High-Dimensional Statistical Analysis of Interstellar and Intergalactic Matter

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

First Stars and First Galaxies 2023, Sapporo, 20-22 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

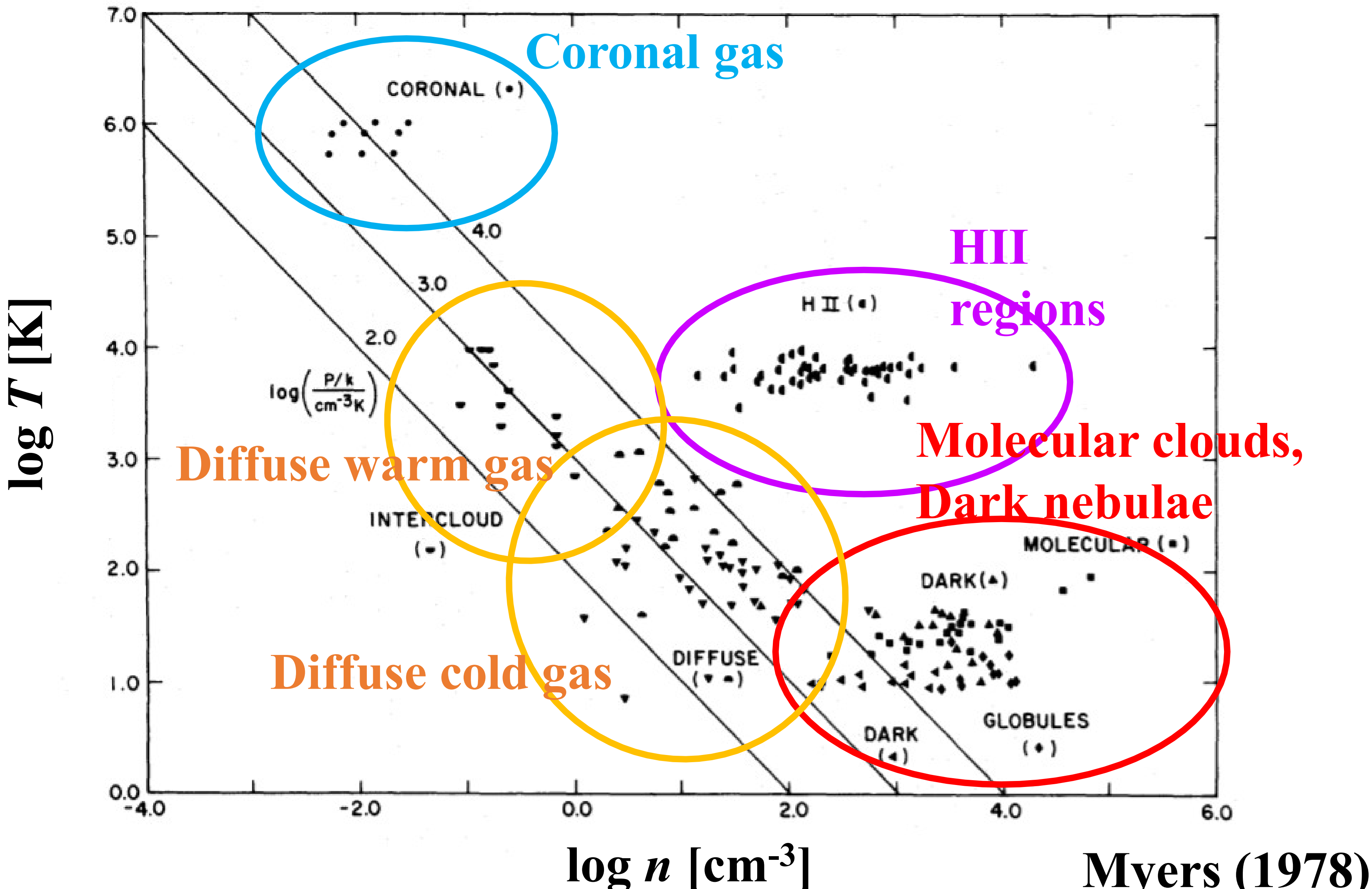
Kohji YOSHIKAWA (吉川 耕司)
Center for Computational Sciences, University of Tsukuba, Japan

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. 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

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. 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 $d/n \to 0$**

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

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

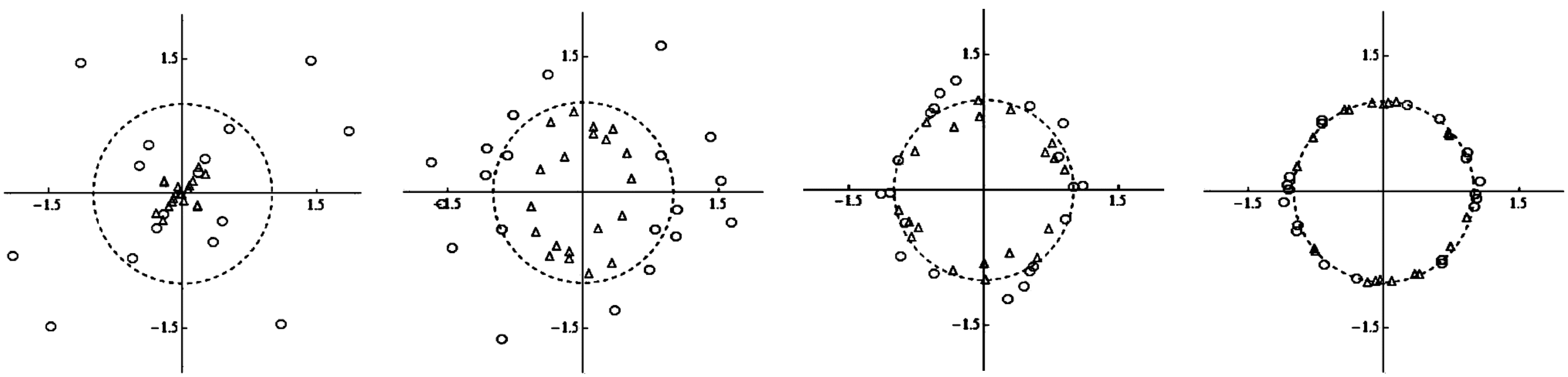
$$\| \bar{\vec{x}} - \vec{\mu} \| \xrightarrow{\text{P}} \infty$$

This striking property is referred to as the strong inconsistency.

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!!



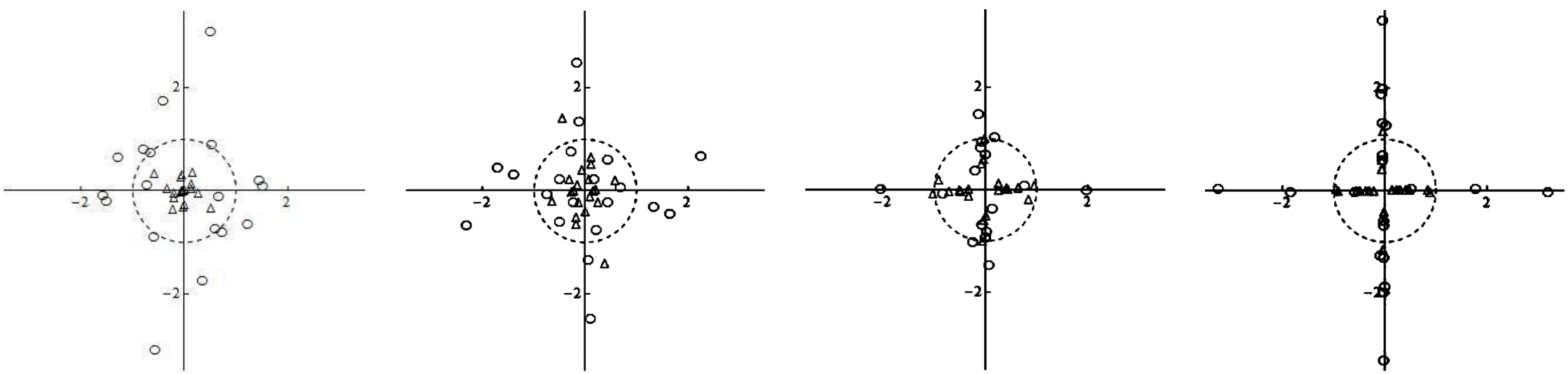
$d = 2$ $d = 20$ $d = 200$ $d = 2000$

Yata & Aoshima (2012)

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!!

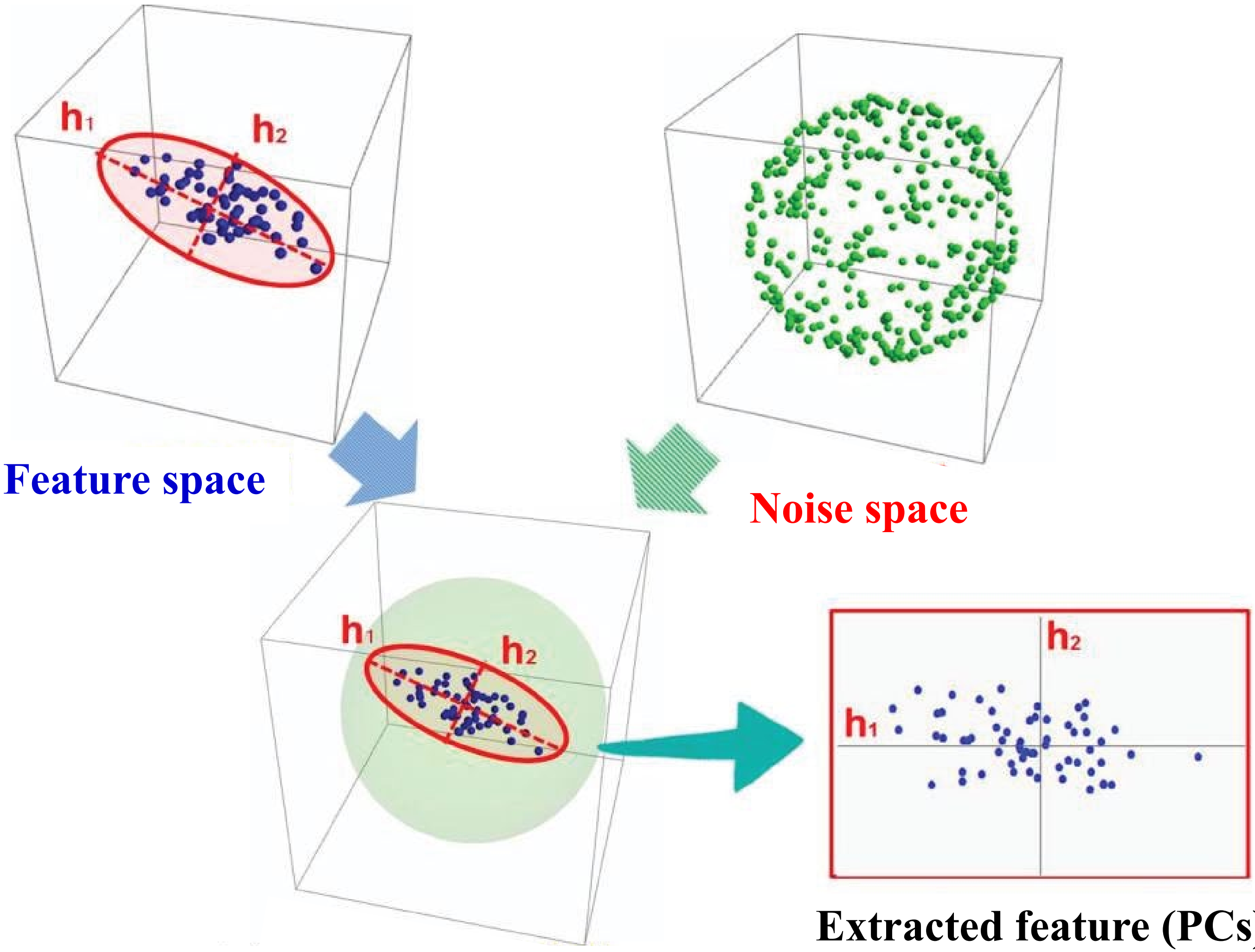


$d = 2$ $d = 20$ $d = 200$ $d = 2000$

Yata & Aoshima (2012)

