High Dimensional Statistical Analysis and its Application to ALMA Map of NGC 253

TSUTOMU T. TAKEUCHI 1,2, KAZUYOSHI YATA 3, KENTO EGASHIRA 4,5, MAKOTO AOSHIMA 3, AKI ISHII 4, SUCHETHA COORAY 6,7, KOUCHIRO NAKANISHI 6,8, KOTARO KOHNO 9, and KAI T. KONO 1

1 Division of Particle and Astrophysical Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464–8602, Japan
2 The Research Center for Statistical Machine Learning, the Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190—8562, Japan
3 Institute of Mathematics, University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki 305–8571, Japan
4 Department of Information Sciences, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba 278—8510, Japan
5 Graduate School of Science and Technology, University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki 305–8571, Japan
6 National Astronomical Observatory of Japan, National Institutes of Natural Sciences (NINS), 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan
7 Research Fellow of the JSPS (PD)
8 Department of Astronomy, School of Science, Graduate University for Advanced Studies (SOKENDAI), 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan
9 Institute of Astronomy, Graduate School of Science, The University of Tokyo 2-21-1, Osawa, Mitaka, Tokyo, 181-0015, Japan

(Received; Revised; Accepted)

Submitted to ApJS

ABSTRACT

In astronomy, 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, such a situation is regarded as ill-posed, and 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 to the astronomical community. 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 R-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

Corresponding author: Tsutomu T. Takeuchi
tsutomu.takeuchi.ttt@gmail.com
NGC 253. This method can also be applied not only to spectroscopic survey data, but also any type of data with small sample size and large dimension.

**Keywords:** methods: statistical — ISM: lines and bands — galaxies: ISM — galaxies: star formation — galaxies: individual (NGC 253) — radio lines: galaxies

1. **INTRODUCTION**

In the past decade, the term “Big Data” has been widely recognized and observed almost all the field of scientific researches, of course also in astronomy. Current astronomical instruments provide us with overwhelmingly large data, providing quantitative information of their physical condition. The most typical big data in astronomy is the product from large surveys, such as the Sloan Digital Sky Survey (SDSS)\(^1\), the Dark Energy Spectroscopic Instrument (DESI) Survey\(^2\), Gaia\(^3\), the Panoramic Survey Telescope And Rapid Response System (Pan-STARRS) Surveys\(^4\), among many others.

We should note, however, that there is another type of big data in astronomy. Detailed observations are often very time-consuming, and it is not easy to map objects to obtain many independently sampled measurements. Such a situation is frequently found in the integral field spectroscopy (e.g., UVES at the Very Large Telescope (VLT)\(^5\), SAURON on the William Herschel Telescope\(^6\), SDSS MaNGA data\(^7\), among others. Another typical example of this kind is spectroscopic mapping surveys by radio interferometers (e.g., Atacama Large Millimeter/Submillimeter Array) or by satellite instruments at X-ray, ultraviolet or mid/far-infrared wavelengths. In such a case, if we denote the dimension in the wavelength (or frequency) with \(d\) and the number of samples with \(n\), we often find that \(n \ll d\).

Classical statistical analysis implicitly assumed a sample size \(n\) is much larger than data dimension \(d\) (\(n \gg d\)). As we mentioned, this is not true for many cases in scientific researches. Traditionally in astrophysics, such a situation is regarded as an ill-posed problem, and there was no choice but to throw away most of the information in wavelength direction to let \(d < n\). It was more or less the same for other research fields before 2010. Many researchers have simply given up further analysis, or even did not find a need to deal with such problems before the advent of the Big Data era. For example, Hand et al. (1994) published a well-known compilation of statistical data, containing more than 500, but most of them satisfied the classical condition of the assumption \(d < n\). However, in late 1990’s, data with \(d \gg n\) have suddenly started to appear along with the development of the information science. One of the epoch-making example of such kind of researches was Golub et al. (1999). They have used the gene expression monitoring by DNA microarrays applied to human acute leukemias as a test case. The dimension of the DNA microarray was \(d = 7129\) and the sample size \(n = 72\). Statistics in 1990’s could not guarantee the accuracy for the estimation in such a case, since it was on the basis of \(d < n\).

Various fundamental problems were known for data with the condition of \(n \ll d\):

1. Sample covariance matrix is usually numerically unstable or does not exist, and classical multivariate analysis cannot work.

2. We meet various type of “curse of dimensionality”.

---

\(^1\)https://www.sdss.org/surveys/.

\(^2\)https://www.desi.lbl.gov/.

\(^3\)https://www.cosmos.esa.int/web/gaia.

\(^4\)https://outerspace.stsci.edu/display/PANSTARRS/.

\(^5\)https://www.eso.org/sci/facilities/paranal/instruments/uves.html.

\(^6\)https://www.ing.iac.es/PR/SH/SH2007/sauron.html.

\(^7\)https://www.sdss.org/surveys/manga/.
3. Calculation cost is often unreachably tremendous.

A method to handle such a problem have been desired since then. In order to study such type of data with a very high dimension and small sample data, first of all we had to push back the boundaries of the traditional technique of multivariate statistical analysis. This is because traditional methods are not able to capture the characteristics of the high-dimensional data space and hence, not able to make use of rich information which the data intrinsically contain.

The situation has drastically changed in 2000’s. A turning point was brought from the field of probability theory and theoretical physics based on the random matrix theory. In these studies, the asymptotic behavior of eigenvalues of the sample covariance matrix in the limit as $d \to \infty$ was explored under Gaussian or Gaussian-type assumptions on the population distribution when $d$ and $n$ increase at the same rate, i.e., $n/d \to c > 0$ (c: const.) (e.g., Baik et al. 2005; Baik & Silverstein 2006; Johnstone 2001; Paul 2007).

In our current interest, however, the assumption as $n/d \to c > 0$ and Gaussian(-type) are actually unrealistically restrictive. A seminal paper by Hall et al. (2005) introduced a new concept, high dimension low sample size (HDLSS) asymptotics\(^8\). The limits they have studied were taken as the dimension of the data $d \to \infty$, while the sample size $n$ was fixed. At that time, this was a revolutionary departure from the traditional asymptotics based on $n \to \infty$ with $d$ fixed.

Hall et al. (2005) and Ahn et al. (2007) discussed an important representation of the HDLSS data, but still under the assumption of Gaussian(-type). In contrast, Yata & Aoshima (2012) developed a new theory by clearing away this assumption, and discovered a completely different type of representation when the Gaussian assumption is not satisfied (see § 2 for the details). These works were the foundation of research in mathematical statistics on very high dimensional data analysis, and evolved into an important part of the currently fashionable research area, the Big Data Science. The new insight shed light to various areas of studies as genomics, medical research, neuroscience, and image and shape analysis. Then, in their sequence of papers, M. Aoshima, K. Yata, A. Ishii and collaborators further developed a new framework of statistical methodology to tackle this type of problems (e.g., Aoshima & Yata 2011; Yata & Aoshima 2012, 2013; Aoshima & Yata 2014, 2015; Ishii et al. 2016; Aoshima & Yata 2019, and references therein). Readers who are interested in the mathematical background of the high-dimensional statistical analysis itself are guided to some reviews,(e.g., Aoshima & Yata 2017; Aoshima et al. 2018).

In spite of the great developments in other research fields, this new methodology of statistics seemed to be completely unknown in astronomy. High dimensional data as the ALMA data are regarded as a sum of informative signals and noise, but we cannot separate them quantitatively without prior information. Supposing that we know the frequencies of emission lines is a kind of such prior information which has been traditionally used in traditional astronomy, but this is not what we prefer. While some frequency regions are considered to contain no information, we proceeded extracting information not by traditional procedure, but by an objective statistical method without a prior information. The high-dimensional statistical analysis is one of such an objective methods.

Thus, in this work, we introduce the method of high-dimensional statistical analysis for the HDLSS data in astronomy. In order to examine the performance of the high-dimensional methods to the astronomical data, we apply it to a spectroscopic map of the central region of NGC 253. NGC 253 is a nearby prototypical starburst galaxy (Rieke et al. 1980), and has been observed in various wavelengths. We apply the high-dimensional principal component analysis (PCA) to the ALMA spectral map (Ando et al. 2017, hereafter

---

\(^8\) The first consideration on the HDLSS asymptotics, however, dates back to 1980’s by the work of Casella & Hwang (1982)
A17 have presented an 8 pc × 5 pc resolution view of the central ∼ 200 pc region of NGC 253, based on the ALMA Band 7 observations. As we introduce in details, the ALMA maps can be regarded as being typically HDLSS. The spectrum of NGC 253 is very rich in molecular lines (e.g., Ando et al. 2017; Martín et al. 2021). Another focus of this work is to classify the spectra full of spectral line features by the PCA. Evolution of galaxies is mainly driven by the star formation, a transition from interstellar medium (ISM) to stars. Various phases of the ISM are related, and the evolution of the ISM is a key to complete the understanding of the galaxy evolution. Spectroscopic observations are of vital importance to extract and interpret the information of matter in galaxies (e.g., Martín et al. 2021). Then, this classification is potentially a powerful tool for the studies of galaxy evolution if it turns out to be useful.

