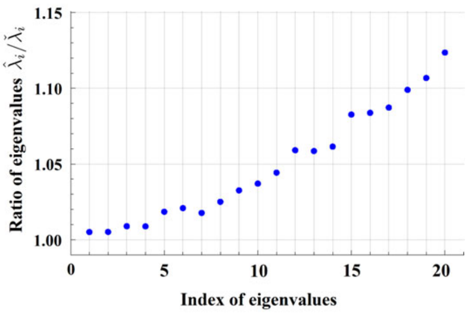
3. Analysis of Starburst Region in NGC253

3.1 Analysis of Raw Data

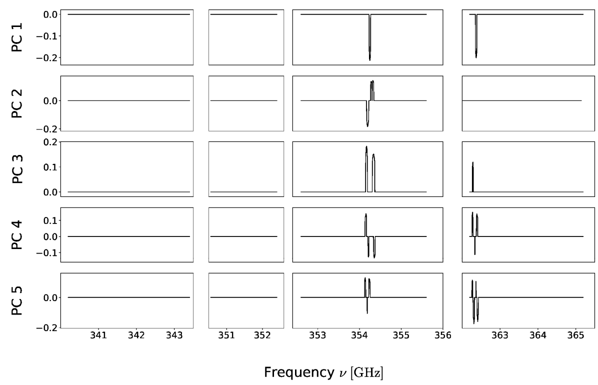
Eigenvalues of the PCA (contribution)



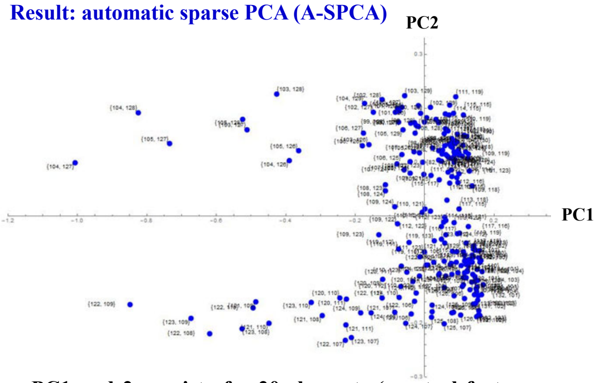
Ratio of eigenvalues obtained by traditional and high-dimensional PCAs (raw data)



Eigenspectra for PC1-5 from A-SPCA

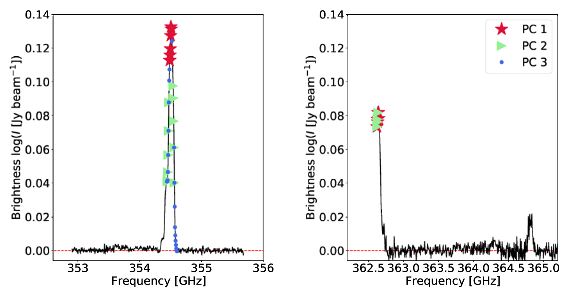


Result: automatic sparse PCA (A-SPCA)



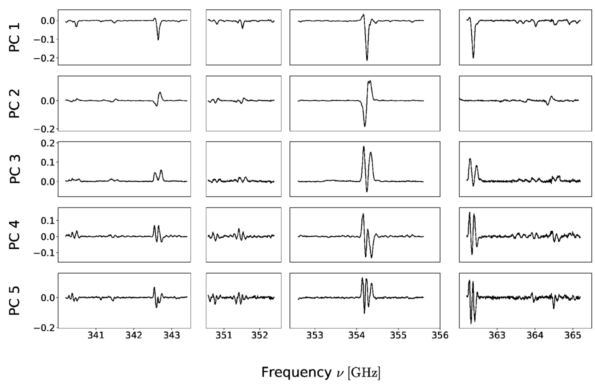
PC1 and 2 consist of ~ 20 elements (spectral features on the resolution units). The key features may be reduced only to a few to several lines!

