Journal of Physics: Conference Series **699** (2016) 012009 doi:10.1088/1742-6596/699/1/012009

Data-driven approach to Type Ia supernovae: variable selection on the peak luminosity and clustering in visual analytics

Makoto Uemura[1], Koji S Kawabata[1], Shiro Ikeda[2], Keiichi Maeda[3], Hsiang-Yun Wu[4], Kazuho Watanabe[5], Shigeo Takahashi[6] and Issei Fujishiro[4]

[1] Hiroshima Astrophysical Science Center, Hiroshima University, Kagamiyama 1-3-1, Higashi-Hiroshima 739-8526, Japan
[2] The Institute of Statistical Mathematics and CREST, JST, Tachikawa 190-8562, Japan
[3] Department of Astronomy, Kyoto University, Kitashirakawa-Oiwake-cho Sakyo-ku, Kyoto 606-8502, Japan
[4] Department of Information and Computer Science, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
[5] Toyohashi University of Technology, 1-1 Hibarigaoka Tempaku-cho Toyohashi 441-8580, Japan
[6] Department of Computer Science and Engineering, University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu 965-8580, Japan

E-mail: uemuram@hiroshima-u.ac.jp

Abstract. Type Ia supernovae (SNIa) have an almost uniform peak luminosity, so that they are used as "standard candle" to estimate distances to galaxies in cosmology. In this article, we introduce our two recent works on SNIa based on data-driven approach. The diversity in the peak luminosity of SNIa can be reduced by corrections in several variables. The color and decay rate have been used as the explanatory variables of the peak luminosity in past studies. However, it is proposed that their spectral data could give a better model of the peak luminosity. We use cross-validation in order to control the generalization error and a LASSO-type estimator in order to choose the set of variables. Using 78 samples and 276 candidates of variables, we confirm that the peak luminosity depends on the color and decay rate. Our analysis does not support adding any other variables in order to have a better generalization error. On the other hand, this analysis is based on the assumption that SNIa originate in a single population, while it is not trivial. Indeed, several sub-types possibly having different nature have been proposed. We used a visual analytics tool for the asymmetric biclustering method to find both a good set of variables and samples at the same time. Using 14 variables and 132 samples, we found that SNIa can be divided into two categories by the expansion velocity of ejecta. Those examples demonstrate that the data-driven approach is useful for high-dimensional large-volume data which becomes common in modern astronomy.

1. Introduction

Type Ia supernovae (SNIa) are thermonuclear explosions that occur when an accreting white dwarf reaches its critical mass, and starts a runaway reaction. The critical mass, so-called the "Chandrasekhar mass", is about 1.38 times the mass of the sun ($M_\odot$). The presence of this

critical mass makes the luminosity of SNIa uniform. Then, the apparent magnitude of SNIa measured on the earth is a function of the distance to the object. Hence, SNIa have been used as a "standard candle" which enables us to measure the distance of galaxies in cosmology. The accelerating expansion of the Universe was discovered based on the distance estimation using SNIa [1, 2, 3]. In recent years, the amount of the SNIa data has rapidly been increasing owing to large survey projects [4, 5]. When the data set was small, it can be handled based on the experience of domain experts. However, the numbers of both samples and variables are large nowadays, so that the data-driven approach is expected to be useful. Here, we introduce two issues of SNIa, that is, the variable selection of the peak magnitude and the classification of SNIa.

The observed peak magnitude of SNIa (M) has a small, but significant diversity. It is due to the interstellar extinction, and hence calibrated by their reddened color. In addition, it is well known that M depends on the decay rate [6]. The corrected magnitude, M_0, is expressed as:

$$M = M_0 + \beta_1 c + \beta_2 x, \tag{1}$$

where c and x denote the color and decay rate, and $\beta_{1,2}$ are their coefficients. Both c and x are values obtained by photometric observations with broad-band filters. The photometric data is more readily obtained compared with spectroscopic ones because the fluxes are integrated over the transmission curve of the broad-band filter in photometric observations. Recent survey projects provide large and uniform spectroscopic data. Using those spectra, a new set of explanatory variables of M has been searched [7, 8, 9]. However, the number of the candidates of explanatory variables, that is, all data points in spectra, is larger than the number of sample in those studies. We need to select the best set of variables for a given data set of M, although the standard least-square method cannot solve this problem.