This paper is organized as follows. In § 2, we try to introduce the concept of HDLSS data and high-dimensional statistical analysis, with an emphasis of the high-dimensional principal component analysis (PCA). There we present some important theorems that characterize the surprising features of the HDLSS data, without detailed mathematical proofs. Then in § 3, we move on to the actual analysis of astronomical data. As already mentioned, we used an ALMA map of NGC 253. We summarize the data properties in this section. Our data analysis is twofold: first we examine if the high-dimensional statistical analysis is really useful for the analysis of spectroscopic mapping data in § 4. After confirming its excellent performance in this type of analysis, we further go to a deeper analysis of the NGC 253 molecular lines in § 5. § 6 is devoted to the conclusions of this work. Some technical details related to both theoretical and practical methods in this work are presented in Appendix. Appendices A and B describe the properties of the dual sample covariance and its geometric representation. In Appendix C, we present the methodology for non-Gaussian data that was not discussed in the main text. Comparison between the results of the traditional and high-dimensional PCAs is shown in Appendix D. The method to correct the Doppler shift is presented in Appendix E.

2. METHOD: HIGH-DIMENSIONAL STATISTICAL ANALYSIS

2.1. High-Dimensional Statistical Analysis

The HDLSS data have a sample size $n$ and data dimension $d$ with $n \ll d$. In this work, we mean the newly developed statistical methodology to deal with the HDLSS data by “high-dimensional statistical analysis”.

2.1.1. Unusual behavior of the high-dimensional statistics

As a pedagogical example, we demonstrate a typical “unusual” properties only appear in the high-dimensional statistical analysis, known as the strong inconsistency. We consider a sample mean $\bar{x}$,

$$\bar{x}_n \equiv \frac{1}{n} \sum_{i=1}^{n} x_i. \quad (1)$$

Suppose that $x_i$s are drawn from a $d$-dimensional distribution with a mean $\mu$. Under the usual condition often supposed in the traditional statistics, $d/n \longrightarrow 0$, we have

$$\|\bar{x} - \mu\| \overset{P}{\longrightarrow} 0 \quad (2)$$

where the superscript $P$ stands for the convergence in probability\(^9\). Equation (2) is referred to as the consistency of the sample average. However, in the high-dimensional statistics, $d/n \longrightarrow \infty$, and shockingly we

\(^9\) More precisely, $\lim_{d/n \to 0} \mathbb{P}(\|\bar{x}_n - \mu\| > \epsilon) = 0$ for any $\epsilon > 0$. Here $\| \cdot \|$ stands for the Euclid norm, and $\mathbb{P}(A)$ is the probability of an event $A$. Similarly, $\lim_{d/n \to \infty} \mathbb{P}(\|\bar{x}_n - \mu\| < \epsilon) = 0$ for any $\epsilon > 0$ for eq. (3).
have
\[ \| \vec{x} - \vec{\mu} \| \xrightarrow{P} \infty . \] (3)

This is the so-called strong inconsistency in the HDLSS framework. What eq. (3) mean is that the noise in the observed data drastically increases with the increase of the dimension \( d \). Traditional statistical techniques suppose that the noise is negligible compared to the data information. Hence, the traditional statistics can never handle the HDLSS data.

In order to have more concrete image on such “counter-intuitive” phenomena, we examine the behavior of this huge noise in high dimension. For this, however, some mathematical background is necessary. In the following we introduce the background mathematical concepts. We do not try to introduce them in a strict manner, either not try to give rigorous mathematical proofs, even if the following sections seem to be rather mathematically inclined.

2.1.2. Dual sample covariance

Consider a parent distribution of dimension \( d \), with a mean vector \( \vec{\mu} = \vec{0} \) and a covariance matrix \( \tilde{\Sigma}_d > \tilde{0} \) (i.e., all the eigenvalues are positive). We can adopt this constraint for the sake of simplicity. Let \( \lambda_i (i = 1, \ldots, d) \) be the eigenvalues of \( \tilde{\Sigma} \), and sort them out as
\[ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d (> 0) . \] (4)

Then, we can make an eigenvalue decomposition of
\[ \tilde{\Sigma}_d = \tilde{H} \tilde{\Lambda} \tilde{H}^\top \] (5)
\[ \tilde{\Lambda} \equiv \text{diag} (\lambda_1, \lambda_2, \ldots \lambda_d) . \] (6)

Here, \( \tilde{H} = [\vec{h}_1, \ldots, \vec{h}_d] \) is an orthogonal matrix of the corresponding eigenvectors. Note that these are all outrageously huge matrices.

Now we draw a set of \( n \) samples from this parent population \( (d > n) \), \( \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \). Suppose that we describe this set of data by a \((d \times n)\) data matrix
\[ \tilde{X} = (\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n) , \] (7)
and \( \vec{x}_i (i = 1, \ldots, n) \) are independently and identically distributed (i.i.d.). We should note that the i.i.d. condition is not always satisfied in practice. However, this condition is often required for the rigorous construction of the mathematical methodology. The sample covariance matrix \((d \times d)\) is expressed as
\[ \tilde{S} = \frac{1}{n} \tilde{X} \tilde{X}^\top . \] (8)

Here, we define the dual sample covariance \((n \times n)\),
\[ \tilde{S}_D \equiv \frac{1}{n} \tilde{X}^\top \tilde{X} . \] (9)

A schematic description of a sample covariance matrix \( \tilde{S} \) and its dual matrix \( \tilde{S}_D \) is shown in Fig. 1. If we consider the sample eigenvalues of \( \tilde{S}_D \), its \( n \) eigenvalues
\[ \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n > 0 \] (10)
are common with the first $n$ eigenvalues of $\tilde{S}$. See Appendix A for the merit of $\tilde{S}_D$.

An important piece of the initial exploration of HDLSS asymptotics in Hall et al. (2005) was the revelation of a surprisingly rigid geometric structure that naturally emerges from the asymptotics in an increasingly random setting, referred to as the geometric representation by them. Recall the eigenvalue decomposition of the covariance matrix $\tilde{\Sigma} = \tilde{H}\tilde{\Lambda}\tilde{H}^\top$. Let $\tilde{X} \equiv \tilde{H}\tilde{\Lambda}^{1/2}\tilde{Z}$.

Then, $\tilde{Z}$ is a $(d \times n)$ sphered data matrix from a distribution with the identity covariance matrix. Here, we write $\tilde{Z} = [\tilde{z}_1, \ldots, \tilde{z}_d]^\top$ and $\tilde{z}_i = (z_{i1}, \ldots, z_{in})^\top$, $(i = 1, \ldots, d)$. Note that $E(z_{ij}z_{i'j}) = 0$ ($i \neq i'$) and $V(\tilde{z}_i) = \tilde{I}_n$, where $E$ and $V$ denote the expectation and variance of a random variable, and $\tilde{I}_n$ is the $n$-dimensional identity matrix. We assume that the fourth moments of each variable in $\tilde{Z}$ are uniformly bounded. Note that if $\tilde{X}$ is Gaussian, the $z_{ij}$s are i.i.d. standard normal random variables. We also define the eigenvalue decomposition of $S_D$ by

$$\tilde{S}_D = \sum_{i=1}^{n} \hat{\lambda}_i \tilde{u}_i \tilde{u}_i^\top.$$  

Here, $\tilde{u}_i$ denotes a unit eigenvector corresponding to $\hat{\lambda}_i$. See Appendix B for details of geometric representations of $\tilde{S}_D$. Let $\hat{h}_i$ be the $i$-th eigenvector of $\tilde{S}$. Note that $\hat{h}_i$ can be calculated by

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

2.2. High-Dimensional Principal Component Analysis

We adopt the same setting as the previous sections for the data distribution: we suppose a random variable with a mean $\tilde{\mu} = \tilde{0}$ and a covariance matrix $\tilde{\Sigma}_d > 0$. We assume the following model, referred to as “the generalized spiked model”, for the eigenvalue distribution of $\tilde{\Sigma}_d$.

$$\lambda_i = \begin{cases} a_id^{\alpha_i} & \text{for } i = 1, \ldots, m \\ c_i & \text{for } i = m + 1, \ldots, d \end{cases}$$
where \( a_i(>0), c_i(>0) \) and \( \alpha_i(\alpha_1 \geq \cdots \geq \alpha_m > 0) \) are unknown constants that preserves the order \( \lambda_1 \geq \cdots \geq \lambda_d \), and \( m \) is an unknown positive fixed integer\(^{10}\). The \( m \) stands for the number of spiked eigenvalues in the sense of quadratic contribution ratio. The details of the spiked model including the generalized spiked model was considered and given in Yata & Aoshima (2013). The eigenvalues \( \lambda_i(i \leq m) \) describe the intrinsic information of the data, and \( \lambda_i(i \geq m + 1) \) represent the noise. Note that

\[
\sum_{i=m+1}^{d} \lambda_i^2 \left( \sum_{i=m+1}^{d} \lambda_i \right)^{-2} \xrightarrow{P} 0 \quad \text{as} \quad d \to \infty
\]

i.e., the noise satisfies the sphericity condition [eq. (B1)]. We set the sample covariance matrix \( \tilde{S} = n^{-1} \tilde{X} \tilde{X}^\top \) as above. Here, we define \( n(d) \) as the size of a sample depending on \( d \).