High-dimensional PCA

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

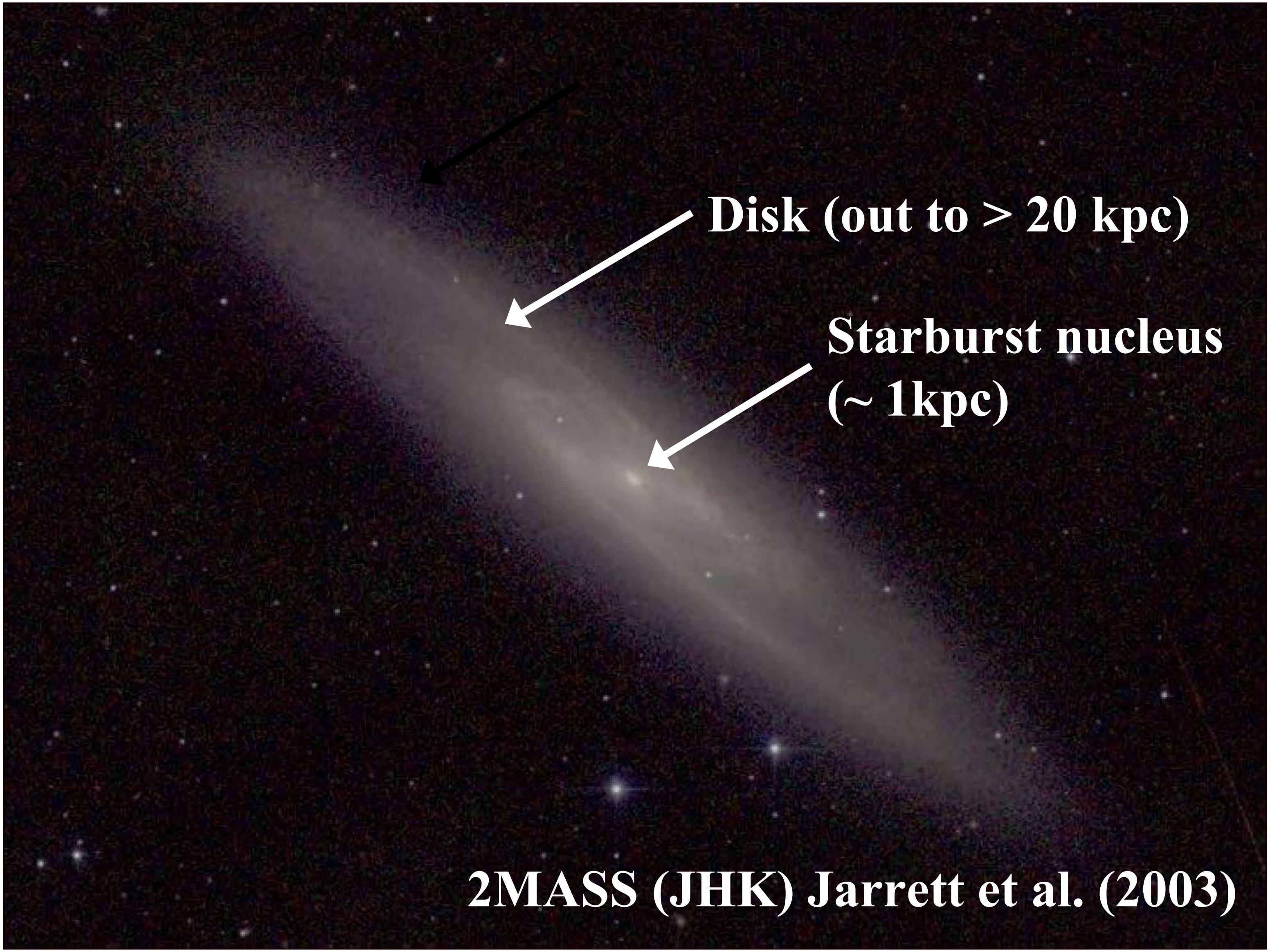


Feature space embedded in a noise space

Aoshima (2012)

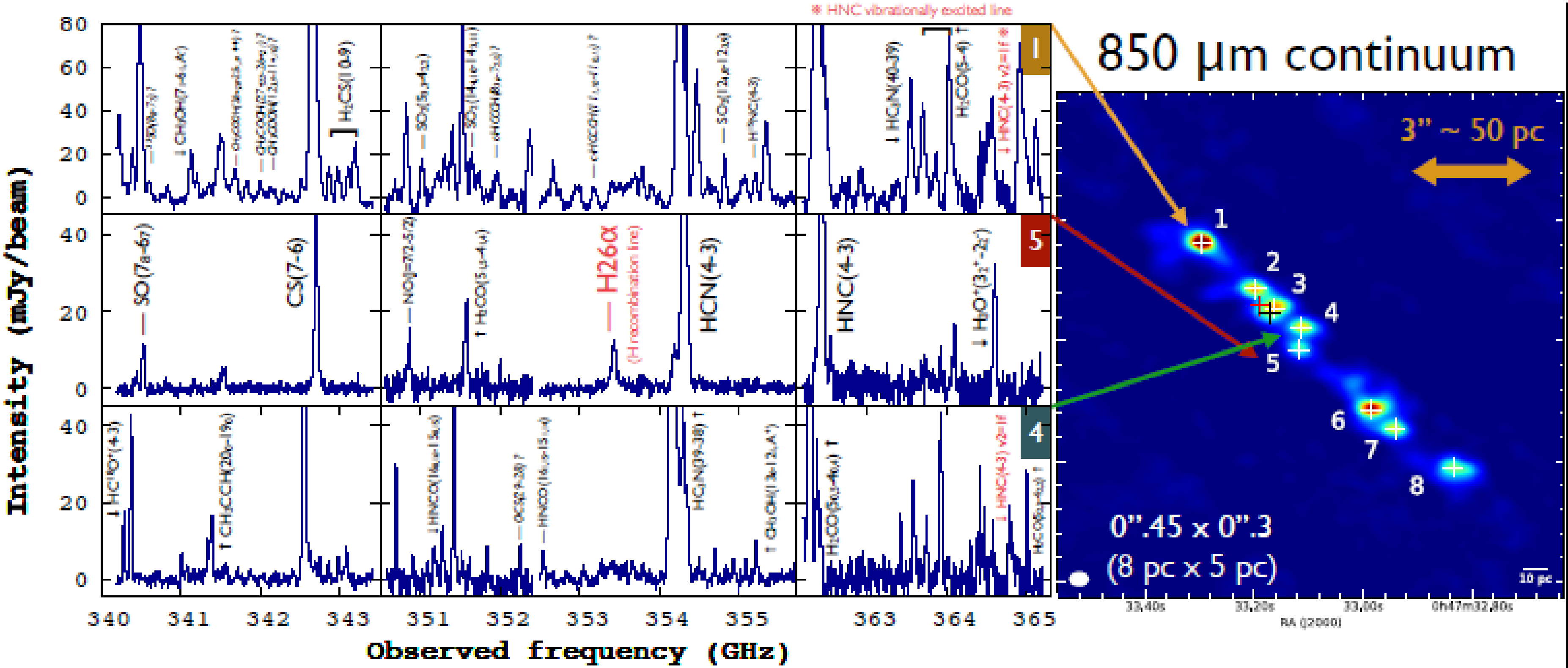
2.3 Actual data: ALMA data cube of NGC253

NGC 253: prototypal starburst



Rich in molecular lines

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



ALMA Band7 spectra

Ando et al. (2017)

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

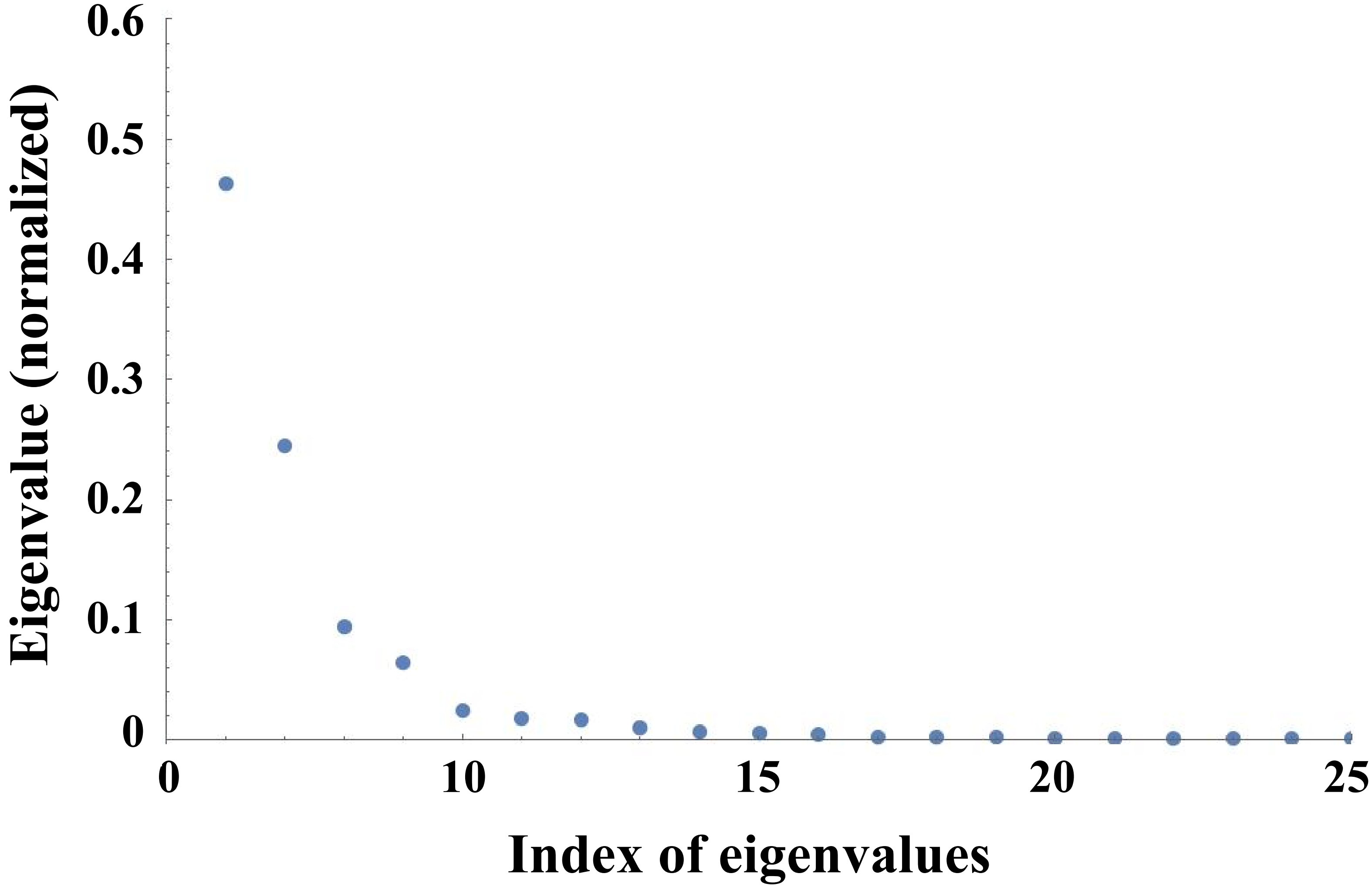
- **Too much information on spectra.**
- **Too large variety of spectral lines compared to n.**