Responsible spectral features for PC1, PC2 and PC3

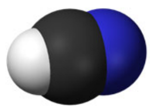


Now PC1 more clearly represents the total intensity, and PC2 and 3 represent smaller-scale velocity structures.
The responsible features are extracted by the A-SPCA (Yata & Aoshima 2024).

Eigenspectra for PC1-5 from NRPCA

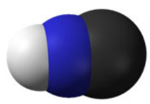


Spectral features corresponding to PC1 and PC2



H—C≡N

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

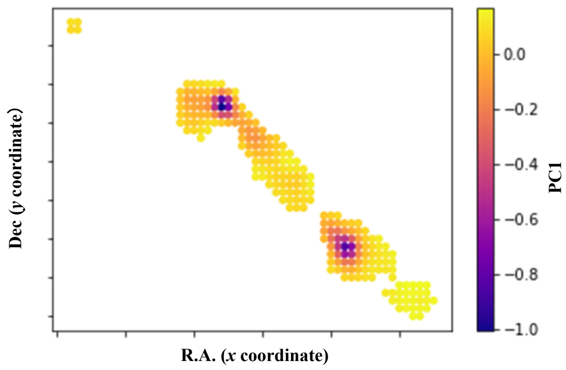


$\mathrm{H{-}\overset{+}{N}{\equiv}C^{-}}$

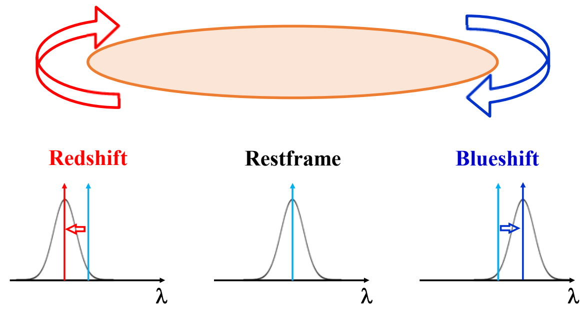
https://en.wikipedia.org/wiki/Hydrogen_isocyanide

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.

Spatial map of PC1

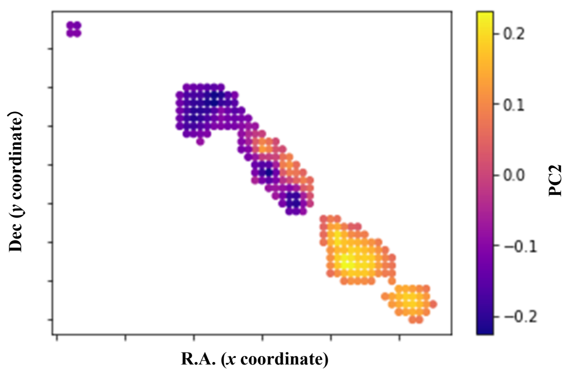


Systemic rotation and Doppler shift



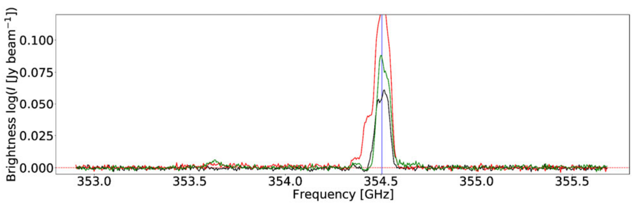
If the system is rotating as a whole, the observed wavelength is affected by the Doppler shift. PC2 beautifully describes the Doppler shift!

Spatial map of PC1 and PC2



3.3 Main analysis

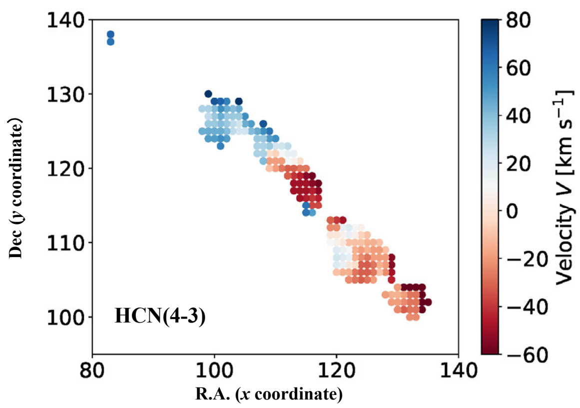
Doppler shift correction



Takeuchi et al. (2021)