### 2.2.1. Limitation of the traditional principal component analysis (PCA)

Yata & Aoshima (2009) investigated the properties of the traditional PCA when it is applied to HDLSS data, and demonstrated the condition so that the estimation by the traditional PCA would have the consistency in the form of \( n(d) = d^\gamma \)\(^{11} \) for a positive constant \( \gamma \) and the limitation of the traditional PCA. They have given the following two theorems on the eigenvalues, depending if the condition would satisfy

\[ z_{ij} \quad (i = 1, \ldots, d, \; j = 1, \ldots, n) \text{ are independent.} \]  

\[ \text{Theorem 2.1} \; [\text{Yata & Aoshima (2009)} \; \text{I}] \]

\[ \text{If } Z \text{ satisfies Condition } (16), \text{ for } i \leq m, \]

\[ \frac{\hat{\lambda}_i}{\lambda_i} = 1 + o_P(1), \]  

under the conditions

1. \( d \to \infty \) and \( n \to \infty \) for \( i \) that satisfies \( \alpha_i > 1 \).

2. \( d \to \infty \) and \( d^{1-\alpha_i}/n(d) \to 0 \) for \( i \) that satisfies \( \alpha_i \in (0, 1] \).

See Appendix C for details of the consistency when \( Z \) does not satisfy Condition (16).

Theorem 2.1 gives the condition that the estimation by the traditional PCA would have the consistency. The condition described by \( d \to \infty \) and \( n \to \infty \) in (1) of these theorems is a mild condition in the sense that one can choose \( n \) free from \( d \) or that \( n \) may be much smaller than \( d \) such as \( n = \log d \). However, in (2), \( n \) is heavily depending on \( d \). Hence, we should be very careful to apply the conventional PCA to the HDLSS data. In order to overcome this fundamental problem of the conventional PCA, Yata & Aoshima (2010a) and YA12 proposed two new PCAs for HDLSS data. In this work, we focus on one of them, the noise-reduction methodology (YA12).

\(^{10}\) One may feel suspicious if such a restrictive model can really describe the general properties of eigenvalue distribution of real data. However, as we see in §§ 4 and 5, the generalized spiked model really well represents the actual eigenvalues.

\(^{11}\) In this context, consistency stand for the situation as follows: Consider an estimator \( t(n) \) (\( n \): sample size) is a consistent estimator for a parameter \( \theta \) if and only if, for all \( \epsilon > 0 \), we have

\[ \lim_{n \to \infty} \mathbb{P}\{|t(n) - \theta| < \epsilon\} = 1. \]
Figure 2. A schematic description of the high-dimensional principal component analysis (PCA). Even if we have significant important features, because of the existence of a huge noise sphere, they are completely embedded in the observed data. The high-dimensional PCA can subtract the noise sphere and enables to extract the desired features.

2.2.2. Noise-reduction PCA (NRPCA): a new method for HDLSS data

Here we introduce the high-dimensional PCA. First we informally present the concept of the high-dimensional PCA, which is schematically described in Fig. 2. Because of the enormous noise that concentrates on the sphere in the data space, the intrinsically important characteristics of the data are completely buried under the noise sphere and obscured. However, it is not hopeless to estimate the characteristics even under such a situation, thanks to the fact that the noise behavior around the sphere can be well analyzed and essentially subtracted. The method to concretely realize this idea is the noise-reduction (NR) methodology proposed by YA12.

We try to explain it in a more rigorous manner. YA12 proposed the NR methodology derived from the geometric representation [eq. (B2)]. First we decompose $\tilde{S}_D$ as

$$n\tilde{S}_D = \sum_{i=1}^{m} \lambda_i \tilde{z}_i \tilde{z}_i^T + \sum_{i=m+1}^{d} \lambda_i \tilde{z}_i \tilde{z}_i^T \tag{18}$$
The second term represents the noise. Then, from Theorem B.1, under Condition (16), we obtain the following geometric representation for the noise

\[
\sum_{i=m+1}^{d} \lambda_j \tilde{z}_i \tilde{z}_i^T \xrightarrow{P} \tilde{I}_n \quad \text{as} \quad d \to \infty,
\]

i.e., the noise part becomes deterministic under the HDLSS condition. This result enables us to remove the noise part and obtain the formula

\[
\hat{\lambda}_i = \lambda_i - \frac{1}{n-i} \left( \text{tr} \tilde{S}_D - \sum_{j=1}^{i} \hat{\lambda}_j \right) \quad (i = 1, \ldots, n-1).
\]

Estimation of \( m \) is given by Aoshima & Yata (2018). We have the following theorem.

**Theorem 2.2** [Yata & Aoshima (2012) II].

When the components of \( \tilde{Z} \) satisfy Condition (16), it holds that for \( i \leq m \),

\[
\frac{\hat{\lambda}_i}{\lambda_i} = 1 + o_p(1),
\]

under the conditions that

1. \( d \to \infty \) and \( n \to \infty \) for \( i \) such that \( \alpha_i > 1/2 \),

2. \( d \to \infty \) and \( d^{1-2\alpha_i}/n(d) \to 0 \) for \( i \) such that \( \alpha_i \in (0, 1/2] \).

If we compare Theorem 2.2 with Theorem 2.1, we see that the consistency holds for \( \hat{\lambda}_i \)'s under milder conditions than \( \hat{\lambda} \) [under Condition (16)] \(^{12}\).

Next, we apply the NR methodology to the PC direction vectors. Define \( \tilde{h}_i \equiv (n\hat{\lambda}_i)^{-1/2} \tilde{X} \hat{u}_i \), and examine its performance as the estimator of the PC direction vector \( \tilde{h}_i \). Note that \( \tilde{h}_i = (\hat{\lambda}_i/\lambda_i)^{-1/2} \tilde{h}_i \). YA12 showed the following theorem.

**Theorem 2.3** [Yata & Aoshima (2012) III]

Assume \( \lambda_i \) (\( i \leq m \)) has multiplicity one. Under conditions in Theorem 2.2, and the components of \( \tilde{Z} \) satisfy Condition (16), it holds that

\[
\tilde{h}_i^\top \tilde{h}_i = 1 + o_p(1).
\]

We also consider the PC scores. The \( i \)-th PC score of \( x_i \) is given by \( \tilde{h}_i^\top \tilde{x}_j = \tilde{z}_{ij} (\lambda_i)^{1/2} \equiv \tilde{s}_{ij} \). Let \( \tilde{u}_i \equiv (\tilde{u}_{i1}, \ldots, \tilde{u}_{in})^T \) (\( i = 1, \ldots, d \)). Consider we estimate the PC score by \( \tilde{u}_{ij} (n\lambda_i)^{1/2} \equiv \tilde{s}_{ij} \). We define the mean squared error (MSE) of the \( i \)-th PC score by

\[
\text{MSE}(\tilde{s}_i) \equiv \frac{1}{n} \sum_{j=1}^{n} (\tilde{s}_{ij} - s_{ij})^2.
\]

Then we have yet another important theorem.

\(^{12}\) If \( \alpha_i > 1/2 \), the NR methodology has the consistency even when \( n \) is much smaller than \( d \) such as \( n = \log d \).
Theorem 2.4 [Yata & Aoshima (2012) V]
Assume that \( \lambda_i \) (\( i \leq m \)) has multiplicity one. Then, under the same conditions as Theorem 2.3,
\[
MSE(\hat{s}_i) = \frac{1}{\lambda_i} = o_p(1)
\]
holds.

All these theorems hold for the cases with nonzero mean, by replacing the setting with zero-centered variables. Thus, now we have demonstrated that the NR methodology yields high performances for the estimation of the PC eigenvalues (contributions), PC vectors and PC scores for HDLSS data. In the following discussions, the presented eigenvalues are obtained by the NRPCA.

2.2.3. Automatic Sparse PCA (A-SPCA)

In this section, we introduce (thresholded) sparse PCA methods for high-dimensional data. The sparse PCA (SPCA) can be summarized as follows: Let \( \hat{\vec{h}}_i = (\hat{h}_i(1), \ldots, \hat{h}_i(d))^\top \) for all \( i \). Given a sequence of threshold values \( \zeta > 0 \), define the thresholded entries as
\[
\hat{h}_{i(s)}^\ast = \begin{cases} 
\hat{h}_{i(s)} & \text{if } |\hat{h}_{i(s)}| \geq \zeta, \\
0 & \text{if } |\hat{h}_{i(s)}| < \zeta
\end{cases}
\]
for \( s = 1, \ldots, d \).

Let \( \hat{h}^\ast_i = (\hat{h}^\ast_i(1), \ldots, \hat{h}^\ast_i(d))^\top \). The thresholded estimator of \( \vec{h}_i \) is defined by
\[
\hat{\vec{h}}^\ast_i = \hat{h}_i / \|\hat{h}_i\|.
\]
However, the estimator heavily depends on a choice of \( \zeta \) and does not hold consistency properties for HDLSS data. Recently, Yata & Aoshima (2022) proposed a SPCA method by using the NR method as follows: For the PC direction by the NR method, \( \hat{\vec{h}}_i = (\hat{h}_i(1), \ldots, \hat{h}_i(d))^\top \), we assume \(|\hat{h}_i(1)| \geq \cdots \geq |\hat{h}_i(d)|\) for the sake of simplicity. For a given constant \( \omega \in (0, 1] \), we consider PC directions whose cumulative contribution ratio is greater than or equal to \( \omega \). There exists a unique integer \( \hat{k}_{i\omega} \in [1, d] \) such that
\[
\sum_{s=1}^{\hat{k}_{i\omega}-1} \hat{h}_{i(s)}^2 < \omega \quad \text{and} \quad \sum_{s=1}^{\hat{k}_{i\omega}} \hat{h}_{i(s)}^2 \geq \omega.
\]
We define that
\[
\tilde{h}_{i\omega(s)} = \begin{cases} 
\hat{h}_{i(s)} & \text{if } |\hat{h}_{i(s)}| \geq |\hat{h}_{i(\hat{k}_{i\omega})}|, \\
0 & \text{otherwise}
\end{cases}
\]
for \( s = 1, \ldots, d \).