The other example is about the classification of SNIa. Several sub-types have been proposed for SNIa based on observed features. The classification of SNIa is crucial for cosmology, because some sub-types possibly have different peak magnitudes or colors. For example, Wang, et al. 2009 classified SNIa using the expansion velocity which is measured from the absorption line of Si II 6355Å [10]. They proposed that SNIa having high expansion velocities form a distinct sub-type which has an intrinsically red color, and appropriate color corrections improve the distance estimation. Now, we can revisit past classifications using large data of samples and variables. A problem is that it is difficult to see the structure in the data in the case that the dimension of the data is high. We need to find both a good set of variables and sub-types at the same time.

In this article, we review our recent works related to the above two problems. Uemura, et al. 2015 use a variable selection approach for modeling the peak magnitude of SNIa [11]. We use cross-validation to control the generalization error and a LASSO-type estimator to choose the set of variables (section 3). In section 4, we present a visual analytics approach for clustering SNIa. This visual analytics tool is based on asymmetric biclustering method, and enable us to visually check the result of clustering to find the potential structures in the data.

2. Data

All data of SNIa was obtained from the supernova database operated by the UC Berkeley group[1] [5]. They present both photometric (the peak magnitude, color, decay rate, and distance) and spectroscopic data. We show an example of spectra in figure 1. The spectra of SNIa are characterized by broad absorption-lines, for example, of Si, Ca, and S. The right panel shows the spectra around Si II 5982 and 6355Å. We also show the definitions of some astronomical values related to the lines.

[1] http://heracles.astro.berkeley.edu/sndb/

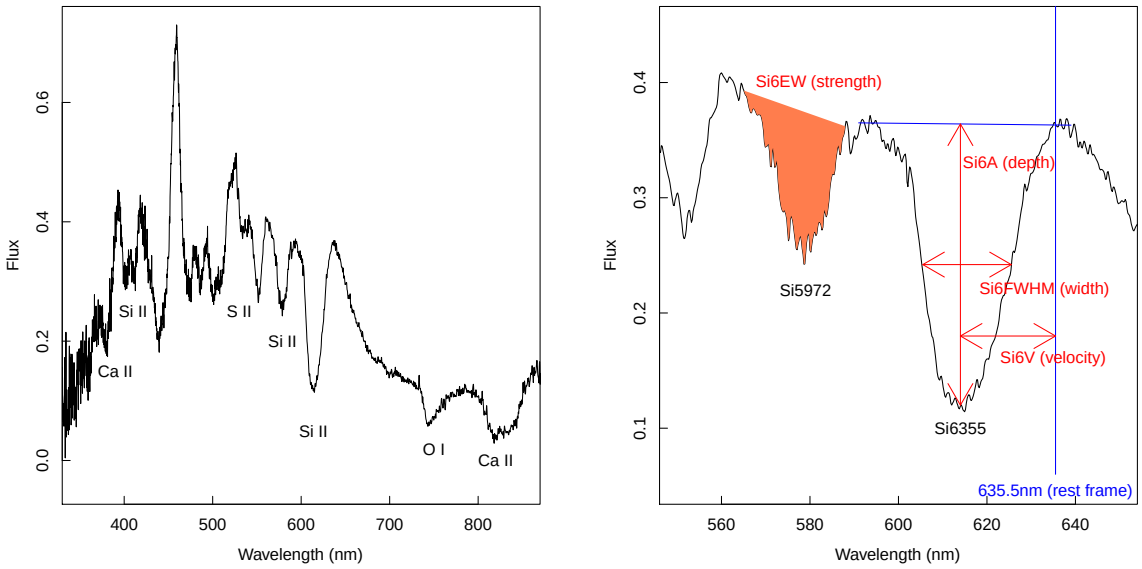

Figure 1. Left panel: Typical spectra of Type Ia supernovae (SNIa) between 350 and 850 nm. Right panel: Close-up view of the spectra around two silicon lines, Si II 5982 and 6355Å. Some definitions of astronomical values are also indicated. The data is from the database of the Berkeley group [5].