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

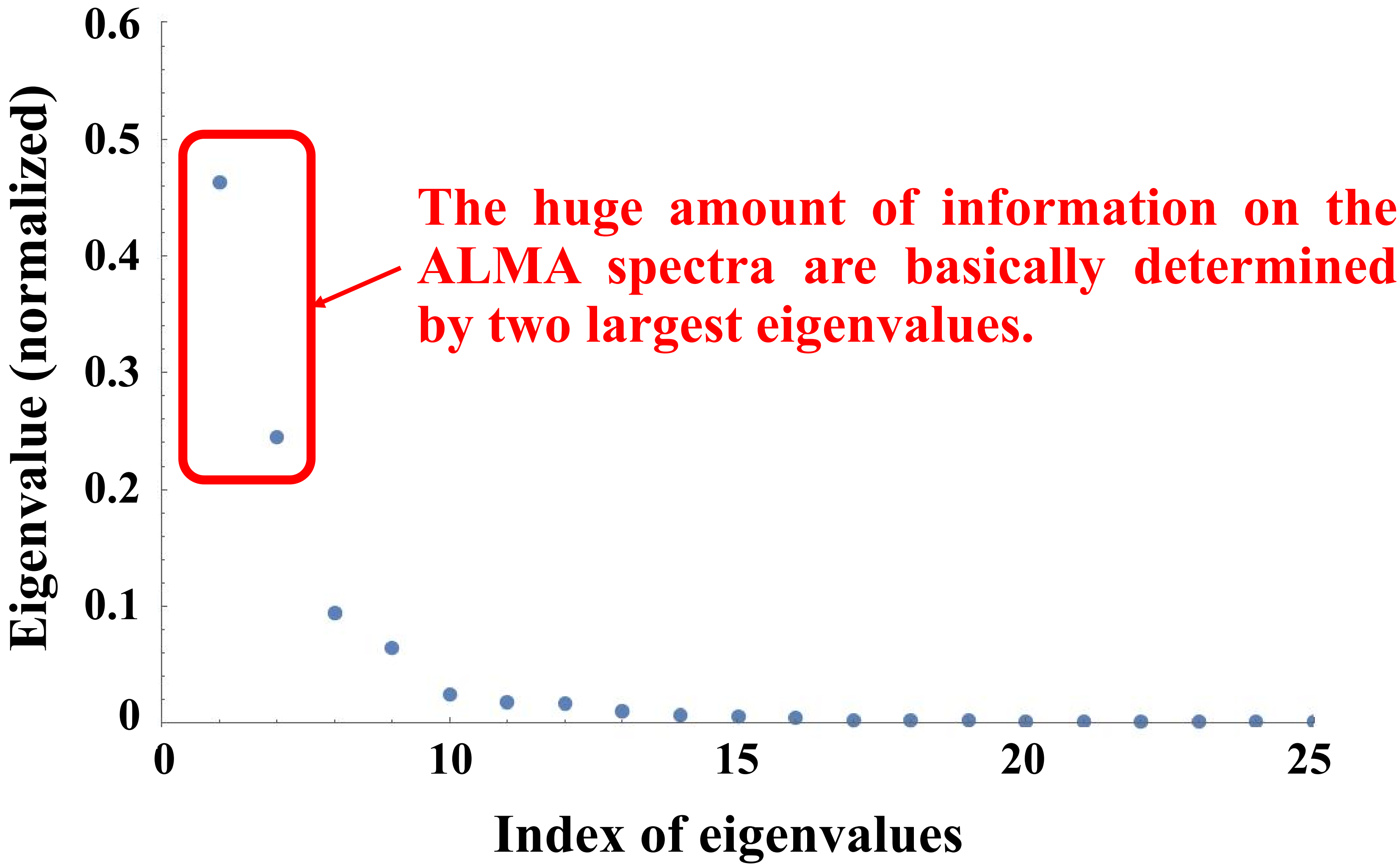
Eigenvalues of the PCA (contribution)



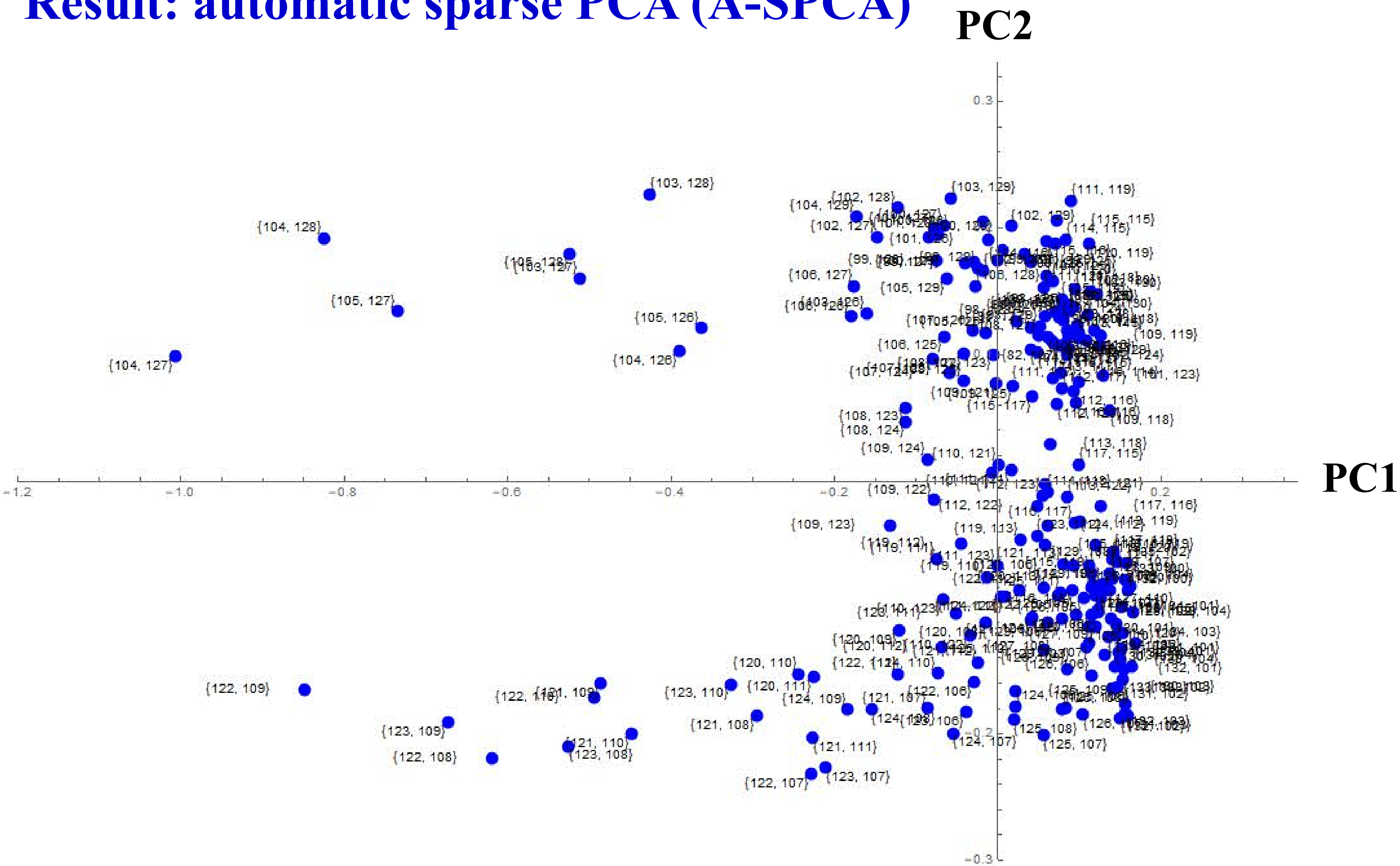
3. Analysis of Starburst Region in NGC253

3.1 Analysis of Raw Data

Eigenvalues of the PCA (contribution)

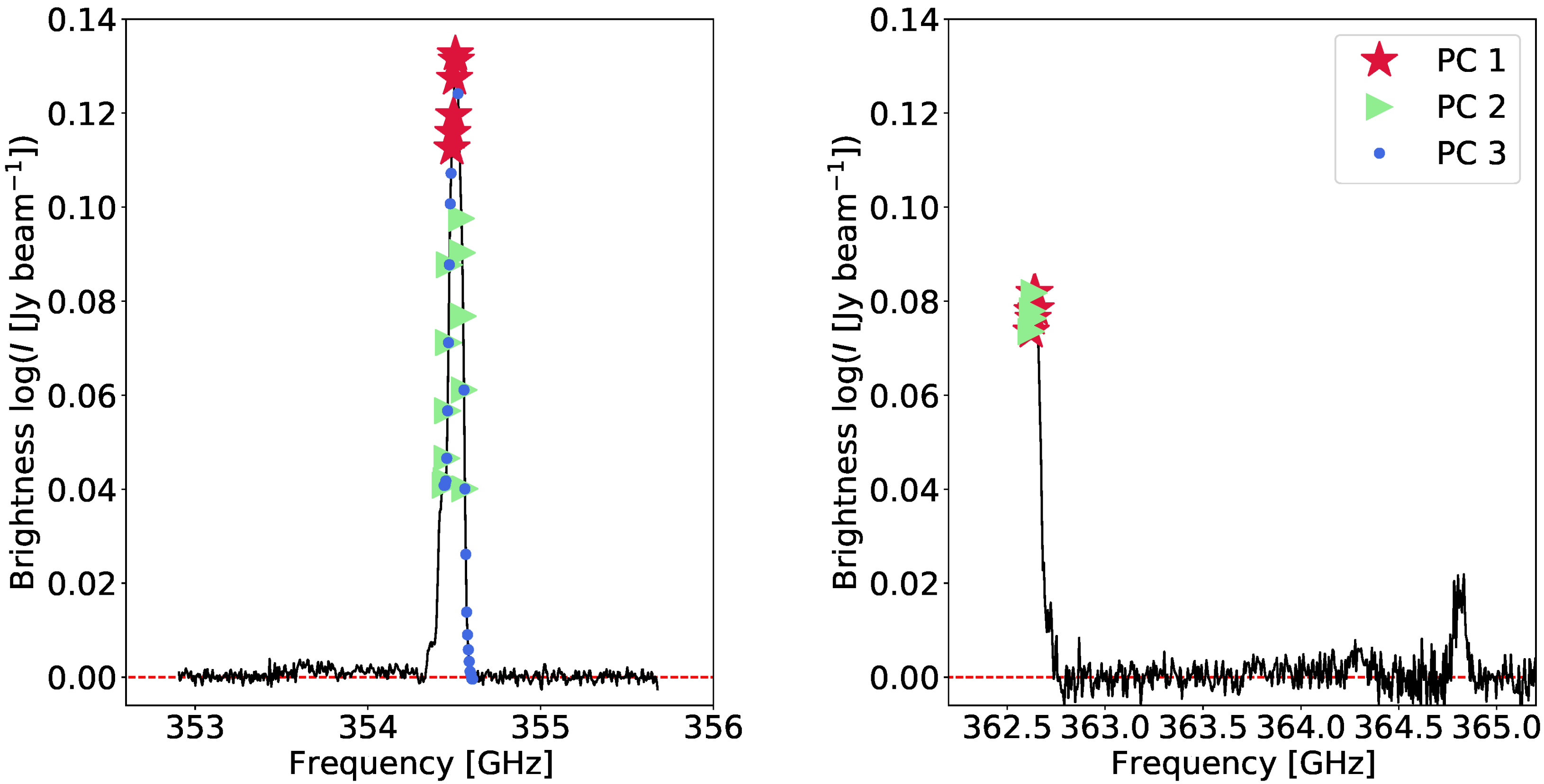


Result: automatic sparse PCA (A-SPCA)



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!

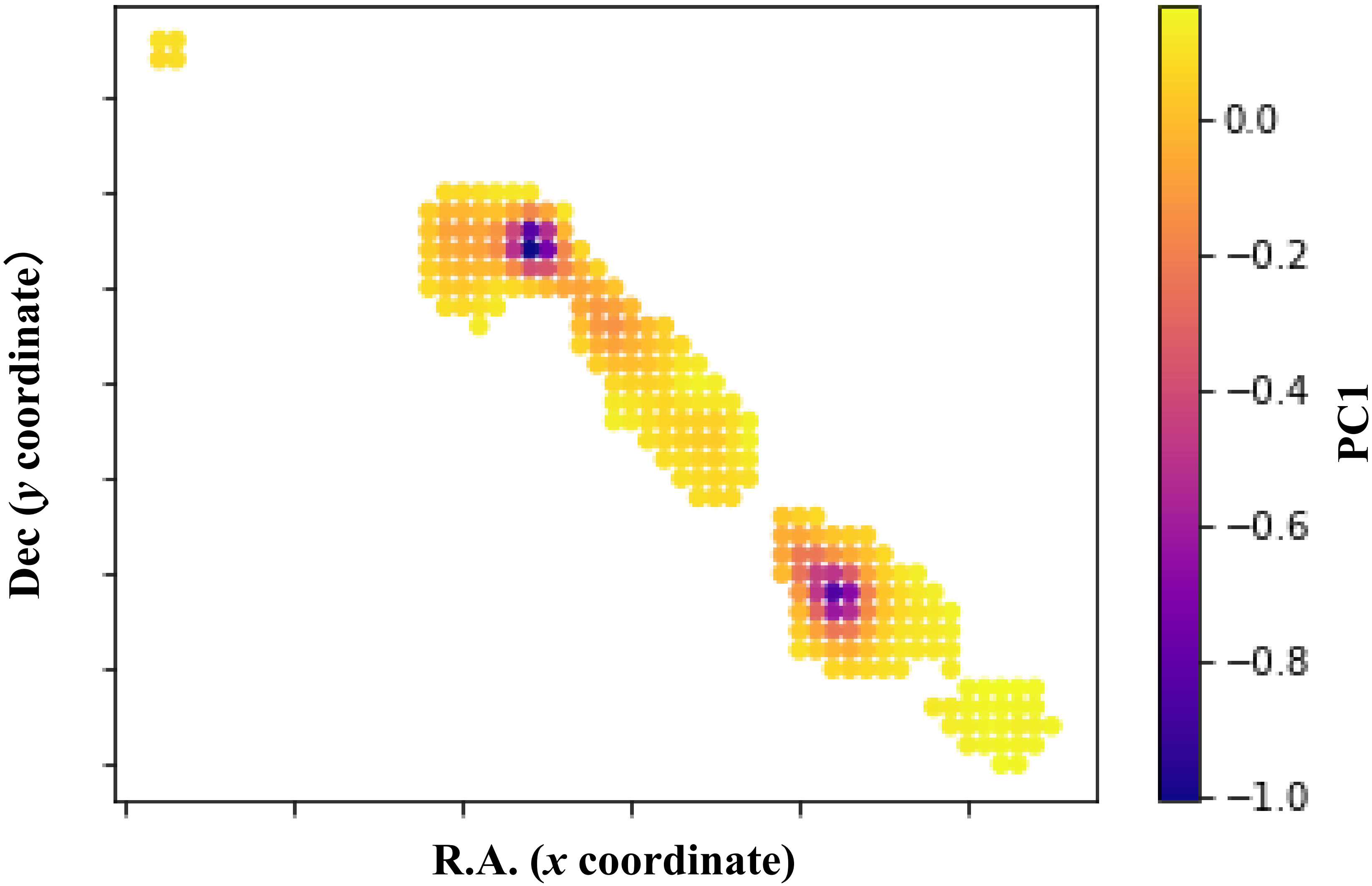
Responsible spectral features for PC1, PC2 and PC3