Let \( \tilde{\vec{h}}_{i\omega} = (\tilde{h}_{i\omega(1)}, \ldots, \tilde{h}_{i\omega(d)})^\top \). Yata & Aoshima (2022) called the new PCA method that uses \( \tilde{h}_{i\omega(s)} \)'s as the “automatic SPCA (A-SPCA)”. They showed some consistency properties of \( \tilde{h}_{i\omega} \) in the HDLSS data and investigated the performance in actual data analyses. Here the A-SPCA uses the same eigenvalues estimated by the NRPCA. In this paper, we set \( \omega = 1/2 \).\(^{13}\) Intuitively, the A-SPCA can extract the most important components of the eigenvectors obtained by the NRPCA. Namely, these components practically controls the PCs.

\(^{13}\) The value of \( \omega = 1/2 \) was obtained by simulations.
3. DATA

Our target galaxy, NGC 253, is a nearby barred spiral galaxy at a distance of 3.5 Mpc, i.e., 1″ corresponds to 17 pc (Rekola et al. 2005). Morphological type Sc is assigned. It has a systemic heliocentric velocity of $243 \pm 2$ km s$^{-1}$ ($z = 0.000811$: Kamphuis et al. 2015), mainly due to the cosmological expansion. This galaxy is one of the brightest sources of far infrared emission except the Magellanic Clouds and has been extensively studied at many wavelengths. It is known as a pure starburst, whose star formation rate (SFR) in the central molecular zone is estimated to be $\text{SFR} \sim 2 M_\odot \text{[yr}^{-1}]$ (Rieke et al. 1980; Keto et al. 1999; Bendo et al. 2015). Massive, dense, and warm ISM (e.g. Sakamoto et al. 2011) were identified in the central region. Super star clusters are identified at optical and near-infrared (NIR) bands (Fernández-Ontiveros et al. 2009). An intense outflow is also observed (Matsubayashi et al. 2009; Bolatto et al. 2013). The super star clusters and outflow are also associated with the starburst activity.

3.1. ALMA map of NGC 253

The data used in this work is a spectroscopic map of a nearby archetype starburst galaxy NGC 253, obtained by A17, at Band 7. They obtained the 340–365 GHz ($\lambda \simeq 0.85$ mm) spectra covering a total frequency range of 11 GHz, compiling the results of two projects in the ALMA Cycle 2 observations: 2013.1.00735.S (PI: Nakanishi) and 2013.1.00099.S (PI: Mangum). The field of view of ALMA at these wavelengths is 16″–17″. The observations cover the frequency range of 340.2–343.4 GHz and 352.5–355.7 GHz for 2013.1.00735.S, and 350.6–352.4 GHz and 362.2–365.2 GHz are covered by 2013.1.00099.S. The uncertainty of the flux calibration in ALMA Band 7 observations is $\sim 10\%$ according to the ALMA Cycle 2 Proposer’s Guide. Data reduction is conducted with Common Astronomy Software Applications (CASA: McMullin et al. 2007), versions 4.2.1 and 4.2.2. The image of the central region of NGC 253 was created by using task clean. The synthesized beam size of the final images is $0.′′45 \times 0.′′3$, which corresponds to 8 pc $\times$ 5 pc at the distance to NGC 253. With a velocity resolution of 5.0 km s$^{-1}$, the rms noise levels are 1.0–2.4 mJy beam$^{-1}$. Various emission lines are detected in the spectra (see Fig. 2 Table 1 of A17). The number of spectroscopic resolution unit (sampling number) is 2248.

The A17 original map consists of $864 \times 864$ points in the spatial domain (R.A and dec.), and 2248 components along the frequency domain (in units of [Hz]) in total. However, since the pixel scale was smaller than the synthesized beam size of the map $0.′′45 \times 0.′′3$ (oversampled), each spatial grid in the map is not independent of each other. To overcome this issue, we regridded the map in the spatial dimension. The equivalent circular beam size corresponding to the elliptical beam of $0.′′45 \times 0.′′3$ is $0.′′37$. We used this effective beam size to regrid the image map. The resulting map with independent grids consists of $230 \times 230$ pixels. We then should cut out low-signal regions. For this, instead of estimating S/N for some lines, we integrated the intensity over the whole range of frequency and calculated the total intensity $I$ [Jy beam$^{-1}$] and chose the regions with $I > 1.0$, based on the rms level. We show this “mask” region, together with the significantly bright regions in Fig. 3. The number of remaining pixels $n$ is 231.

3.2. Doppler shifts due to the systemic Hubble velocity and rotation

The central region of NGC 253 is strongly inclined ($i \sim 78°$). This causes a fairly coherent rotation pattern on the spectroscopic map through the Doppler shift. We made a Gaussian fit to three prominent emission lines HCN(4–3), HNC(4–3), and CS(7–6) in the spectra to estimate the line center wavelengths. The map of the velocity field is presented in Fig. 4. We determined the amount of the shift by averaging the shifts obtained from three lines. Then we shifted the frequencies with respect to the systemic velocity of
Figure 3. The bright regions of NGC 253 map cut out by the mask. Left: the mask region map. White regions have significant intensity signals. Center: the cut-out region with significantly bright emission. Right: the bird’s view of the signal.

Figure 4. The velocity map of NGC 253 estimated from HCN(4–3) (Top), HNC(4–3) (Middle), and CS(7–6) (Bottom). The systemic velocity of 243 \( \text{[km s}^{-1} \text{]} \) is subtracted.

NGC 253 caused by the Hubble velocity. After this correction, we have a map without the systematic effect of the coherent rotation and we can analyze more subtle properties of the submillimeter emission lines.
Figure 5. The eigenspectra corresponding to the 1st–5th principal components (PCs) constructed from the original ALMA spectral map of NGC 253. Upper five panels: eigenspectra obtained by the noise-reduction principal component analysis (NRPCA). Lower five panels: same as Upper panels, but obtained by the automatic sparse principal component analysis (A-SPCA). Only the controlling features have nonzero values.
Figure 6. Normalized eigenvalues (contributions) obtained from the analysis of the ALMA spectroscopic map of NGC 253 by the NRPCA.

We demonstrate an example of the emission lines in Fig. 22 in Appendix E. We have frequency gaps in our ALMA data cube. If we try to shift the spectrum according to the velocity, the boundaries of the spectra are also shifted. We simply adopted frequency ranges common to all the pixels. The final size of the data is 231 along the spatial position, and 1971 in spectral dispersion after the all corrections.

3.3. Spectral map as HDLSS data

The resulting map of NGC 253 has spatial dimension $231 \times$ spectral dimension $1971$, i.e., $n = 231$ and $d = 1971$, and clearly $n \ll d$. Though it is not as extreme as DNA microarray data (typically $n \sim 10–100$ and $d \sim 10^4–10^5$), our ALMA map is typical HDLSS data. Problems from astrophysical side is that a very large amount of information is contained in the spectra, and a large variety of spectral lines are observed compared to $n$. This situation is not special in the high-dimensional analysis.

However, the application of the high-dimensional methods to spectroscopic data is not completely straightforward. Spectral lines are broadened and/or shifted by a turbulent motion and other physical mechanisms in the ISM. Then, the information carried by each spectral resolution unit is not independent. Such a complication has never been examined in other field of science about HDLSS data. However, since it is commonly seen in spectroscopic data in any field of astronomy, before going to the detailed analysis, we first examine if the high-dimensional statistical analysis really works for the astronomical spectroscopic map, with line shifts and broadening.

4. PREPARATORY ANALYSIS

As we have already mentioned above, there are some potential problems specific to astronomical spectroscopic data, which are different from typical examples of HDLSS data in genomics and other fields. The most fundamental issue is that, when we regard a spectroscopic data as a vector data $\vec{x} = (x_1, \ldots, x_d)^T$, the information of neighboring frequency (or wavelength) index $i$ and $i + 1$ ($i = 1, \ldots, d - 1$) is not independent. Namely, we often have internal motion in the ISM of a galaxy that leads to a Doppler broadening along the frequency axis. We also find a systemic rotation of the system over the whole observed region of

---

14 Note that $n > \log d = 7.59$ when $d = 1971$ and $n = 231$, so that the NR methodology works well for the HDLSS data. See Footnote 12.

15 The existence of continuum radiation makes the situation even more complicated, but we do not try to examine this problem in the current work.
an object (§ 3). In order to examine if the high-dimensional statistical analysis would work for such type of astronomical spectroscopy data, we made a preparatory analysis of the original spectral mapping data of NGC 253 with the high-dimensional PCA. By the term “original”, we mean that we started this analysis without correcting the Doppler shift on the map due to the internal systemic rotation. Namely, we blindly applied the high-dimensional PCA methods to the data and examined what feature would be detected. For this, we sort the 2-dimensional coordinates along the R.A.- direction and then dec. direction, and renumbered so that we can regard the data as a 1-dimensional vector. Though we can apply the PCA directly to a 2-dimensional image, we adopted this procedure throughout this paper, as a first simplistic analysis.