In section 3, we use the spectra, as well as the photometric data as the candidates of explanatory variables of the peak magnitude. Among the database, we selected samples having spectral data between 3500 and 8500Åtaken within five days from the maximum. This selection reduced the number of samples to 78. Each spectrum contains 138 data points. In our sample, we excluded Type Iax supernovae, which have recently been recognized as a peculiar faint class of SNIa [12]. In section 4, we use the absorption line parameters which are measured from the spectral data. In conjunction with the photometric data, such as the distance, decay rate, color, and peak magnitude, we use them as the candidates of variables for classification of SNIa.

3. Variable selection for modeling the peak magnitude of SNIa

In this section, we review our recent work about the variable selection of the peak absolute magnitude (M) of SNIa [11]. We consider a linear model of M:

$$\boldsymbol{M} = X\boldsymbol{\beta} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{M}$ is a vector containing N samples of the peak absolute magnitude, M. X is a $N \times L$ matrix of the explanatory variables, including a constant term, and $\boldsymbol{\beta}$ is co-efficients. We consider a Gaussian noise, $\boldsymbol{e}$. The number of samples, N, is larger than that of the candidate variables, L if the spectral data are included as the explanatory variables [7, 8, 9]. However, our interest focuses on a solution having a few variables which control the peak magnitude of SNIa. In other words, most of elements of $\boldsymbol{\beta}$ should be zero. In order to obtain such a solution, we use a LASSO-type estimator:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{M} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{3}$$

where λ is a tunable constant. Using LASSO, we can obtain a sparse solution, $\hat{\boldsymbol{\beta}}$, and choose the best set of explanatory variables. We use cross-validation in order to control the generalization error. A 10-fold cross-validation was applied to determine λ in the following results.

Table 1. Coefficients of variables [11].

	Coefficients		
Variables	Model 1	Model 2	Model 3
c	0.376	—	—
$f_{\rm tot}(6373)$	0.100	0	0
x	−0.050	−0.014	—
$f_{\rm cnt}(6084)$	−0.034	0	0
$f_{\rm cnt}(6289)$	−0.045	0	0
$f_{\rm cnt}(6631)$	−0.061	0	0
$R(3780/4580)$	−0.050	0	0
$f_{\rm tot}(3752)$	0.063	0	0

The candidates of the explanatory variables are the color (c), decay rate (x), six flux ratios that were previously proposed ($R(\lambda_1/\lambda_2)$), and spectral data. About the spectral data, we used two kinds of normalized spectra instead of the arbitrary flux ratios that were used in the previous studies. First, the variables of most interest are the flux ratios of the line areas to the continuum level. It can be obtained by the spectra normalized by the continuum level ($\boldsymbol{f_{\rm cnt}}(\lambda)$). Second, the local colors in the continuum which may have information independent of the broadband colors (c) are also variables of interest. They can be obtained by the spectra normalized by the total flux between 3500 and 8500Å($\boldsymbol{f_{\rm tot}}(\lambda)$). Each spectra was re-binned into 134 wavelength bins between 3500 and 8500Å, as in the past studies [9]. The number of all candidate variables is $L = 276$. The number of sample is $N = 78$. Each variable was normalized to have zero mean and unit variance by a linear scaling.

We start the model with all candidate variables (Model 1). Table 1 shows the selected variables having non-zero coefficients. As proposed in many past studies, the color, c, and decay rate, x have non-zero coefficients. $f_{\rm tot}(6373)$ denotes the flux at 6373Å of the total flux normalized spectra, having a relatively large coefficient. This variable represent a local continuum color at 6373Å, and could be interesting if it has information independent of the broad-band color, c. The other non-zero elements are related to the absorption lines: $f_{\rm cnt}(6084)$ and $f_{\rm cnt}(6289)$,the fluxes at 6084 and 6289Å of the continuum normalized spectra, are related to the absorption line, Si II 6355Å. $R(3780/4580)$, a flux ratio between 3780 and 4580Å, and $f_{\rm tot}(3752)$ are related to Ca II. $f_{\rm cnt}(6631)$ can be considered as a noise because it is a continuum flux of continuum normalized spectra.