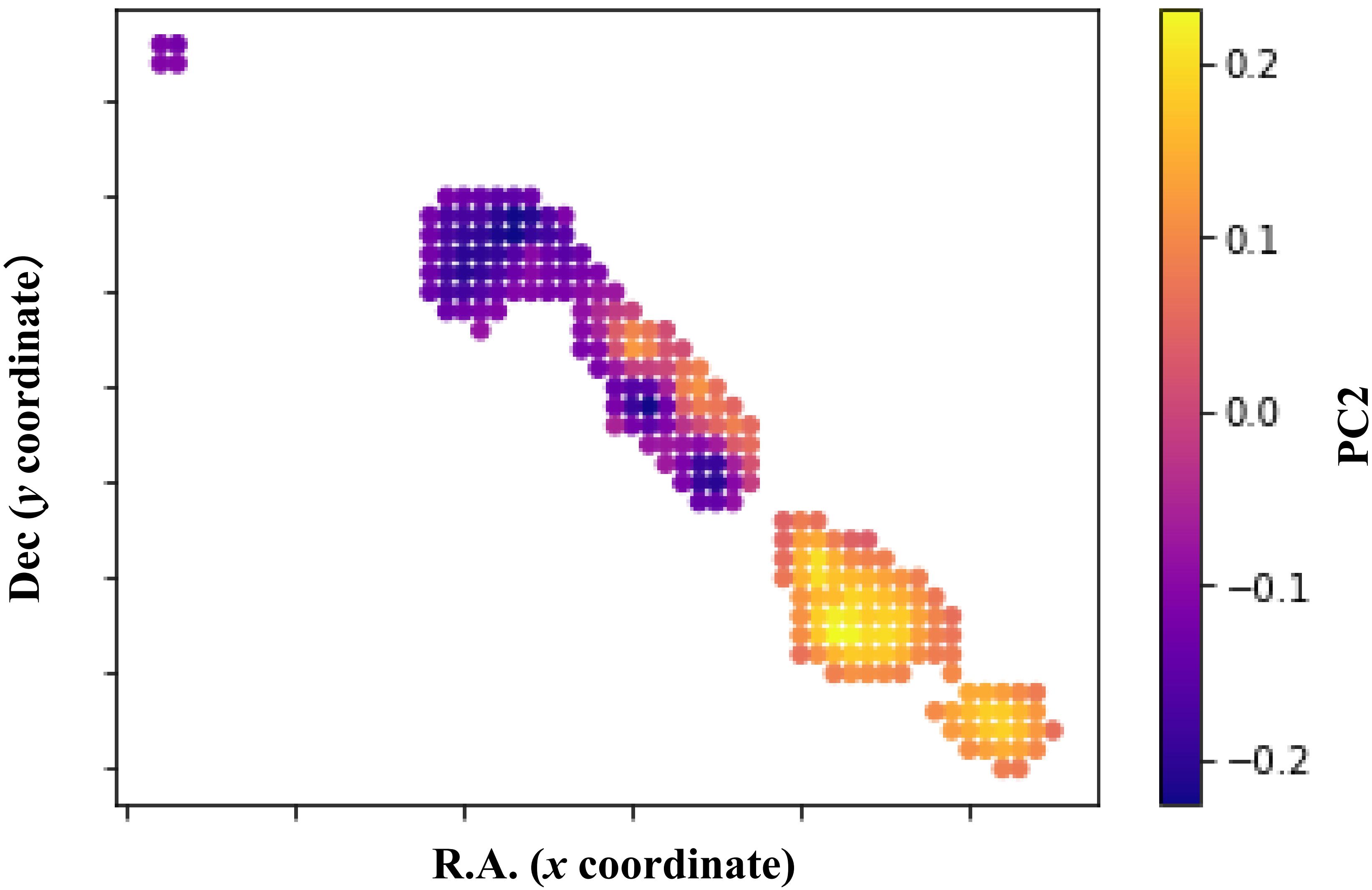
Takeuchi et al. (2022)