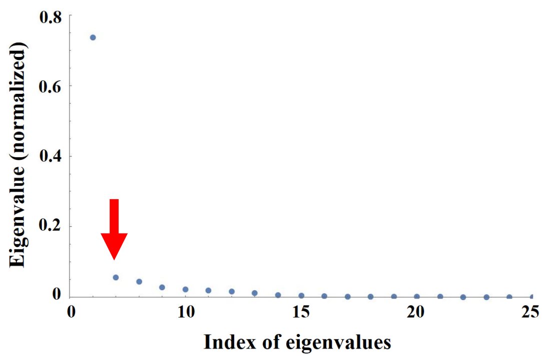
We estimated the peculiar velocity field (mainly due to the systemic rotation of the central region of NGC253) by averaging the results from HCN(4-3), HNC(4-3) and CS(7-6) lines, and corrected the Doppler shift.
Due to this correction, the final data dimension is $d = 1971$.

Velocity field of the systemic rotation



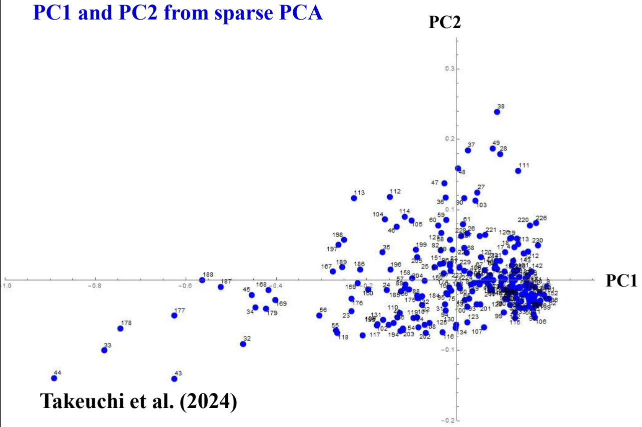
⇒ Doppler shift correction to remove the systemic rotation.

Eigenvalues of the NGC253 after Doppler correction



Takeuchi et al. (2024)

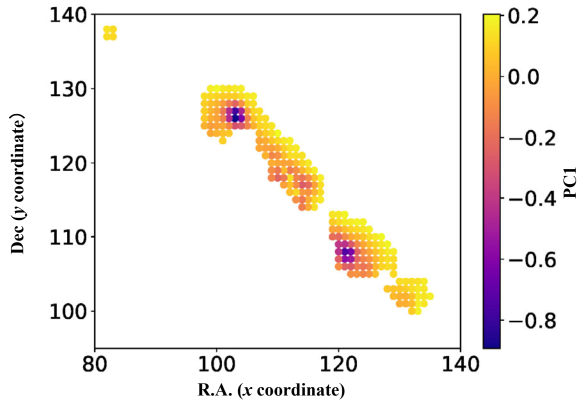
PC1 and PC2 from sparse PCA



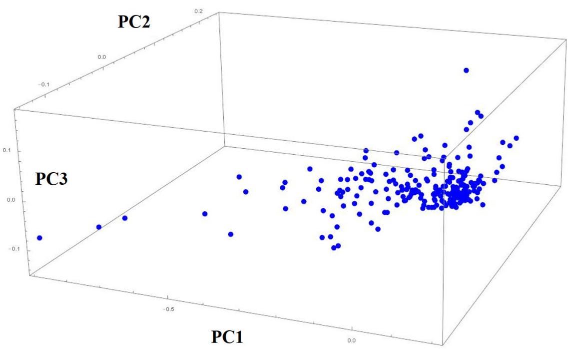
Takeuchi et al. (2024)

Butterfly-like pattern completely disappeared.