4.1. Eigenspectra

In the following part of this work, we mainly focus on the A-SPCA results. However, since the NRPCA deal with all the spectral features without extracting only the most important ones, we also present the eigenspectra (eigenvectors constructed from the observed spectra by PCA) from the NRPCA. Upper panel of Fig. 5 shows the eigenspectra constructed by the NRPCA from the ALMA spectral map of NGC 253. We observe that specific features are commonly found in similar frequency ranges of the spectra for all the PCs. This suggests that some spectral features contain richer information on the whole spectra than other spectral regions.

The A-SPCA specifies and localizes the most important components that correspond to these particular features. Lower panel of Fig. 5 shows the eigenspectra obtained by the A-SPCA. As seen in Fig. 5, the A-SPCA only leaves the values if the features are important to determine the spectra, and forces the other components to be zero. Hence, the A-SPCA eigenspectra are not used for the reconstruction of the spectra.

\[\text{Figure 7. The distribution of PC1 and PC2 of the ALMA map of NGC 253 by the A-SPCA.}\]

\[\text{Figure 7. The distribution of PC1 and PC2 of the ALMA map of NGC 253 by the A-SPCA.}\]
with all features, and the NRPCA is used for this purpose. We make use of this outstanding advantage of the A-SPCA to specify the controlling spectral features of the PCA in the following analysis.

4.2. Distribution of the contribution of PCs

First we show the eigenvalue distribution of the PCs sorted with their contribution obtained by the NRPCA in Fig. 6. First several eigenvalues are prominently larger than the rest. This behavior is referred to as “spiked” in the context of the high-dimensional statistical analysis. Recall that the eigenvalues of the PCA represent the contribution of each PC, namely, the PCA decomposes the whole covariance matrix into contributions from each PC. Since the obtained eigenvalue distribution shows this spiked structure, we see that the high-dimensional PCA detected some characteristics of the ALMA spectral map of NGC 253. This guarantees that the current data can be safely regraded as typically HDLSS, and that the assumptions of the high-dimensional PCA apply.

Astrophysically, Fig. 6 demonstrates a very promising possibility of the high-dimensional statistical analysis to the spectral mapping data. A17 identified more than 30 molecular/radical lines on the map (see their Fig. 1), as well as some broad or continuum features. Even if we restrict our interest to the lines, it is implausible that we classify them only by physical intuition. Traditionally, for example, a line ratio has been used to extract physical information from spectral data. However, brute-force trials may not work if we try to diagnose many lines simultaneously, because the number of line ratios would increase prohibitively with a number of features, known as a combinatorial explosion. In a clear contrast, Fig. 6 implies that the whole information of the complicated spectral map can be represented by first several PCs. This means that the complicated spectral features including more than 30 lines are controlled by several bases represented by the PCs. This advantage itself is a well-known benefit to use the PCA in general as a dimensionality reduction method (e.g., Galaz & de Lapparent 1998; Ronen et al. 1999; Wang et al. 2011; Pace et al. 2019; Portillo et al. 2020, among many others), but all these previous applications were on integrated 1-d spectra with sufficiently large number of \( n \), namely, not HDLSS data. We stress that the analysis in this work is substantially different from these works: the current method is able to be applied to a spectral map which is a typical HDLSS dataset and traditional PCA would not have worked due to the huge noise sphere, as mentioned in Section 2.

Since we confirmed that the methodology of high-dimensional statistical analysis worked well for the astrophysical spectral map data, we safely proceed to further analysis by the high-dimensional PCA.

4.3. Physical meaning of first two PCs

Next step is to examine what PC1 and 2 represent. Figure 7 shows the scatter of PC1 and PC2. The most striking feature is the butterfly-like pattern symmetric with respect to the \( x \)-axis. It reminds us of the coherent rotation of the central region of NGC 253. We reconstructed the 2-dimensional structure to map PC1 and 2 in Fig. 8.

It is clear that PC1 represents the intensity of lines, and PC2 represents the symmetric and coherent pattern of the center of NGC 253. We observe a nice agreement between Lower panel of Fig. 8 with directly estimated velocity field of the same region in Fig. 4.

To have a deeper look at the relation, we present the correlation between PCs and observed properties in Fig. 9. Upper panel of Fig. 9 is the scatter plot between PC1 and the total intensity integrated over the whole frequency range for each pixel. We indeed observe a good correlation between them (Fig. 9). The correlation is not linear, because the PCA extracts only the linear features of the data, while the relation to the physical quantities would not be necessarily linear. We also examined the relation between the peculiar velocity and
The 2-dimensional structure to map PC1 and PC2 obtained by the A-SPCA. Left panel shows the map of PC1, and Right panel describes the map of PC2, respectively.

The scatter plot between PCs and observed physical properties. Left: the correlation between PC1 and the integrated intensity over the whole frequency range $I$ [Jy beam$^{-1}$] on the ALMA map of NGC 253. Right the correlation between PC2 and the peculiar velocity on the ALMA map.

PC2 in Right panel of Fig. 9. As we already presented in 2-dimensional map, PC2 approximately delineate the velocity field in Fig. 9. However, we see some outliers where the PC2 and the peculiar velocity are not coherent. These regions may be affected by some local phenomenon, for example, a strong outflow. We will revisit this issue later with a detailed analysis.

4.4. Spectral features corresponding to PCs: original data

It is interesting to examine what features are concretely responsible for PCs. Actually the A-SPCA can specify the responsible features to characterize the PCs. We show the responsible spectral features in Figs. 10. The spectra are integrated over the whole analyzed regions for both figures. In Fig. 10, stars represent the features responsible for PC1, and triangles are for PC2. We do not show higher-order contributions for clarity. PC1-related features are assigned to the HCN(4–3) and HNC(4–3) lines, particularly
Figure 10. Responsible features to characterize PCs from the A-SPCA for the ALMA map of NGC 253. Stars represent the features responsible for PC1, and triangles are for PC2.

close to the central part of the lines. PC2-related features are related to HCN(4–3), but in contrast to PC1, they are connected to the wing part of the line. Since the spectra are integrated over the analyzed regions, the wing is made actually by the Doppler shifts. This is clearly consistent with the fact that PC2 map (Fig. 8) agrees well with the Doppler velocity field of NGC 253 map (Fig. 4).

The performance of the A-SPCA to the ALMA spectroscopic map as HDLSS data is certainly guaranteed by this preparatory analysis. We can safely proceed to the detailed analysis of the molecular lines, as in the next Section.

5. RESULTS AND DISCUSSION

After examining the performance of the high-dimensional statistical analysis in § 4, now we can proceed to a further physical analysis.

5.1. Application of the high-dimensional PCA to the Doppler shift-corrected map of NGC 253

We apply the high-dimensional PCA to the Doppler-corrected map of NGC 253, exactly in the same manner as we did in § 4. Note that, though we use the same notation for PC2, now PC2 is the value obtained from the Doppler-corrected map and different from those in § 4 from the NRPCA. We should also note that since we applied the NRPCA and A-SPCA to the Doppler-corrected map, all the contributions of PCs are affected by the correction.

Same as the preparatory analysis, we first show the eigenspectra of the Doppler-corrected spectral map in Fig. 11. Comparing with Fig. 5, the meaning of the shape of each eigenspectrum becomes clearer. We stress that since we have already subtracted the global shift of the whole spectra, it does not represent the systemic velocity of the position anymore. Thus, the velocity structure is always the first-order deviation from the systemic rotation pattern, and appears as a broadening or asymmetric distortion of the line profiles. Another important difference is that an important spectral line is identified in PC5 (see PC5 of Bottom panel in Fig. 11) which was not specified from the original data. This is because the dominant effect of the
Figure 11. The eigenspectra corresponding to the 1st–5th principal components (PCs) constructed from the Doppler-corrected ALMA spectral map of NGC 253. Upper five panels present the eigenspectra obtained by the NRPCA, and lower five panels show those obtained by the A-SPCA. Systemic rotation was successfully removed from the map, and more subtle features appear more prominent.
Figure 12. The eigenspectra corresponding to the 1st–5th principal components (PCs) constructed from the Doppler-corrected ALMA spectral map of NGC 253. Upper five panels present the eigenspectra obtained by the A-SPCA with 0.25, and Lower five panels shows that obtained by the A-SPCA with 0.75.

in the Doppler-corrected map. Now very clearly PC2 represents the wing of the HCN and some other lines,
indicating that it reflects a very local expansion of the gas. PC3 and 4 describe the blueward distortion of the lines, caused by a gas motion toward us faster than the expansion velocity. Though PC5 is less prominent, but similarly to PC3 and 4, it represents the redward distortion of the lines.

We present the distribution of the contribution of PCs in Fig. 13. Now the contribution of new PC2 is much smaller than old PC2 in Fig. 6. This is also because of the successful removal of the systemic rotation. Instead, we have a smaller but non-negligible contribution of new PC2. Now PC2 would describe much more subtle and local features of the spectral line map of NGC 253, which is consistent with the implications from the eigenspectra.

This is more drastically represented on the scatter plot of PCs. We demonstrate the relations between PCs in Figs. 14–16. Figure 14 shows the 2-dimensional distribution of PC1 and PC2. The butterfly-like pattern seen in Fig. 7 completely disappeared. Figures 15 and 16 are the relations between PC1 and PC3, and PC2 and PC3, respectively. Again we do not find any symmetric pattern in Figs. 15 and 16. From these figures, we do not have an obvious symmetric pattern in the intrinsic spectral features in the map. Both in 2- and 3-d, we observe a continuous distributions among PCs.