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

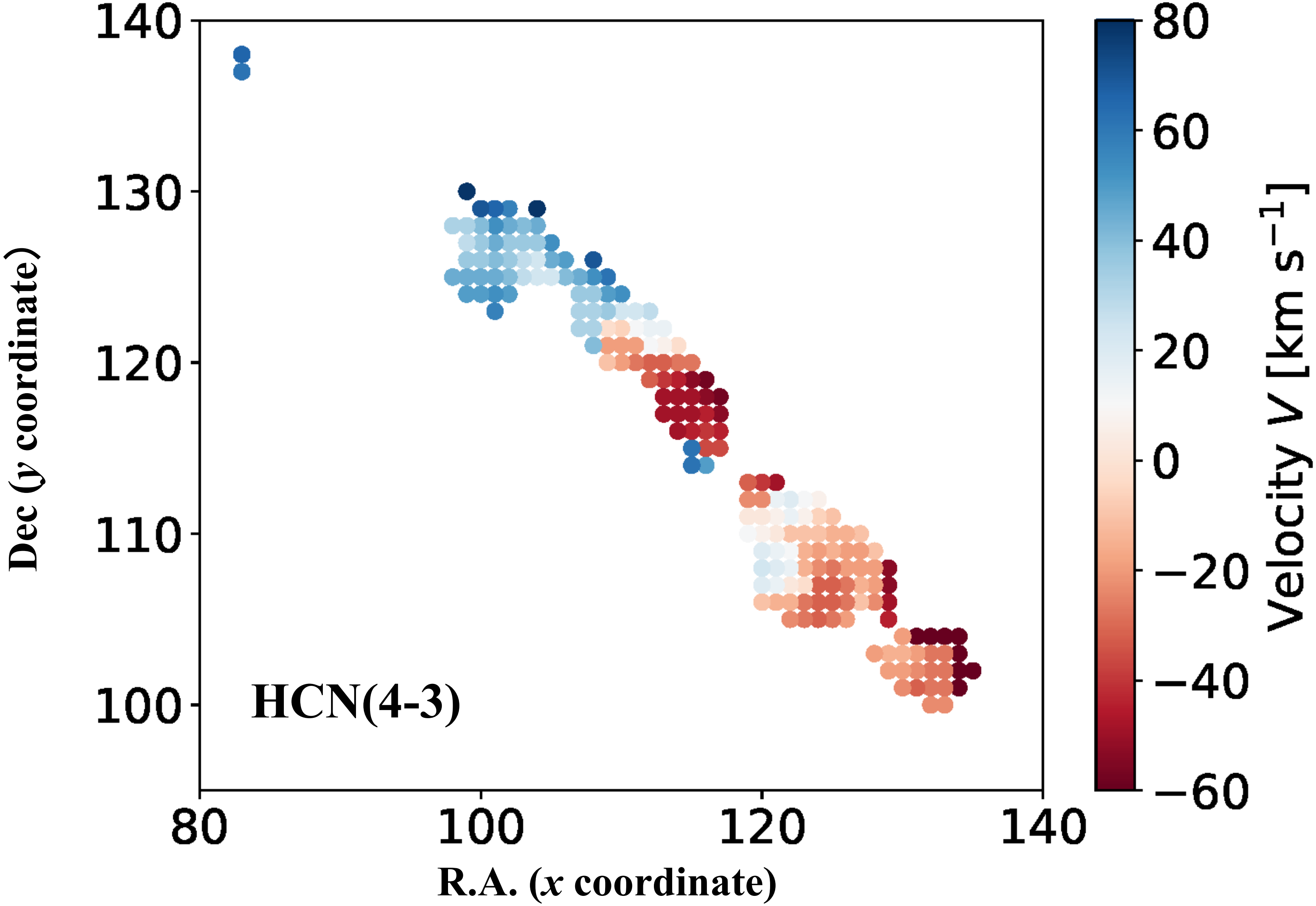
Spatial map of PC1



Spatial map of PC1 and PC2



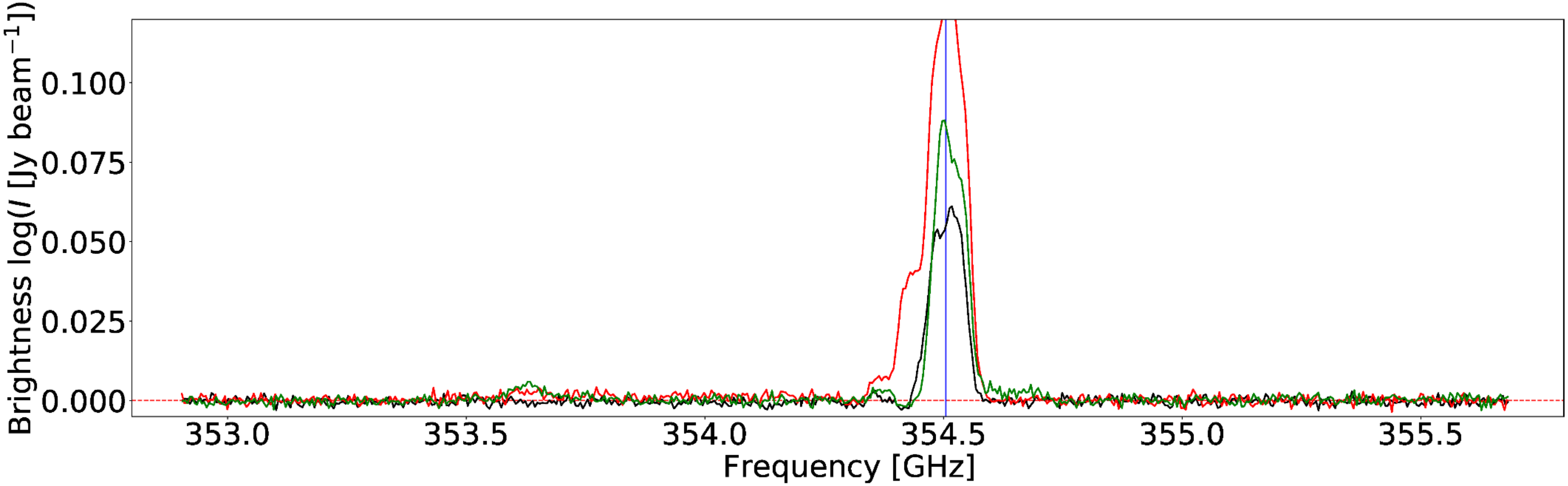
Velocity field of the systemic rotation



⇒ **Doppler shift correction to remove the systemic rotation.**

3.3 Main analysis

Doppler shift correction

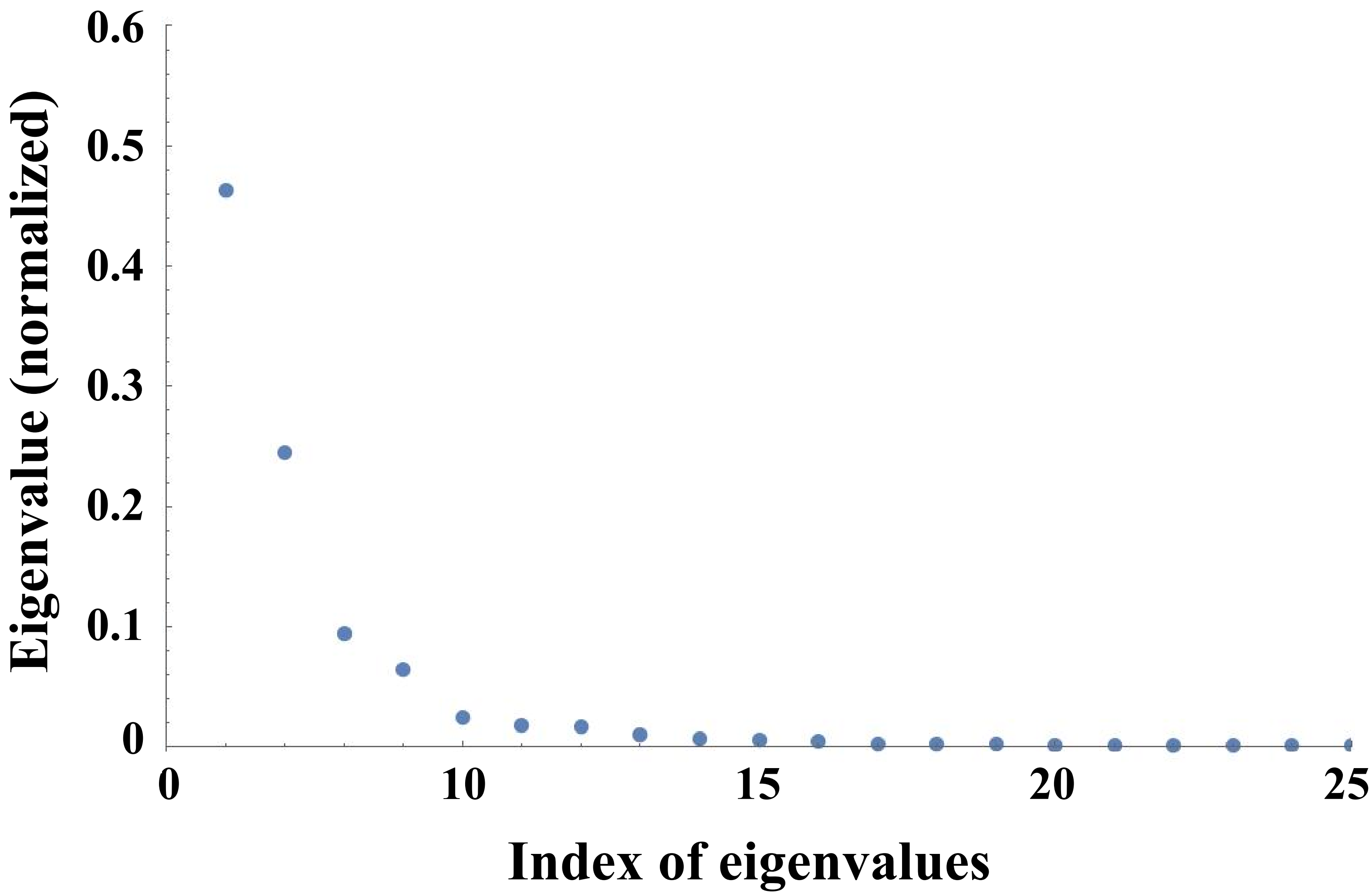


Takeuchi et al. (2021)

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.

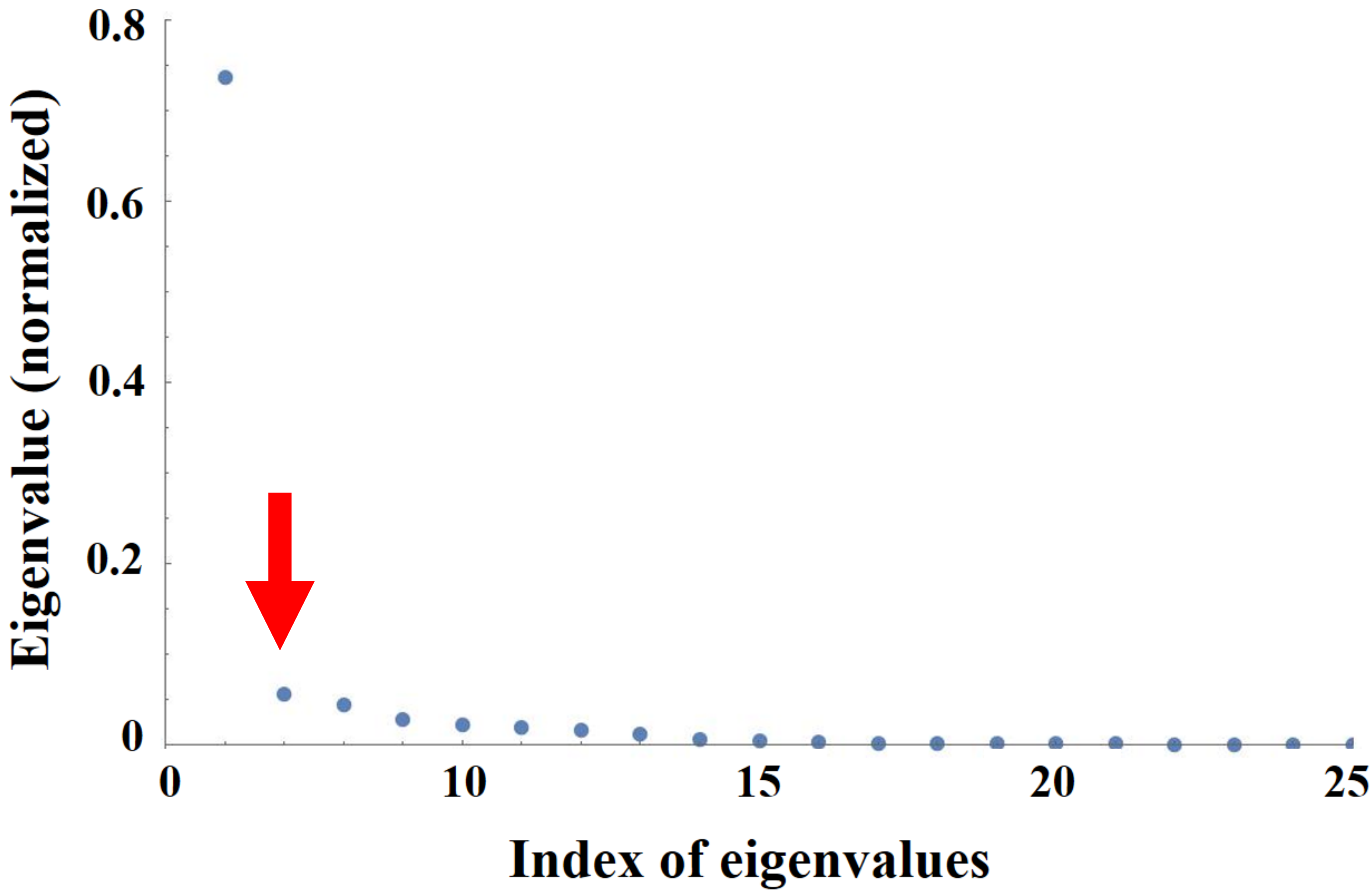
Due to this correction, the final data dimension is *d* = 1971.

Eigenvalues of the NGC253 before Doppler correction



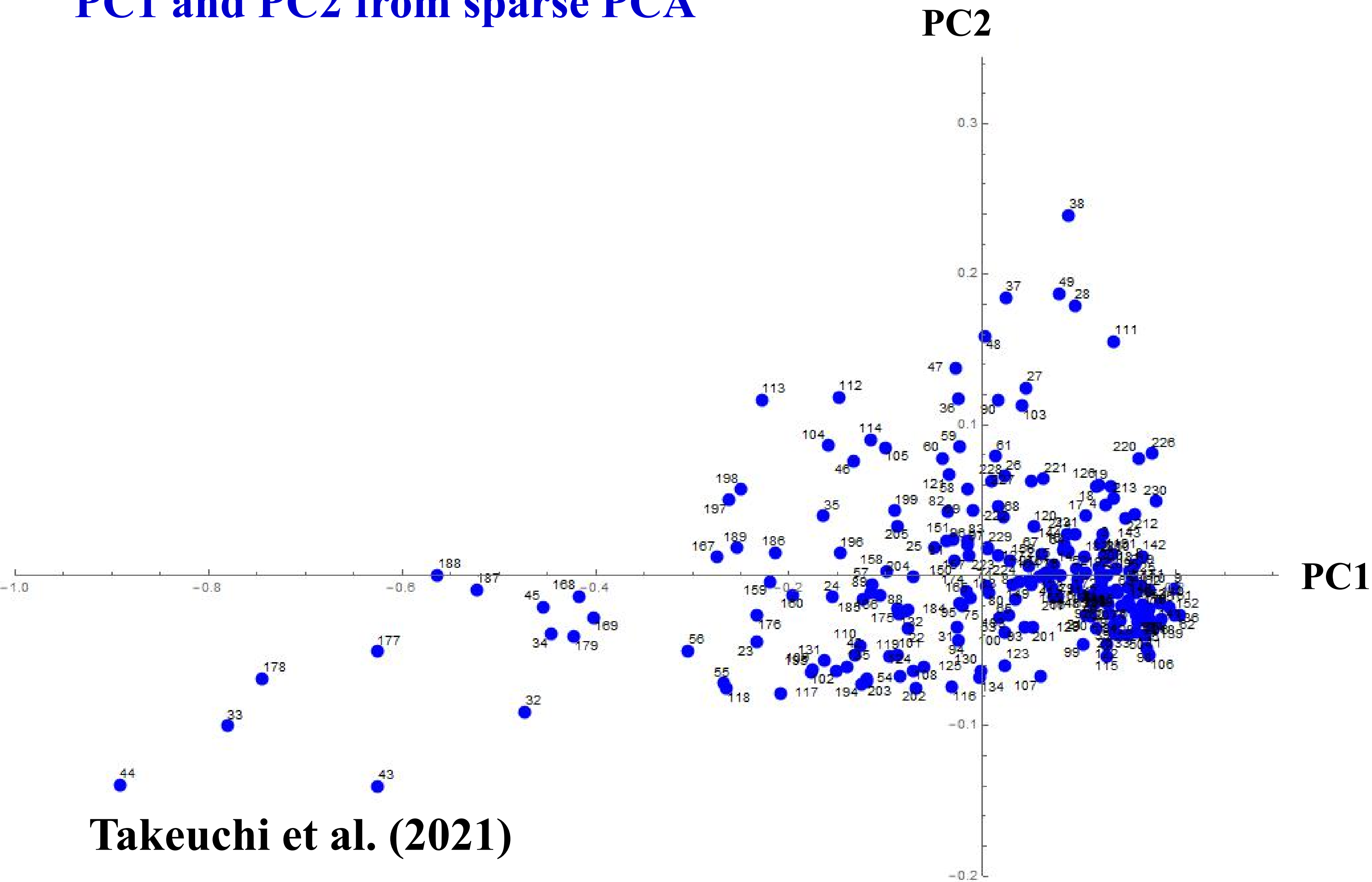
Takeuchi et al. (2023)

Eigenvalues of the NGC253 after Doppler correction



Takeuchi et al. (2023)