Spatial map of PC1 after Doppler correction

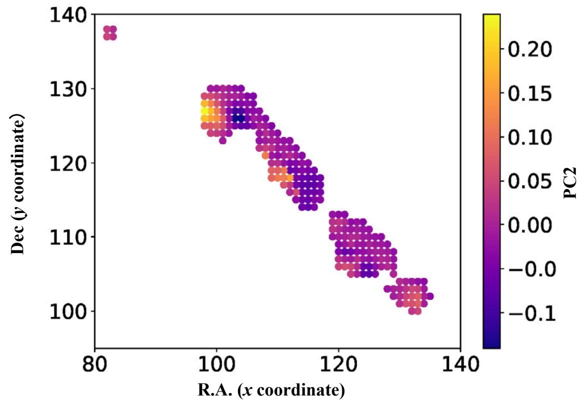


PC1, PC2, and PC3 from sparse PCA

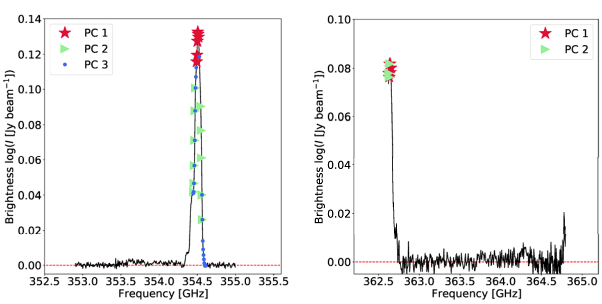


Takeuchi et al. (2024)

Spatial map of PC2 after Doppler correction



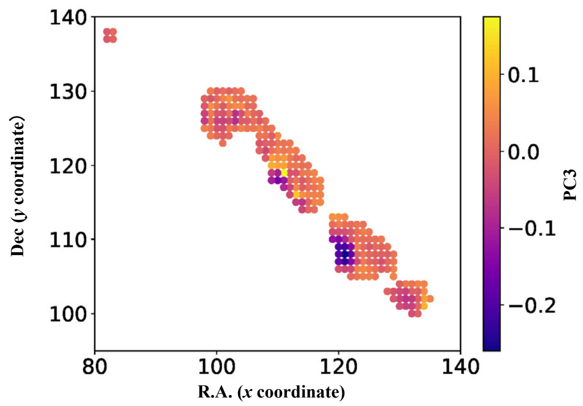
Responsible spectral features for PC1, PC2 and PC3



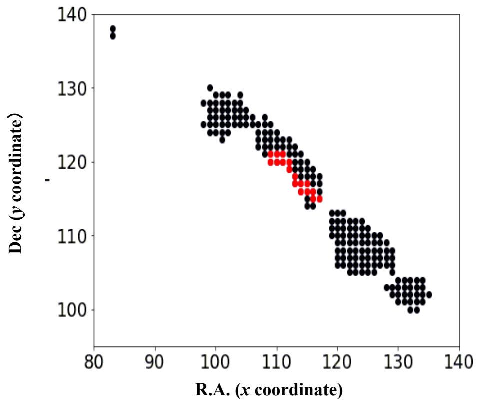
Takeuchi et al. (2024)

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

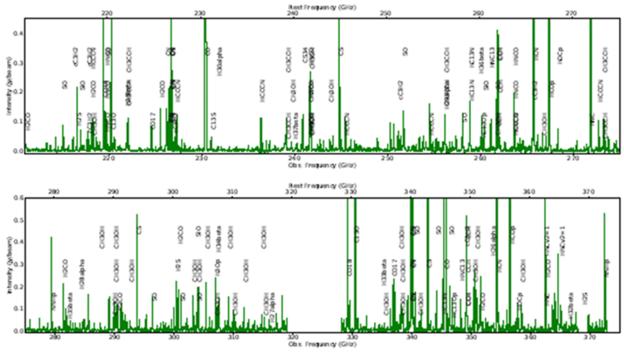
Spatial map of PC3 after Doppler correction