The local color, $f_{\rm tot}(6373)$, is potentially interesting, while its non-zero coefficients could be due to a high correlation between c and $f_{\rm tot}(6373)$. To evaluate it, we used a new model in which M is corrected for the effect of c (Model 2). As a result, no variable was selected, except for the decay rate, x. The result was confirmed another model in which M is corrected for the effect of c and x (Model 3). No variable has non-zero coefficient in this model. This result suggests that $f_{\rm tot}(6373)$ and the other spectroscopic variables were selected in the first model simply because of their high correlations with c or x. Our analysis confirmed the past understanding about SNIa, that is, the peak absolute magnitude depends on the color and decay rate, and does not support adding any other variables in order to have a better generalization error.

In recent studies, arbitrary flux ratios were taken into account as the candidates of the explanatory variables of the peak magnitude [7, 8, 9]. Each spectra contains 138 data points, and the number of the flux combination is $138 \times 137 = 18906$, which is over two orders of magnitude larger than the number of samples. The results of those studies are not completely consistent. Our approach has two advantages against those past studies. First, we reduced the

Table 2. Variables used in the ABC analysis

Variable	Definition
Si6EW	Equivalent width of Si6355.
Si6DEW	Corrected Si6EW for the observation epoch.
Si6V	Velocity of Si6355.
Si6A	Depth of Si6355.
Si6FWHM	Full-width of half maximum of Si6355.
Si5EW	Equivalent width of Si5972.
Si5DEW	Corrected Si5EW for the observation epoch.
z	Redshift, corresponding to the distance to the object.
x	Decay rate.
c	Color.
mB	Apparent magnitude.
MB	Absolute magnitude calculated from z and mB.
MBc	Color (c) corrected absolute magnitude.
MBcx	Color (c) and decay-rate (x) corrected absolute magnitude.

number of candidate variables by using normalized spectra, instead of the arbitrary flux ratios. The reduction of the candidate variables leads to the reduction of false positive signals. Second, this data-driven approach enable us to determine the number of explanatory variables from the data itself. In past studies, only one or two variables were evaluated. In our study, we conclude that no spectroscopic variable can provide a better model for M. Of course, our conclusion is based on our samples. A better model could be found by using further larger data sets in future.

4. Classification of SNIa via visual analytics for asymmetric biclustering

In the last section, we developed a model of the peak magnitude of SNIa. This modeling was based on the assumption that our samples of SNIa have the common peak magnitude, while it is not trivial as introduced in section 1. A search for possible sub-types of SNIa and the axes for their classification is crucial not only for the cosmology, but also for the understanding their explosion mechanism. In this section, we present our analysis of SNIa using a visual analytics tool for this classification problem.

We use the Asymmetric Bi-Clustering (abbreviated as ABC hereafter) tool for our analysis [13]. This tool has recently been developed for an asymmetric biclustering-based subspace search of multivariate data. Highly correlated dimensions are automatically grouped to form feature subspaces in an interactive and progressive manner. The asymmetric biclustering combines spherical k-means for grouping highly correlated dimensions, together with ordinary k-means for identifying subsets of data samples. Lower dimensional representations of data in feature subspaces are successfully visualized by parallel coordinate plot (PCP).

Using the ABC tool, we analyzed 132 SNIa samples from the Berkeley database and 14 variables, which are summarized in table 2. The definitions of the equivalent width (EW), velocity (V), depth (A), and full-width of half maximum (FWHM) of lines are depicted in figure 1. We show an example set of the analysis process step by step. Figure 2 shows the initial state. The 132 samples are plotted in 14 parallel coordinates. Figure 3 shows the results of biclustering for 6 data and 6 axis clusters. In the upper panel, axes having high correlation are clustered, and the sample clusters are expressed by colors. The block matrix diagram shown in the lower panel indicates block errors of each axis and sample cluster. We can see that a small, blue-colored cluster has a large block error in the axis cluster of x and MBcx, as shown in