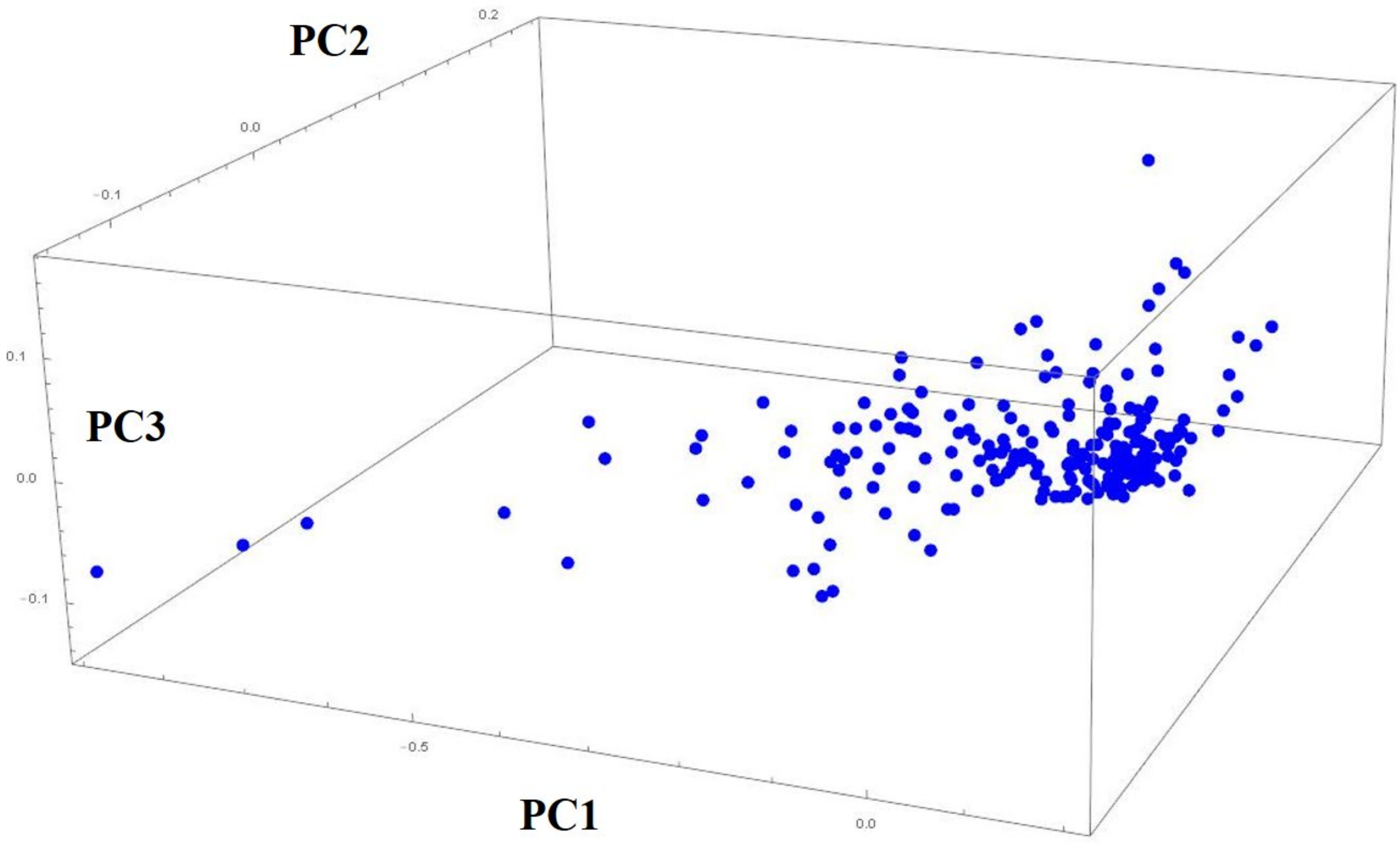
PC1 and PC2 from sparse PCA



Takeuchi et al. (2021)

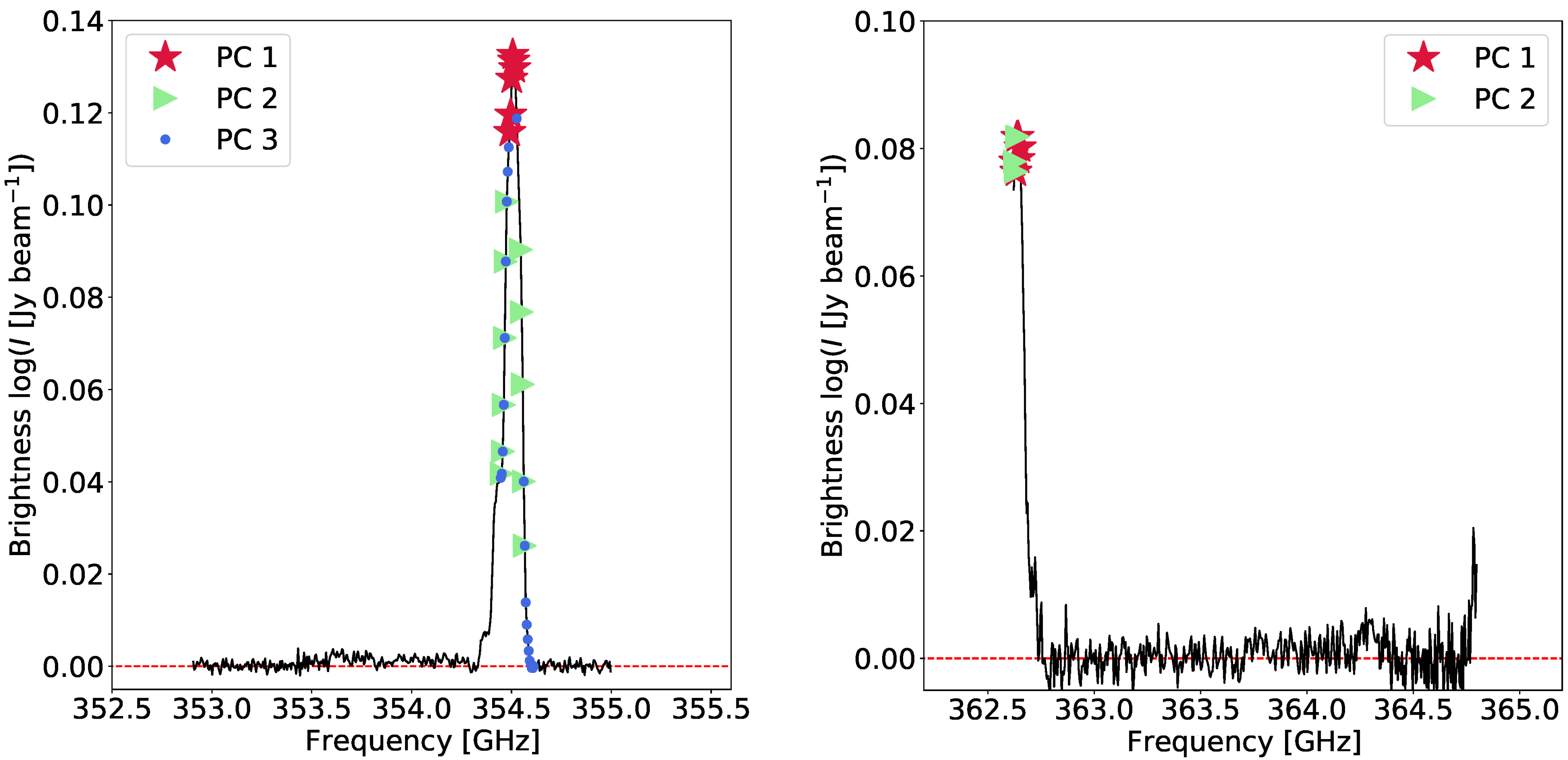
Butterfly-like pattern completely disappeared.

PC1, PC2, and PC3 from sparse PCA



Takeuchi et al. (2021)

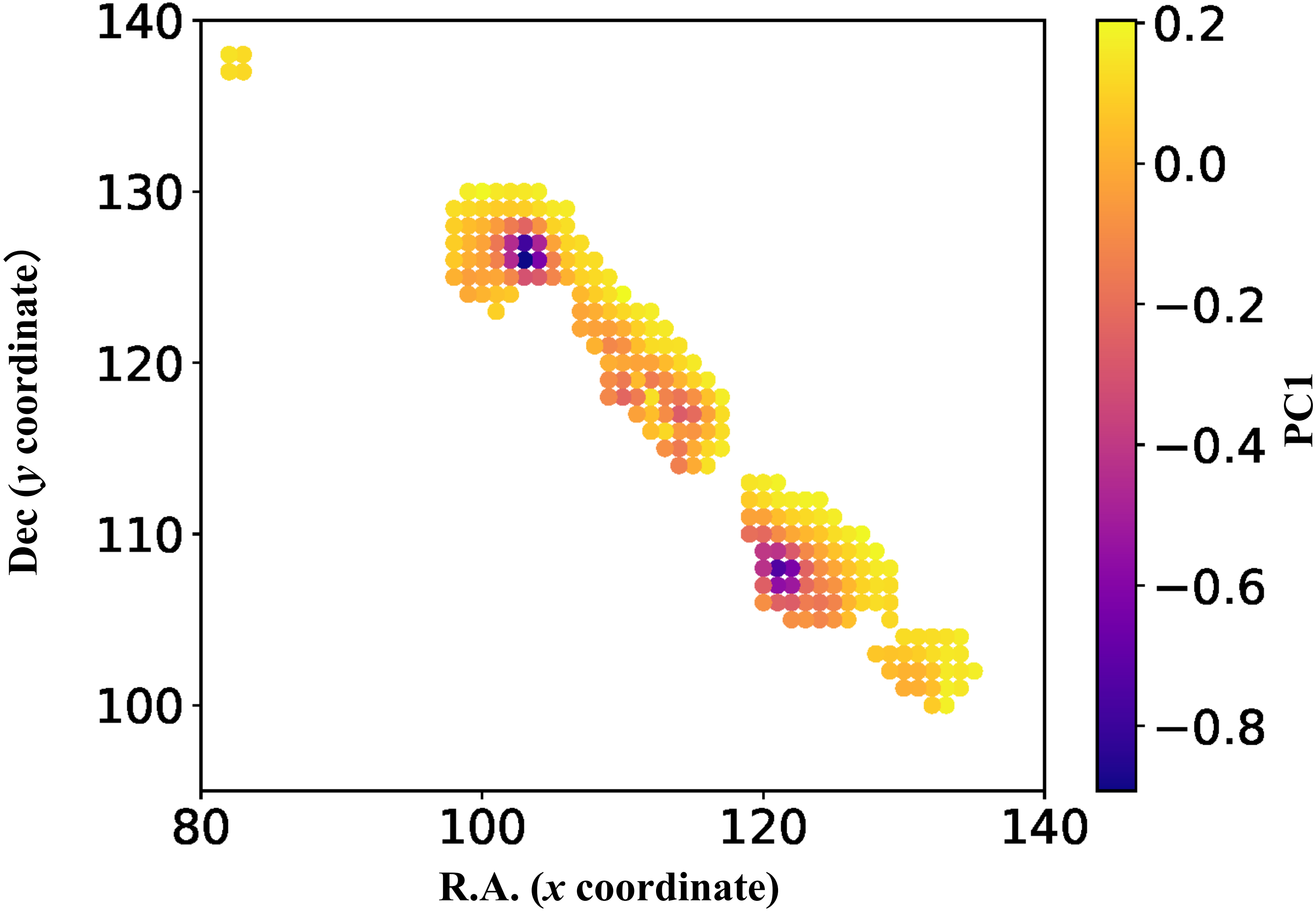
Responsible spectral features for PC1, PC2 and PC3



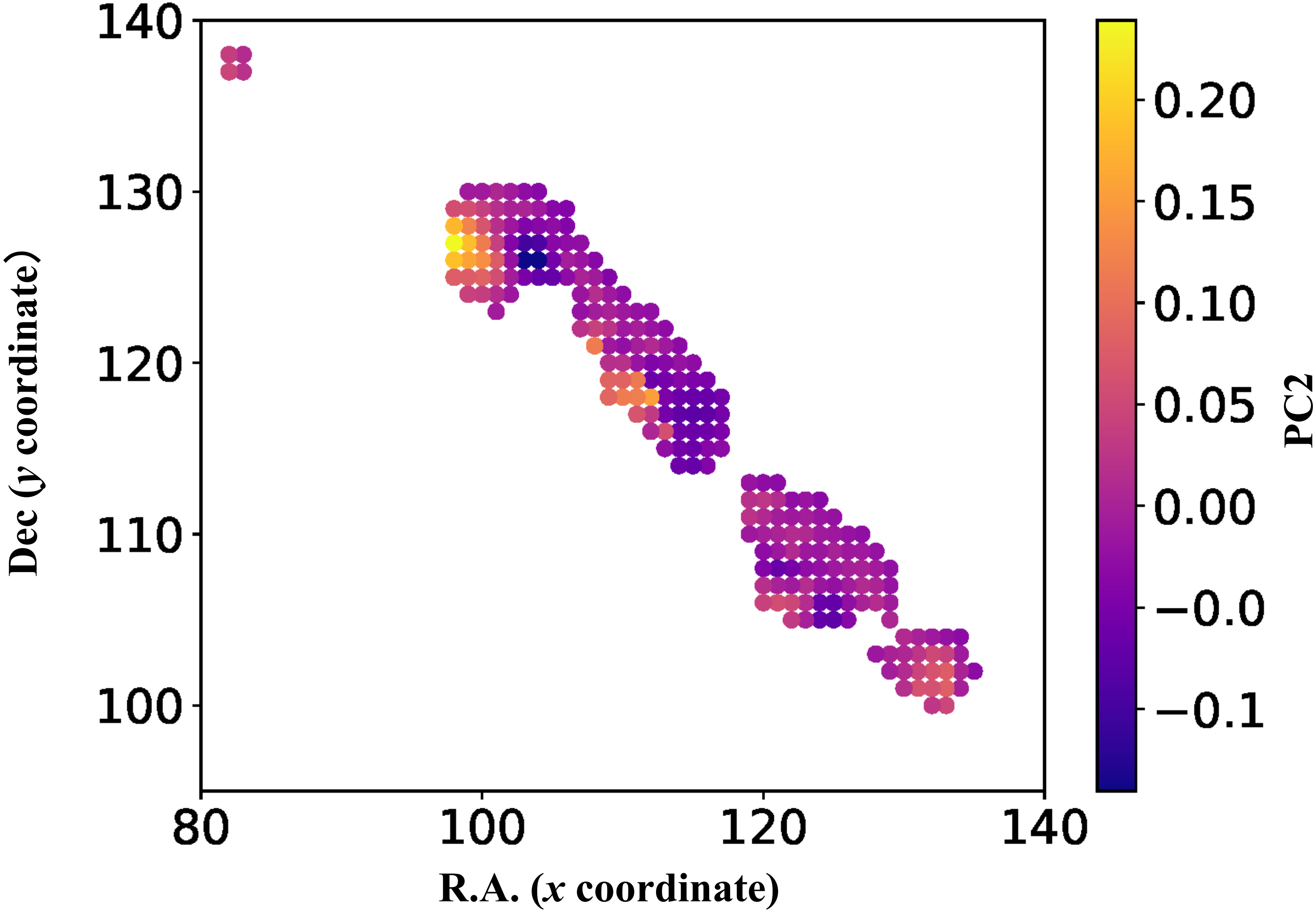
Takeuchi et al. (2021)