5.2. Map of PCs of the NGC 253 spectra

We reconstructed the 2-dimensional structure to map PC1, 2, and 3 in Fig. 17. Again, PC1 represents the intensity of lines. In contrast to Fig. 8, new PC2 shows a more complicated, localized pattern in NGC 253. The map of PC3 also shows smaller scale structures. Recalling that PC2 suggest the existence of small-scale outflow. According to A17, the spatial scale (diameter) of the most prominent region with large PC2 is $\sim 10$ pc, namely it corresponds to a small-scale phenomenon, indeed.

Particularly interesting is the region with negative PC3 in Bottom panel of Fig. 17. The eigenspectra show that PC3 and partially PC4 describes a local but larger-scale flow toward us, which appears in blueshift of the lines. Bolatto et al. (2013) reported a global molecular outflow from the center of radio continuum. More recently, Walter et al. (2017) and Krieger et al. (2019) performed detailed ALMA observations of this region. They made a detailed analysis and named this outflow as “SW streamer”. This outflow is going out toward us, and the position of the root of the SW streamer precisely agrees with the region with negative PC3 in Fig. 17 (Krieger et al. 2019, see their Fig. 1). The anomalous region specified by the peculiarity in velocity in Fig. 18 also agrees well with the root of the SW streamer. In addition to the
Figure 14. The 2-dimensional structures of the distribution of PC1 and PC2 of the NGC 253 obtained by the A-SPCA, after correcting the effect of the Doppler shifts of the systemic Bottom: same as Top panel, but with PC1 and PC3. Note that the PC2 in this figure is different from that in Fig. 8.

Figure 15. Same as Fig. 14, but for PC1 and PC3.
molecular observations, an outflow is also observed in Hα and other optical emission lines by a Fabry-Perot spectroscopy (Matsubayashi et al. 2009). Though this delineates the flow of ionized gas, the position, shape, and direction of the outflow agree fairly well with the molecular outflow. All these clues indicate that the high-dimensional PCA efficiently extract an outflow phenomenon from an ALMA spectroscopic map, in a purely objective way.

5.3. Spectral features of the Doppler-corrected data corresponding to the PCs

In Fig. 19, stars and triangles represent PC1- and PC2-related responsible spectral features similarly to Fig. 10, and filled circles are PC3-related. First we observe a global shift of lines compared to Fig. 10. This is, of course, due to the recession velocity. Since the boundary between ALMA Bands are fixed at the observer’s frame, some parts of the spectra go out of the boundary after the Doppler shift correction. As a result, we lose significant parts of the spectra in Fig. 19. After the Doppler correction, PC1 describes the intensity of HCN(4–3) and HNC(4–3) lines more specifically, which is reflected to the fact that responsible features concentrate on the peak of these lines. PC2 is assigned to the part of the profile slightly far from the line center for HCN(4–3). However, it is related to the lower-frequency side of HNC(4–3) line, though it is difficult to see clearly because this side is cut out by the correction. PC3 delineate further side of the higher-frequency side of HCN(4–3) line, while it has some overlap with PC2-related features on the lower-frequency side. These properties were already suggested from eigenspectra (Fig. 11), and we confirm by the detailed information from the A-SPCA. Thus, even if the spectra look formidably complicated, the high-dimensional PCA methods like NRPCA and A-SPCA can disentangle the information and pick up the controlling spectral features from the spectroscopic map. Rather surprisingly, the detailed profile of HCN and HNC lines govern almost all spectral characteristics of the ALMA map of NGC 253, including outflow phenomenon. Further investigation on detailed astrophysical processes from the ALMA map of NGC 253 is left to our future work.
Figure 17. The 2-dimensional structure to map PC1, PC2, and PC3 obtained by the NRPCA.

Figure 18. The region with velocity peculiarity in Fig. 9.

6. CONCLUSION

Evolution of galaxies is mainly driven by the star formation, a transition from ISM to stars. Various phases of the ISM are related, and the evolution of the ISM is a key to complete the understanding of the galaxy
evolution. Spectroscopic observations are of vital importance to extract and interpret the information of matter in galaxies.

Though spectroscopic mapping and integral field spectroscopy are fundamentally important to reveal the ISM physics, detailed observations are often very time-consuming, and it is not easy to map objects to obtain many independently sampled measurements. Consequently, we often meet a situation where the dimension of data $d$ is much larger than the sample size $n$, i.e., $n \ll d$. This is referred to as high-dimensional low sample size (HDLSS). The HDLSS data are different from typical big data in astronomy (usually $n$ is huge), but $d$ is extremely large. In traditional studies in astronomy, the HDLSS problem has been regraded as being ill-posed and simply given up to analyze. However, in the end of the last millennium, a strong need of an analysis method for such data has been arisen (e.g., Golub et al. 1999). Then, during the subsequent decade, the new statistical method to tackle HDLSS problems has been vastly developed, known as the high-dimensional statistical analysis (e.g. Aoshima & Yata 2017; Aoshima et al. 2018). This paper is organized twofold. First we introduced the general framework of high-dimensional statistical methods, and then we applied the methodology to the ALMA spectroscopic map data.

For HDLSS data, the full information of the sample covariance is expressed by a huge $d \times d$ covariance matrix, which is implausible to handle in various aspect. This specific difficulty of HDLSS data is often referred to as “the curse of dimensionality”. The high-dimensional statistical analysis overcomes this problem by introducing the dual representation of sample covariance matrix which is a much smaller, $n \times n$ dimension square matrix. The remarkable property of the dual matrix is that it contains practically the same information on the first $n$ eigenvalues of the original sample covariance matrix. Nevertheless, the dual sample covariance matrix is small enough to handle without the curse of dimensionality. In addition to this most difficult issue, many unusual behaviors are known for HDLSS data are known, such as strong inconsistency, spherical concentration for Gaussian variables, and axis concentration for non-Gaussian variables.
The high-dimensional statistics is the framework to make use of these specific aspect of the HDLSS data and construct a high-dimensional counterpart of traditional methods.

After introducing the concept of the HDLSS data and corresponding methodology, we applied the high-dimensional version of the principal component analysis (PCA) on the NGC253 ALMA spectral map (Ando et al. 2017). ALMA mapping data are typically HDLSS in general, and as for the data in this work, $n = 231$ and $d = 1971$, clearly $n \ll d$ even though it is less extreme than the case in genomics, for example. The ALMA spectra are very rich in various molecular lines, and there is a very large variety in the molecular line intensities (Ando et al. 2017).

We first applied the noise-reduction PCA (NRPCA) and automatic sparse PCA (A-SPCA), one of the latest methodology in high-dimensional statistical analysis, to the original ALMA map of NGC 253. We found that the high-dimensional PCA worked very well, and it could describe the global trend of the spectra only by the first two PCs (the contribution of the 1st and 2nd PCs is $> 50\%$). PC1 approximately represents the total intensity of the spectra, and PC2 describes the Doppler shifts due to the systemic rotation of the observed region. Each PC consists of $\sim 5$–$20$ elements, much fewer than $d$. We note that practically the number of the controlling factors reduces further, since these contributing elements are part of the same spectral features. The controlling features were HCN(4–3) and HNC(4–3) rotational lines. This result strongly guarantees that the high-dimensional statistical analysis methodology works appropriately to the spectral map as HDLSS data.

Then we applied the high-dimensional PCA to the Doppler-corrected ALMA map. The resulting eigen-spectra of the NRPCA correspond to more specific spectral features, especially to the core or wings of the emission lines. While PC1 again represent the total intensity of the spectral map, PC2 appear to delineate a 10-pc scale outflow phenomenon of molecular gas. More impressive is that PC3 and 4 represent a larger-scale coherent flow causing a blueshift of emission lines, and PC5 the redshift. The region of the ALMA map where PC3 is negative concentrates around the position of the radio continuum of NGC 253. This map then implies that the blueshift reflects the global larger-scale outflow from the radio center of the galaxy moving toward us. To examine this, we compared the region with the observed molecular outflow reported by Krieger et al. (2019). We find that the negative PC3 region agrees well with the root of the molecular outflow, named “SW streamer”. This was also confirmed by the anomalous region of the velocity field in the Doppler-uncorrected map. Further this region also give a close agreement with the outflow of ionized gas (Matsubayashi et al. 2009). All these suggest that the PCA indeed extracted the dynamical features from the ALMA spectroscopic map, a typical HDLSS data.

We stress that this result is not obtained by choosing a handful of features by hand, but by making use of the full information of the high-dimensional data. Spectroscopic maps are typical HDLSS data, and we can apply the high-dimensional statistical analysis on various unsolved problems in astrophysics. As mentioned in § 1, the condition also applies to a spectroscopic mapping with satellite instruments, mainly due to the strong limitation of observation time and mission lifetime. The high-dimensional methodology will shed light to the spectral map of future space telescopes. For example, such surveys of nearby galaxies with SPICA were discussed in details. Though SPICA has been cancelled, this methodology will remain useful for a future and/or possible replacement project. The high-dimensional statistical analysis can also be applied any type of data with small sample size and large dimension, e.g., spectroscopy of rare objects.
ACKNOWLEDGMENTS

We deeply thank the anonymous referee for her/his careful reading of the manuscript and constructive comments that improved the clarity of the paper much. This paper makes use of the following ALMA data: ADS/JAO.ALMA#2013.1.00099.S, ADS/JAO.ALMA#2013.1.00735.S. ALMA is a partnership of ESO (representing its member states), NSF (USA) and NINS (Japan), together with NRC (Canada), MOST and ASIAA (Taiwan), and KASI (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ. This work has been supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (19H05076, 21H01128, 19K03937, and JP17H06130). This work has also been supported in part by the Sumitomo Foundation Fiscal 2018 Grant for Basic Science Research Projects (180923), the Collaboration Funding of the Institute of Statistical Mathematics “New Perspective of the Cosmology Pioneered by the Fusion of Data Science and Physics”, and the NAOJ ALMA Scientific Research Grant Number 2017-06B.