Anomaly regions in the velocity field



NGC 253 spectra 211-373 GHz



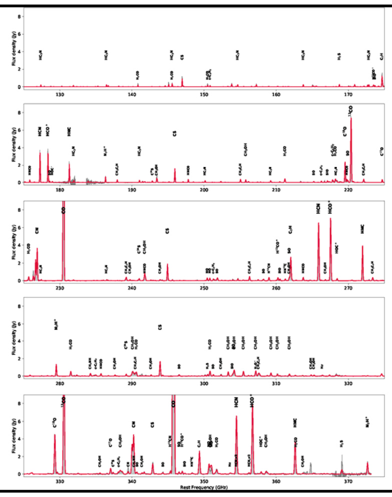
Now, much higher-quality data have been obtained. Extracted characteristics would not be necessarily linear.

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)

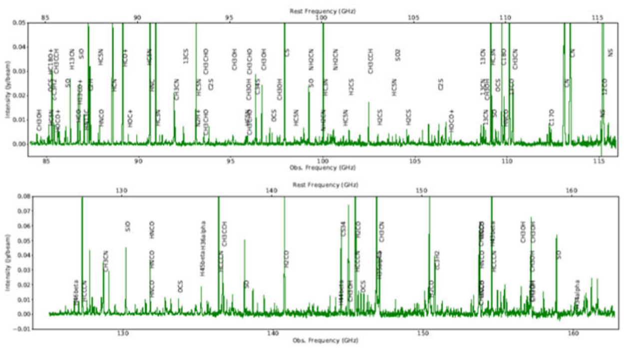
Indeed the outflow phenomenon is mainly delineated by PC3.