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

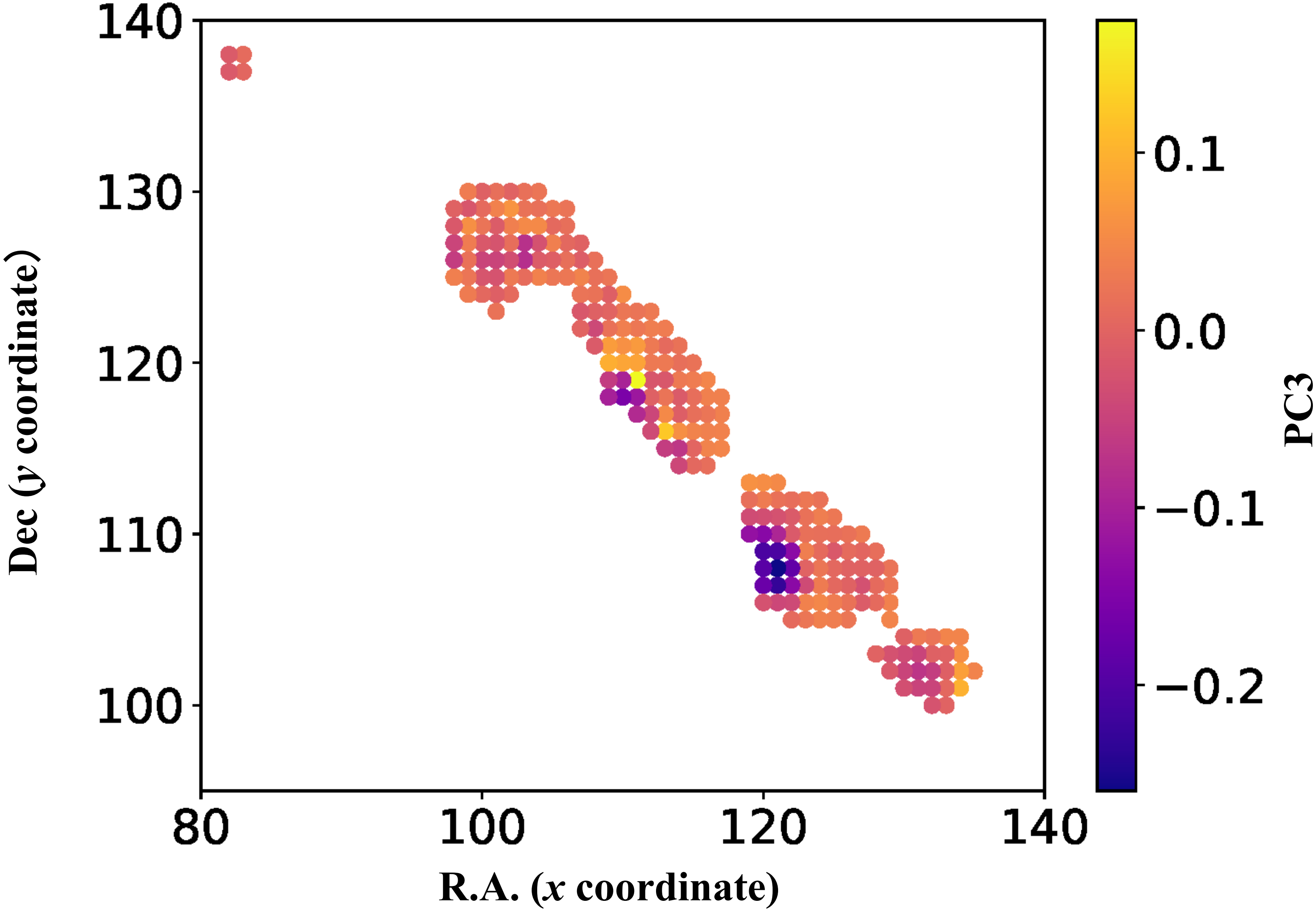
Spatial map of PC1 after Doppler correction



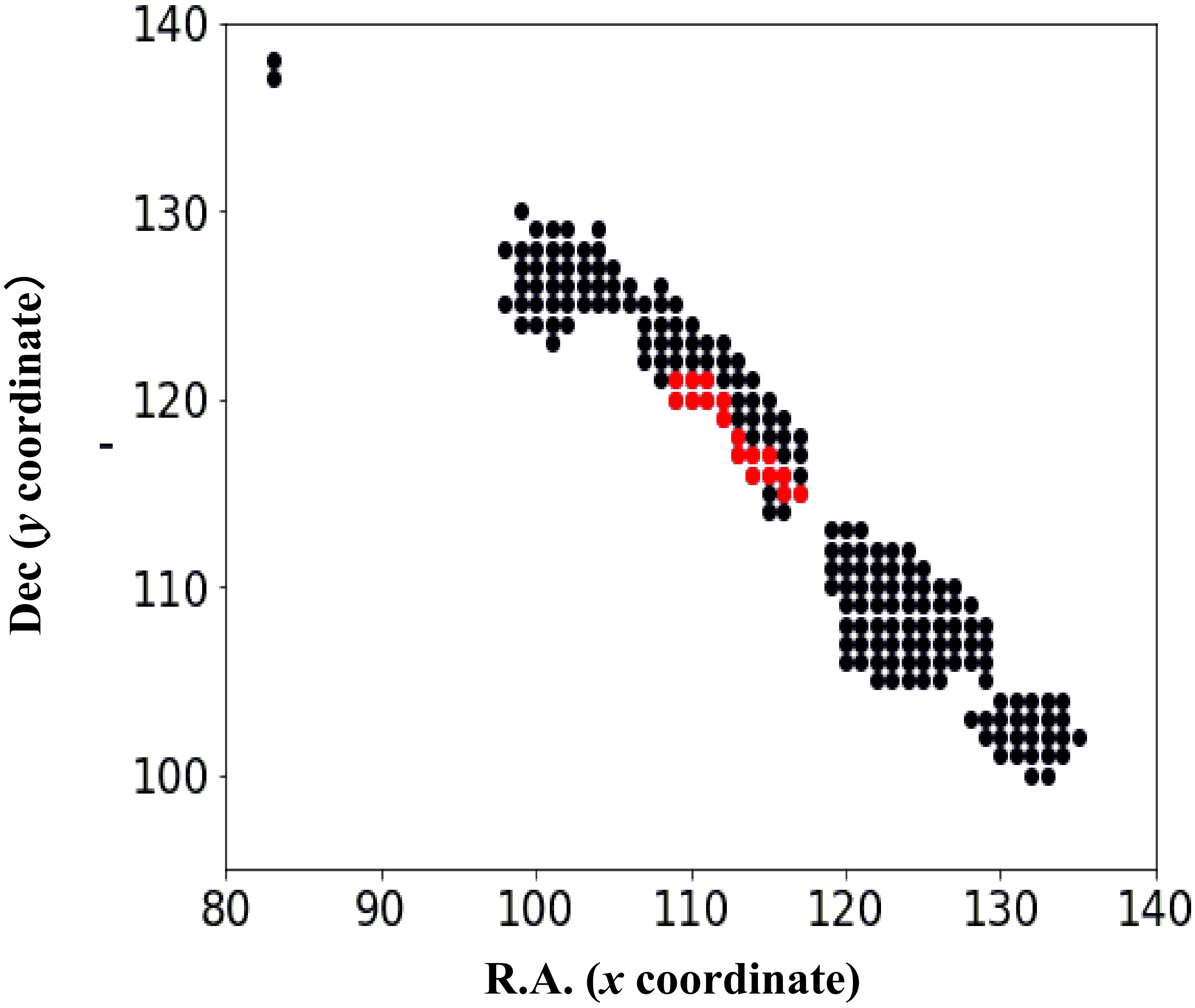
Spatial map of PC2 after Doppler correction



Spatial map of PC3 after Doppler correction



Anomaly regions in the velocity field



What do we see from the Doppler-corrected map?

NGC253

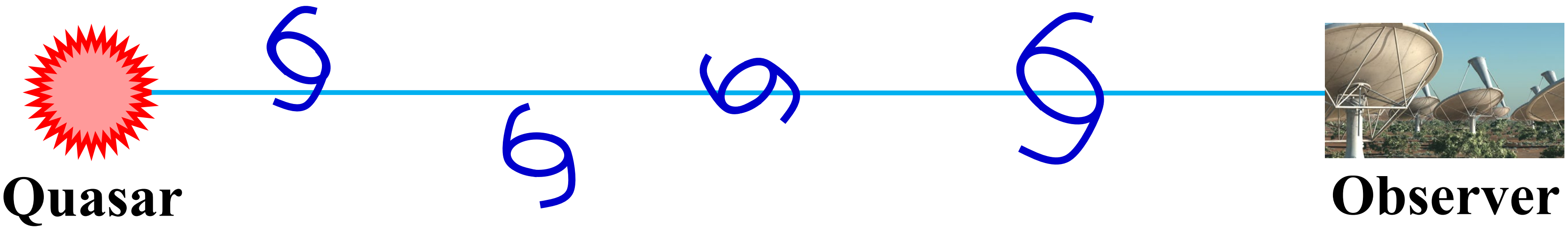
- **Pure starburst: SFR in the central molecular zone is 2 $M_{\odot}$ yr^{-1} (Rieke et al. 1980; Keto et al. 1999)**

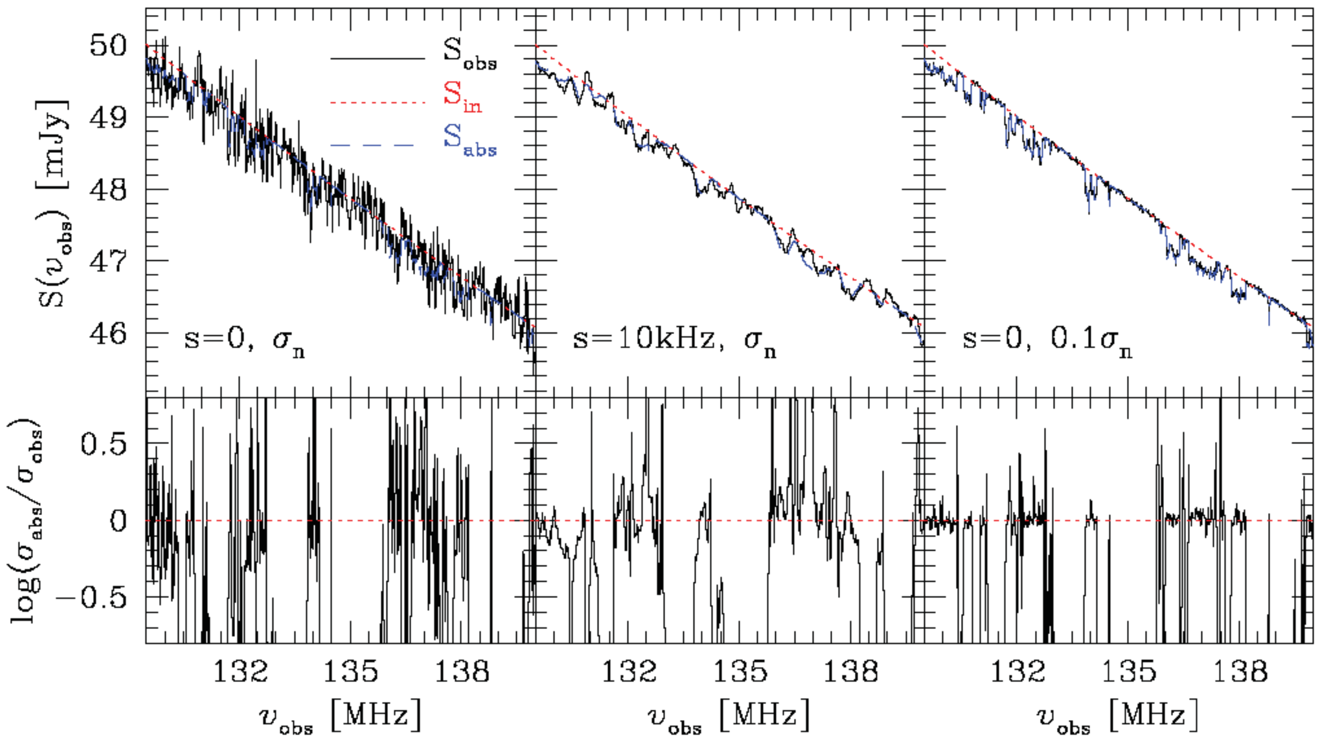
- **Intense outflow (Matsubayashi et al. 2009; Bolatto et al. 2013)**

Indeed the outflow phenomenon is mainly delineated by PC3.