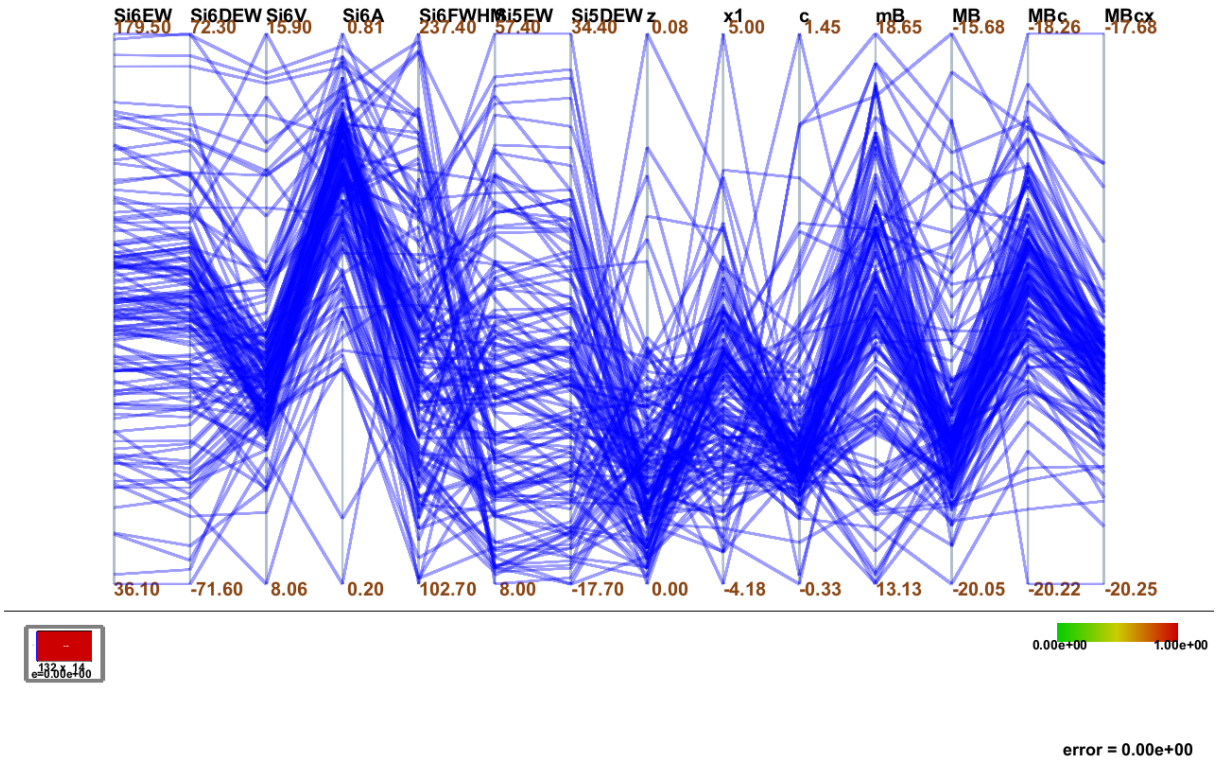

Figure 2. Initial state of the target SNIa dataset of 132 samples in 14 variables.

the lower panel. These are outliers which appears due to their peculiar variation pattern. We excluded them in the ABC tool, and then obtained a new result shown in figure 4. We found the axis cluster {MBc, MBcx} has large block errors in several sample clusters. Both of these variables are the absolute magnitudes, that is, the intrinsic luminosity of SNIa. It is well known that MBc correlates with x [6]. MBcx is the corrected value of MBc for this effect. Clustering those two variables can be understood because, in our data, the trend in MBc against x is small compared with the dispersion in MBc. Judging from the large block errors, we eliminated this axis cluster.

After further eliminating axes or axis clusters which have large block errors, we obtain three axis and three sample clusters, as shown in figure 5. We consider this state as the final result because the total error, which is indicated at the right-bottom area in figures 3–5, reaches its minimum at this state. We note that the order of the process can change run by run because it uses random values to set the initial cluster number for each sample. We have confirmed that most trials converge to the result shown in figure 5. In figure 5, we show the results in contracted PCP with both polylines (upper panels) and strip rendering (lower panels). We can see three axis and three sample clusters. The axis clusters are about the expansion velocity of Si II 6355Å (the left axis), the strength of Si II 6355 (middle), and 5972Å (right). We found that the objects having high velocity, forming a cyan-colored cluster, have strong Si II 6355 and weak Si II 5972, while the low velocity clusters, indicated by blue and magenta, exhibits weak, but positive correlation between Si II 6355 and 5972Å.