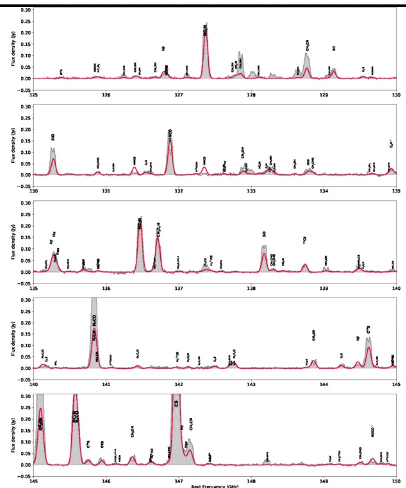


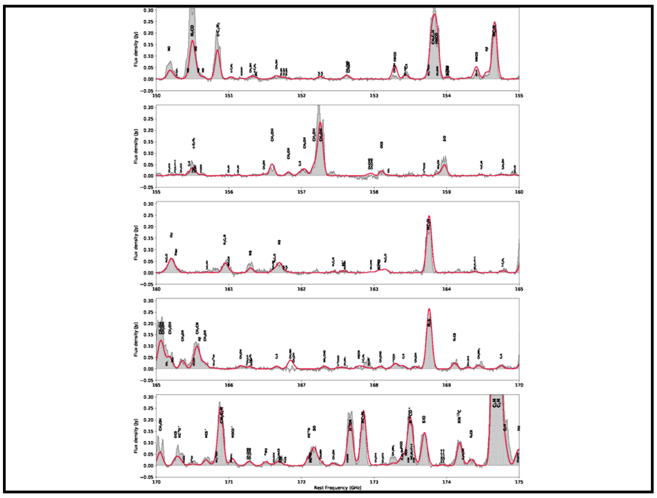
4 Further Analysis toward New Data

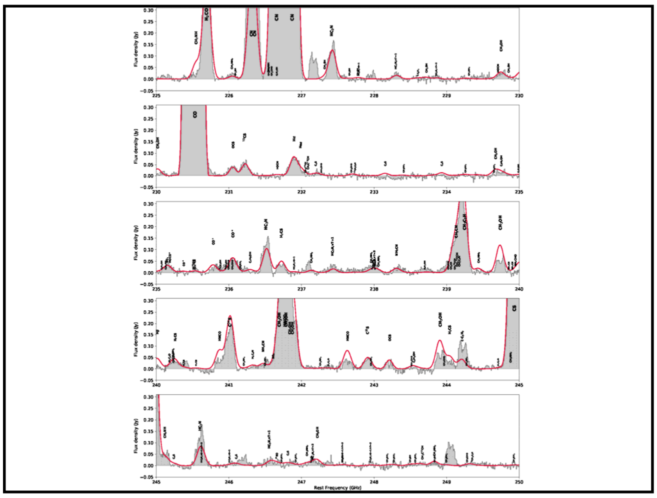
4.1 New data: ARCHEMI

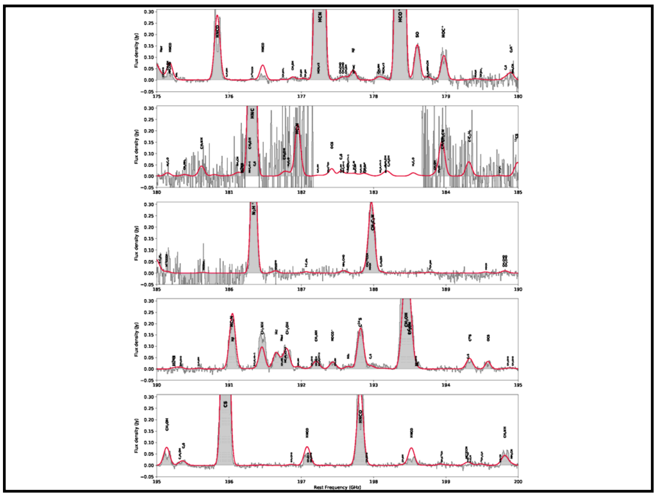
NGC 253 spectra 85-163 GHz

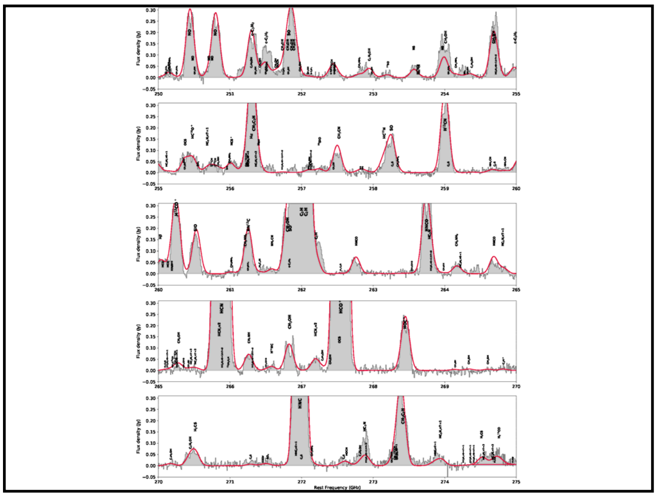


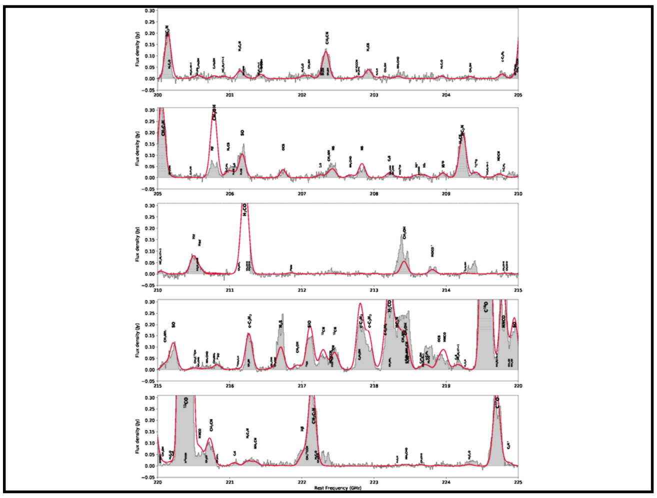


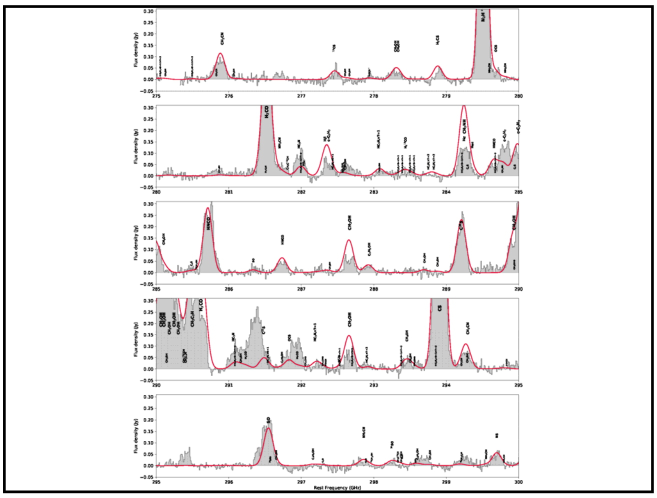


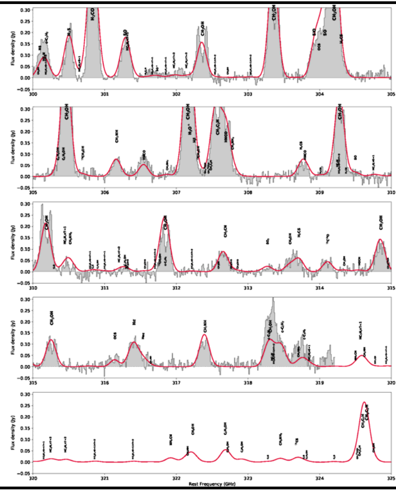












4.2 PCA in feature space: kernel PCA

Kernel trick: how to make PCA nonlinear

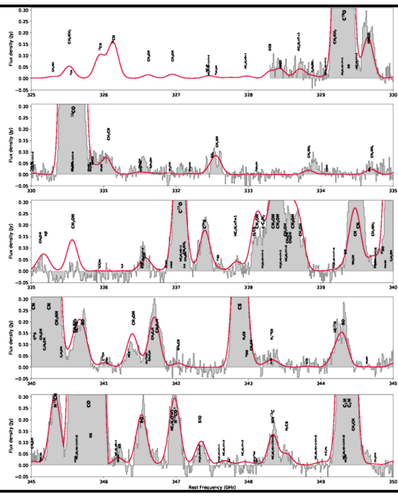
Suppose that instead of using the points $\mathbf{x}_i$ as is, we wanted to go to some different feature space $\phi(\mathbf{x}_i) \in \mathbb{R}^N$.

For example, using polar coordinates, instead of cartesian coordinates, would help us deal with a circle.