4. Analysis of the HI Forest

4.1 What is the HI forest?

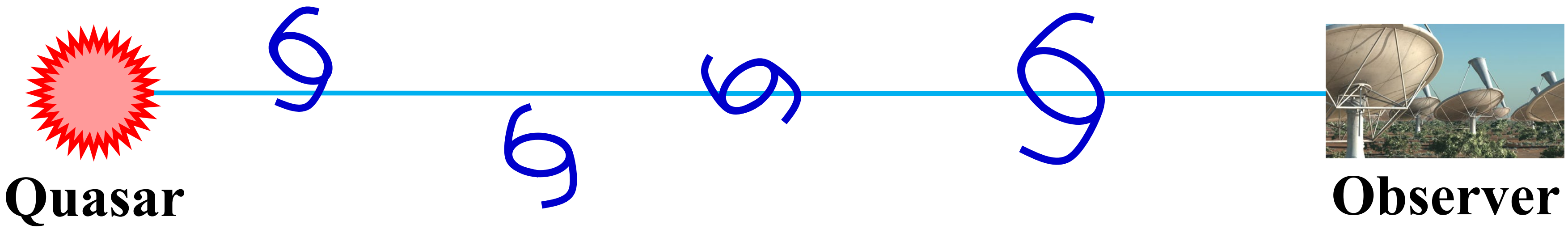




Ciardi et al. (2013)

4.2 Prospect and difficulty in the analysis of HI forest

Prospect

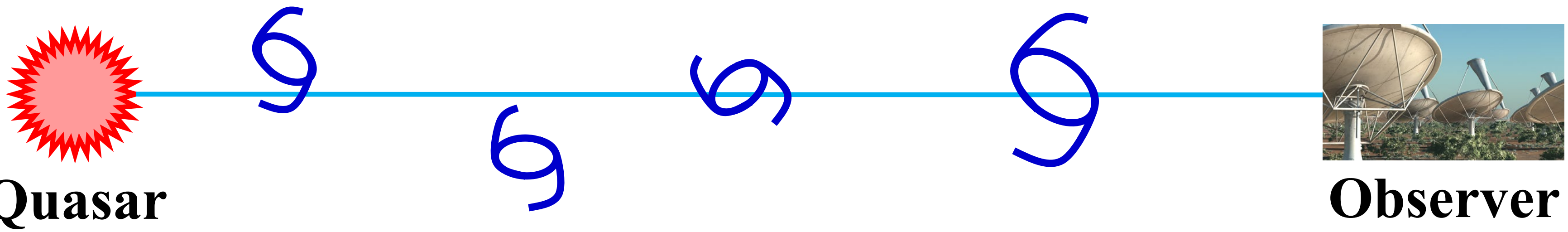


- The HI forest carry the information on the spatial distribution of primordial galaxies. This provides a very important clue to the formation of first galaxies.

- 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.

4.2 Prospect and difficulty in the analysis of HI forest

Difficulty



The background quasars are very rare.

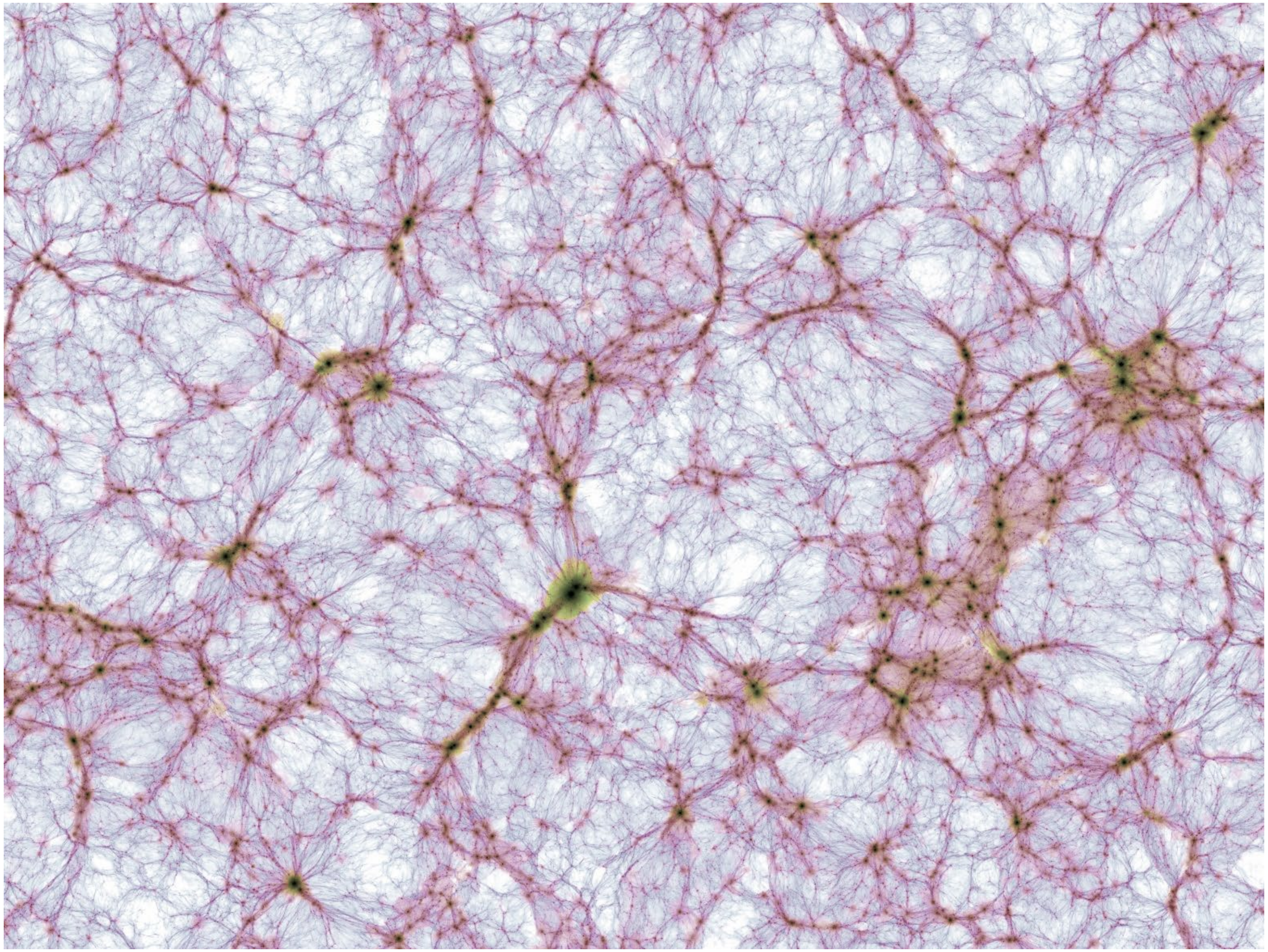
Even by the next-generation radio observational facility Square Kilometre Array (SKA), only a few tens of quasars are expected., while the absorption lines are numerous.

⇒ HDLSS data!

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

4.3 Basic analysis with cosmological simulation

Illustris TNG simulation



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

4.3 Basic analysis with cosmological simulation

Quantification of the spatial distribution of HI gas

Absorption line frequency is described as

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

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

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

4.3 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

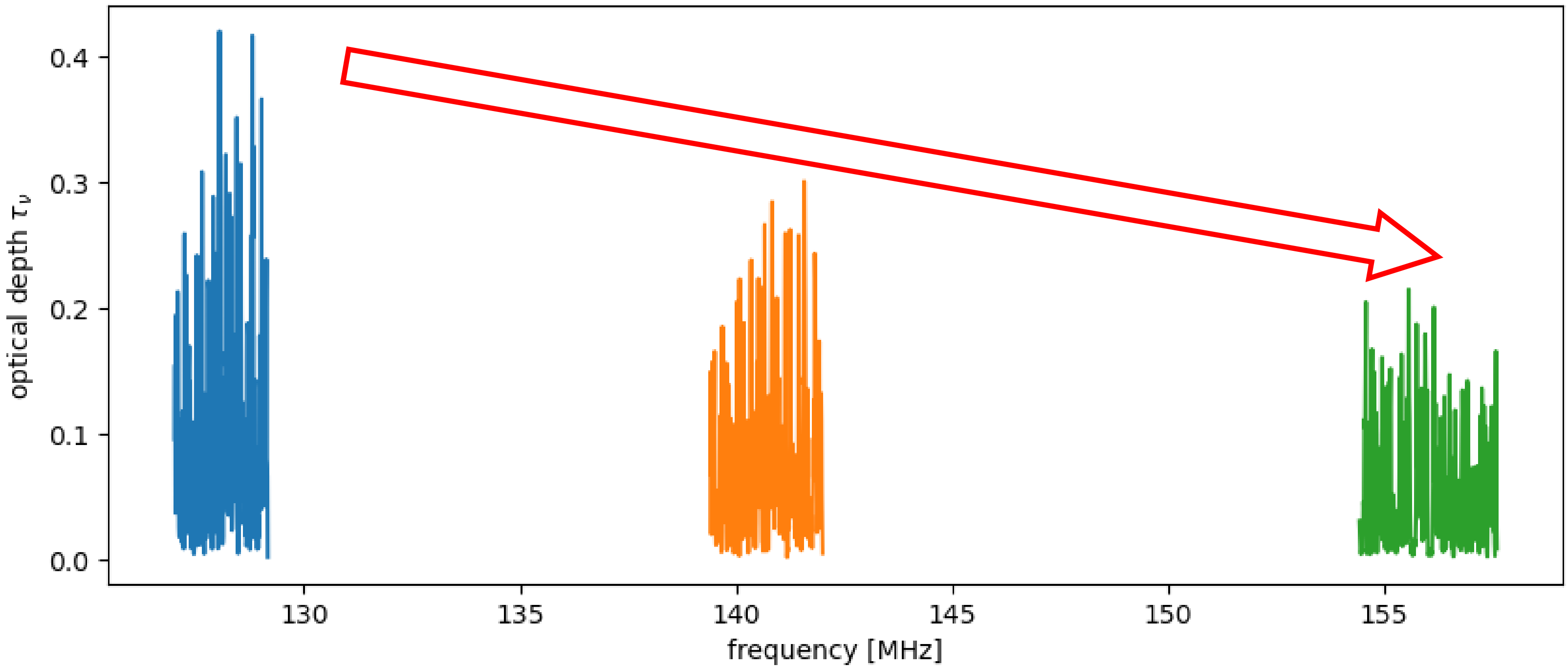
$$\Xi = \begin{pmatrix} \xi^{(1)}(0) & \xi^{(2)}(|z_1 - z_2|) & \dots & \xi^{(n)}(|z_1 - z_d|) \\ \xi^{(1)}(|z_2 - z_1|) & \xi^{(2)}(0) & \dots & \xi^{(n)}(|z_2 - z_d|) \\ \vdots & & \ddots & \\ \xi^{(1)}(|z_d - z_1|) & \xi^{(2)}(|z_d - z_2|) & \dots & \xi^{(n)}(0) \end{pmatrix}$$

This contains fundamental information on the cosmic matter density field.

4.3 Basic analysis with cosmological simulation

Quantification of the spatial distribution of HI gas

High-dimensional PCA of the HI forest absorption lines



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

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.**

4. 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. The spatial distribution and evolution of the hydrogen absorption line systems (HI forest) can be efficiently explored by the high-dimensional statistical analysis.

The high-dimensional statistical analysis can be applied to a vast range of astronomical problems with small sample size. Stay tuned!

If you are interested in details, see astro-ph/2203.04535.