SC has been financially supported by the JSPS as a research fellow (DC1). We deeply thank the SPICA Science Team, especially members of the SPICA the Galaxy and Nearby Galaxies Science Working Group (Fumi Egusa, Hanae Inami, Hiroyuki Kaneko, Itsuki Sakon, Yoichi Tamura, Junichi Baba, Kentaro Motohara, Yoshimasa Watanabe), to which TTT and KN belong. We are also grateful to Yugo Nakayama, Shiro Ikeda, and Nanase Harada for fruitful discussions.

Facilities: ALMA
Software: astropy (Astropy Collaboration et al. 2013, 2018), CASA (McMullin et al. 2007), TOPCAT (Taylor 2005)

APPENDIX

A. THE MERIT OF THE DUAL SAMPLE COVARIANCE

We note that the sample covariance matrix \( \tilde{S} \) and the dual sample covariance \( \tilde{S}_D \) share their non-zero eigenvalues. We also note that the eigenvectors of \( \tilde{S} \) can be calculated by the eigenvectors of \( \tilde{S}_D \). See eq. (13). Thus the merit of the dual sample covariance \((n \times n)\) is rather obvious:

1. We can reduce the computational cost significantly.
2. We can visualize the behavior of data easier than the original data matrix \((d \times d)\).
3. Asymptotic theory can be developed on the dual space.

B. GEOMETRIC REPRESENTATIONS OF \( \tilde{S}_D \)

In this section, we give geometric representations of \( \tilde{S}_D \) when \( d \rightarrow \infty \) while \( n \) is fixed (Yata & Aoshima 2010b, 2012). The geometric representations are keys to deal with the HDLSS data by “high-dimensional statistical analysis. We consider the following sphericity condition

\[
\frac{\text{tr} \left( \tilde{\Sigma}^2 \right)^2}{\left( \text{tr} \tilde{\Sigma} \right)^2} = \frac{\sum_{i=1}^{d} \lambda_i^2}{\left( \sum_{i=1}^{d} \lambda_i \right)^2} \rightarrow 0 \quad \text{as} \quad d \rightarrow \infty.
\]
Figure 20. A sample distribution drawn from a two typical probability density distributions. Left: a sample generated from a $d$-dimensional Gaussian distribution with a mean $\vec{\mu} = \vec{0}$ and a covariance matrix $\tilde{\Sigma}_d = \tilde{I}_d$ (a $d$-dimensional identity matrix). Right: a sample generated from a $d$-dimensional $t$-distribution with $t_d(0, I_d, \nu)$, with a mean $\vec{\mu} = \vec{0}$, a covariance matrix $I_d$ and a degrees of freedom $\nu = 5$.

When $\tilde{X}$ is Gaussian, Ahn et al. (2007), Jung & Marron (2009) and YA12 showed a geometric representation as

$$n \left( \sum_{i=1}^{d} \lambda_i \right)^{-1} \tilde{S}_D \xrightarrow{P} \tilde{I}_n \text{ as } d \to \infty .$$

(B2)

Here the $\lambda_i$s are eigenvalues of the covariance matrix $\tilde{\Sigma}$, i.e., not the sample covariance $\tilde{S}$. Let

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

(B3)

and $S^{n-1} \equiv \{ \vec{x} \in \mathbb{R}^n \mid \| \vec{x} \| = 1 \}$: a unit $(n-1)$-sphere. YA12 showed that

$$\vec{w}_i \in S^{n-1} \text{ in probability as } d \to \infty ,$$

(B4)

that is, the surface concentration of the unit $n$-sphere. From the geometric representation by eq. (B4), we observe that the sample eigenvalues become deterministic and converge to the same value. This means that it becomes impossible to estimate the eigenvalues and the eigenvectors by using $\tilde{S}_D$ (or $\tilde{S}$) as in the conventional method.

\(^{17}\) YA12 denoted $S^{n-1}$ as $\mathbb{R}^n$. We used a more familiar notation in geometry in this work.
When $\tilde{X}$ is non-Gaussian, Yata & Aoshima (2010b) and YA12 discovered a different type of geometric representation,

$$n\left(\sum_{i=1}^{d} \lambda_i\right)^{-1} \tilde{S}_D - \tilde{D}_n \xrightarrow{P} \tilde{O}_n \quad \text{as} \quad d \to \infty.$$  \hfill (B5)

where $\tilde{D}_n$ is a diagonal matrix whose diagonal elements are of $O_P(1)$, and $\tilde{O}_n$ is a zero matrix of dimension $n \times n$. Let

$$\mathcal{A}^n \equiv \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \cup \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \cup \cdots \cup \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\},$$  \hfill (B6)

($\mathcal{A}^n \subset \mathbb{R}^n$). When $\tilde{X}$ is non-Gaussian, YA12 showed that

$$\tilde{u}_i \in \mathcal{A}^n \quad \text{in probability as} \quad d \to \infty,$$  \hfill (B7)

that is, the axis concentration of the $n$-dimensional dual space. From the geometric representation by eq. (B7), we can see that the eigenvectors of the data matrix converge to the unit vector of one of the axes of $\mathbb{R}^n$ and the eigenvalues would be indefinite.

We consider $(z_{1k}^2 - 1, \ldots, z_{dk}^2 - 1)$ $(k = 1, \ldots, n)$ and its covariance matrix $\tilde{\Phi} = (\phi_{ij})$. YA12 considered a boundary that separates the two cases described by eq. (B4) and eq. (B7) as follows.

$$\sum_{i,j=1}^{d} \lambda_i \lambda_j \phi_{ij} \left(\sum_{i=1}^{d} \lambda_i\right)^{-2} \to 0 \quad \text{as} \quad d \to \infty.$$  \hfill (B8)

Note that eq. (B8) is equivalent to eq. (B1) under Condition (16). Based on this, YA12 proposed the following important theorem.

**Theorem B.1** [Yata & Aoshima (2012) I]

Assume the sphericity condition [eq. (B1)]. If the elements of $\tilde{Z}$ satisfy eq. (B8), we have eqs. (B2) and (B4) as $d \to \infty$. Otherwise, eq. (B5) holds as $d \to \infty$, and under some regularity conditions, we have eq. (B7) as $d \to \infty$.

Now we are ready to examine the geometric behavior of the noise in the dual space. Consider a sample with a $d$-dimensional Gaussian distribution with a mean $\mu = \bar{0}$ and covariance matrix $\Sigma_d = \bar{I}_d$. If we generate a sample $n = 3$ from this distribution, we obtain a sample distribution displayed in Fig. 20. For a low-$d$, we have a sample distribution not very far from our intuition, i.e., the sample data concentrate on the origin and have a certain dispersion around it. However, for a higher-$d$, the sample data concentrate on a sphere with a radius of $(d/n)^{\frac{1}{2}}$ (Left panel of Fig. 20). This clearly shows the situation of high-dimensional
data, namely the intrinsic feature of the data is completely overwhelmed by a tremendous disturbance from the huge noise sphere.

The behavior of the sample is even more counter-intuitive for a non-Gaussian case. If we generate a sample from a $d$-dimensional $t$-distribution $t_d(0, I_d, \nu)$, with a mean $\bar{\mu} = 0$, a covariance matrix $I_d$ and a degrees of freedom $\nu = 5$, we observe a very particular behavior (Right panel of Fig. 20). All the data concentrate on one of the axes for high-$d$ cases. This is referred to as the axis concentration, typically seen for a non-Gaussian density distribution in high dimensions.

Thus, we understood the mathematical reason why we have the counter-intuitive behaviors of the HDLSS data in Fig. 20. This also gives the important basis of the central method we apply in this work.

C. CONSISTENCY FOR A NON GAUSSIAN CASE

In this section, we give the consistency eq. (17) when $Z$ does not satisfy Condition. (16).

**Theorem C.1 [Yata & Aoshima (2009) II]**

If $Z$ does not satisfy Condition (16), eq. (17) holds for $i \leq m$ under the conditions

1. $d \to \infty$ and $n \to \infty$ for $i$ that satisfies $\alpha_i > 1$.

2. $d \to \infty$ and $d^2 - 2\alpha_i / n(d) \to 0$ for $i$ that satisfies $\alpha_i \in (0, 1]$.

Theorems 2.1 and C.1 give the condition that the estimation by the traditional PCA would have the consistency. In (2), $n$ is heavily depending on $d$ when the components of $\tilde{Z}$ do not satisfy Condition (16). For the non-Gaussian case, Yata & Aoshima (2010a) proposed a new robust PCA called the cross-data-matrix methodology.