In the higher-dimensional space, we can then do PCA.

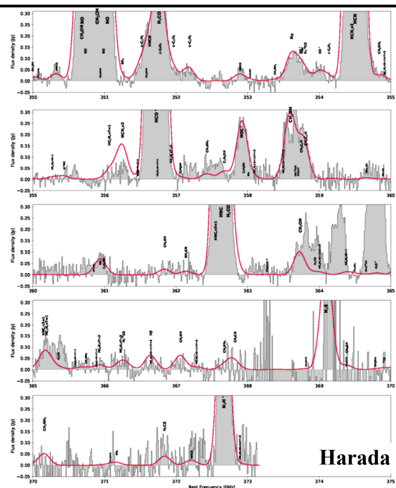
The result will be nonlinear in the original data space.

Possible problem is its intrinsic difficulty in interpretability. We continue to explore its efficiency.



5. 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. **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.**



Harada et al. (2024)

5. Summary

4. **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.**
5. **Kernel PCA is a powerful tool to characterize nonlinear relations in the data. It can provide us with supplementary information to the linear PCA, but since the interpretation is not easy, we need to explore its potential.**

If you are interested in details, see

Takeuchi, T. T., et al. 2024a, ApJS, 271, 44

Takeuchi, T. T., et al. 2024b, Toukei Suuri, in press (in Japanese)