D. COMPARISON BETWEEN TRADITIONAL AND HIGH-DIMENSIONAL PCA

As we mentioned in the main text, high-dimensional data as the ALMA data are regarded as a sum of informative signals and noise. Particularly for the case of HDLSS data, the traditional statistical estimation is seriously hampered by nuisance from the noise. In order to demonstrate the contribution of the noise sphere to the statistical estimation of an HDLSS data, we present ratios of PCA eigenvalues estimated with the traditional and high-dimensional PCA. For the latter, we show the estimates obtained by the NRPCA. As we mentioned in the main text, the eigenvalues are significantly overestimated by the traditional PCA when the data are HDLSS-type [eq. (20)]. This effect can be visualized by taking a ratio of eigenvalues obtained by the traditional PCA ($\hat{\lambda}_i$) and that obtained by the NRPCA ($\tilde{\lambda}_i$).

The result is presented in Fig. 21. Left panel of Fig. 21 is the result of the raw data (without Doppler velocity subtraction), while Right panel shows that of the Doppler-corrected data. In both cases, it is clearly shown that traditionally estimated eigenvalues are systematically larger than NRPCA ones. We should note that the contribution of noise almost monotonically increases for higher-order eigenvalues, both for the raw and Doppler-corrected datasets. This is because the eigenvalues are smaller for higher-order indices with respect to the noise contribution [cf. eq. (20)].

We also note that, since the data dimension $d$ is not extremely higher than data size $n$, the overestimation of the eigenvalues are relatively small. We stress, however, all the values are overestimated due to the additional contribution from noise. Namely, for forthcoming data that have a much stronger HDLSS property as that of ALCHEMI, this bias will become much larger. Thus, this examination shows the potential power of the high-dimensional PCA, and high-dimensional statistical analysis in general.
HIGH DIMENSIONAL STATISTICAL ANALYSIS OF NGC 253 31

Figure 21. Ratio of eigenvalues obtained by the traditional PCA ($\hat{\lambda}_i$) and that obtained by the high-dimensional (NR)PCA ($\tilde{\lambda}_i$). Horizontal axes represent the index of eigenvalues, and vertical axes show the ratio $\hat{\lambda}_i/\tilde{\lambda}_i$ ($i = 1, 2, \ldots$). Left panel of Fig. 21 is the result of the raw data (without Doppler velocity subtraction), while Right panel shows that of the Doppler-corrected data.

E. DOPPLER SHIFT CORRECTION OF A LINE: EXAMPLE

We show an example of the Doppler shift correction, HCN(4–3) line in Fig. 22. Red, green, and blue are the corrected spectra from different spatial position on the map of NGC 253, and we see that the line center agree with its restframe frequency. Since each line has a complicated profile, the line center was determined by the Gaussian fitting. The blue vertical solid line represents the restframe frequency of the HCH(4–3) line, 354.5 GHz.

Figure 22. Example of the Doppler shift-corrected molecular emission line, HCN(4–3). Red, green, and blue are the corrected spectra so that the line center agree with its restframe frequency.

REFERENCES

Ahn, J., Marron, J. S., Muller, K. M., & Chi, Y.-Y. 2007, Biometrika, 94, 760.
http://www.jstor.org/stable/20441411

Ando, R., Nakanishi, K., Kohno, K., et al. 2017, ApJ, 849, 81, doi: 10.3847/1538-4357/aa8fd4
Aoshima, M., Shen, D., Shen, H., et al. 2018, Australian and New Zealand Journal of Statistics, 60, 4, doi: 10.1111/anzs.12212
Aoshima, M., & Yata, K. 2011, Sequential Analysis, 30, 356, doi: 10.1080/07474946.2011.619088
—. 2014, Annals of the Institute of Statistical Mathematics, 983, doi: 10.1007/s10463-013-0435-8
—. 2015, Sequential Analysis, 34, 279, doi: 10.1080/07474946.2015.1063256
—. 2017, Sugaku Expositions, 30, 137, doi: https://doi.org/10.1090/suga/421
—. 2018, Statistica Sinica, 28, 43
—. 2019, Methodology and Computing in Applied Probability, 21, 663, doi: https://doi.org/10.1007/s11009-018-9646-z
Astropy Collaboration, Robitaille, Thomas P., Tollerud, Erik J., et al. 2013, A&A, 558, A33, doi: 10.1051/0004-6361/201322068
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, Astronomical Journal, 156, 123, doi: 10.3847/1538-3881/aabc4f
Baik, J., Ben Arous, G., & Péché, S. 2005, Ann. Probab., 33, 1643, doi: 10.1214/009117905000000233
Baik, J., & Silverstein, J. W. 2006, Journal of Multivariate Analysis, 97, 1382, doi: https://doi.org/10.1016/j.jmva.2005.08.003
Bendo, G. J., Beswick, R. J., D’Cruze, M. J., et al. 2015, MNRAS, 450, L80, doi: 10.1093/mnrasl/slv053
Bolatto, A. D., Warren, S. R., Leroy, A. K., et al. 2013, Nature, 499, 450, doi: 10.1038/nature12351
Casella, G., & Hwang, J. T. 1982, Canadian Journal of Statistics, 10, 305, doi: https://doi.org/10.2307/3556196
Fernández-Ontiveros, J. A., Prieto, M. A., & Acosta-Pulido, J. A. 2009, MNRAS, 392, L16, doi: 10.1111/j.1745-3933.2008.00575.x
Galaz, G., & de Lapparent, V. 1998, A&A, 332, 459, https://archive.org/abs/astro-ph/9711093
Golub, T. R., Slonim, D. K., Tamayo, P., et al. 1999, Science, 286, 531, doi: 10.1126/science.286.5439.531
Hall, P., Marron, J. S., & Neeman, A. 2005, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67, 427, doi: https://doi.org/10.1111/j.1467-9868.2005.00510.x
Hand, D. J., Daly, F., McConway, K., Lunn, D., & Ostrowski, E. 1994, A Handbook of Small Data Sets, 1st edition (Chapman and Hall, London)
Ishii, A., Yata, K., & Aoshima, M. 2016, Journal of Statistical Planning and Inference, 170, 186, doi: https://doi.org/10.1016/j.jspi.2015.10.007
Johnstone, I. M. 2001, Ann. Statist., 29, 295, doi: 10.1214/aos/1009210544
Jung, S., & Marron, J. S. 2009, Ann. Statist., 37, 4104, doi: 10.1214/09-AOS709
Kamphuis, P., Józsa, G. I. G., Oh, S . H., et al. 2015, MNRAS, 452, 3139, doi: 10.1093/mnras/stv1480
Keto, E., Hora, J. L., Fazio, G. G., Hoffmann, W., & Deutsch, L. 1999, ApJ, 518, 183, doi: 10.1086/307246
Krieger, N., Bolatto, A. D., Walter, F., et al. 2019, ApJ, 881, 43, doi: 10.3847/1538-4357/ab2d9c
Martín, S., Mangum, J. G., Harada, N., et al. 2021, A&A, 656, A46, doi: 10.1051/0004-6361/202141567
Matsubayashi, K., Sugai, H., Hattori, T., et al. 2009, ApJ, 701, 1636, doi: 10.1088/0004-637X/701/2/1636
McMullin, J. P., Waters, B., Schiebel, D., Young, W., & Golap, K. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 376, Astronomical Data Analysis Software and Systems XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell, 127
Pace, Z. J., Tremonti, C., Chen, Y., et al. 2019, ApJ, 883, 82, doi: 10.3847/1538-4357/ab3723
Paul, D. 2007, Statistica Sinica, 17, 1617, http://www.jstor.org/stable/24307692
Portillo, S. K. N., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, AJ, 160, 45, doi: 10.3847/1538-3881/ab9644
Rekola, R., Richer, M. G., McCall, M. L., et al. 2005, MNRAS, 361, 330, doi: 10.1111/j.1365-2966.2005.09166.x
Rieke, G. H., Lebofsky, M. J., Thompson, R. I., Low, F. J., & Tokunaga, A. T. 1980, ApJ, 238, 24, doi: 10.1086/157954
Ronen, S., Aragon-Salamanca, A., & Lahav, O. 1999, MNRAS, 303, 284, doi: 10.1046/j.1365-8711.1999.02222.x
Sakamoto, K., Mao, R.-Q., Matsushita, S., et al. 2011, ApJ, 735, 19, doi: 10.1088/0004-637X/735/1/19
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29
Walter, F., Bolatto, A. D., Leroy, A. K., et al. 2017, ApJ, 835, 265, doi: 10.3847/1538-4357/835/2/265
Wang, L., Farrah, D., Connolly, B., et al. 2011, MNRAS, 411, 1809, doi: 10.1111/j.1365-2966.2010.17811.x
Yata, K., & Aoshima, M. 2009, Communications in Statistics - Theory and Methods, 38, 2634, doi: 10.1080/03610910902936083
—. 2010a, Journal of multivariate analysis, 101, 2060. https://ci.nii.ac.jp/naid/120002381162/
—. 2010b, Comm. Statist. Theory Methods. Special Issue Honoring M. Akahira, (ed. M. Aoshima), 39, 1511
—. 2012, Journal of Multivariate Analysis, 105, 193, doi: https://doi.org/10.1016/j.jmva.2011.09.002
—. 2013, Journal of Multivariate Analysis, 122, 334, doi: https://doi.org/10.1016/j.jmva.2013.08.003
—. 2022, To appear in Statistica Sinica (arXiv 2209.14891). https://arxiv.org/abs/2209.14891