The two axes, Si II 6355 and 5972Å have been used for the classification of SNIa [14]. This classification scheme was proposed by a clustering analysis of spectra, and divided SNIa into four sub-types: “Normal”, “Broad line”, “Shallow silicon”, and “Cool” types. The expansion velocity of Si II 6355Å has been used for the definition of the high velocity group of SNIa, which

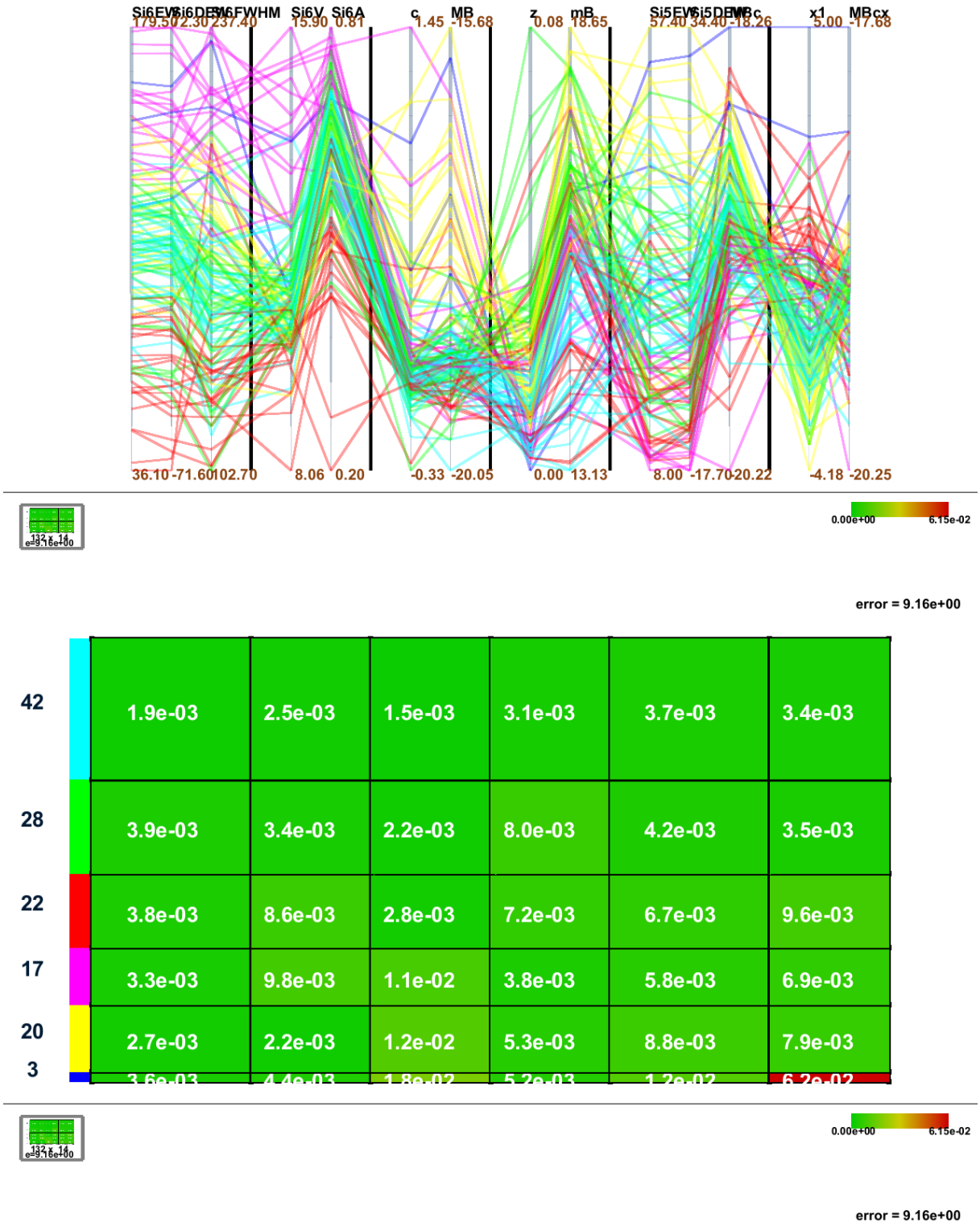

Figure 3. Biclustering 14 axis and 132 samples to 6 axis and 6 sample clusters. Upper panel: Clustered PCP. Lower panel: Block matrix diagram.

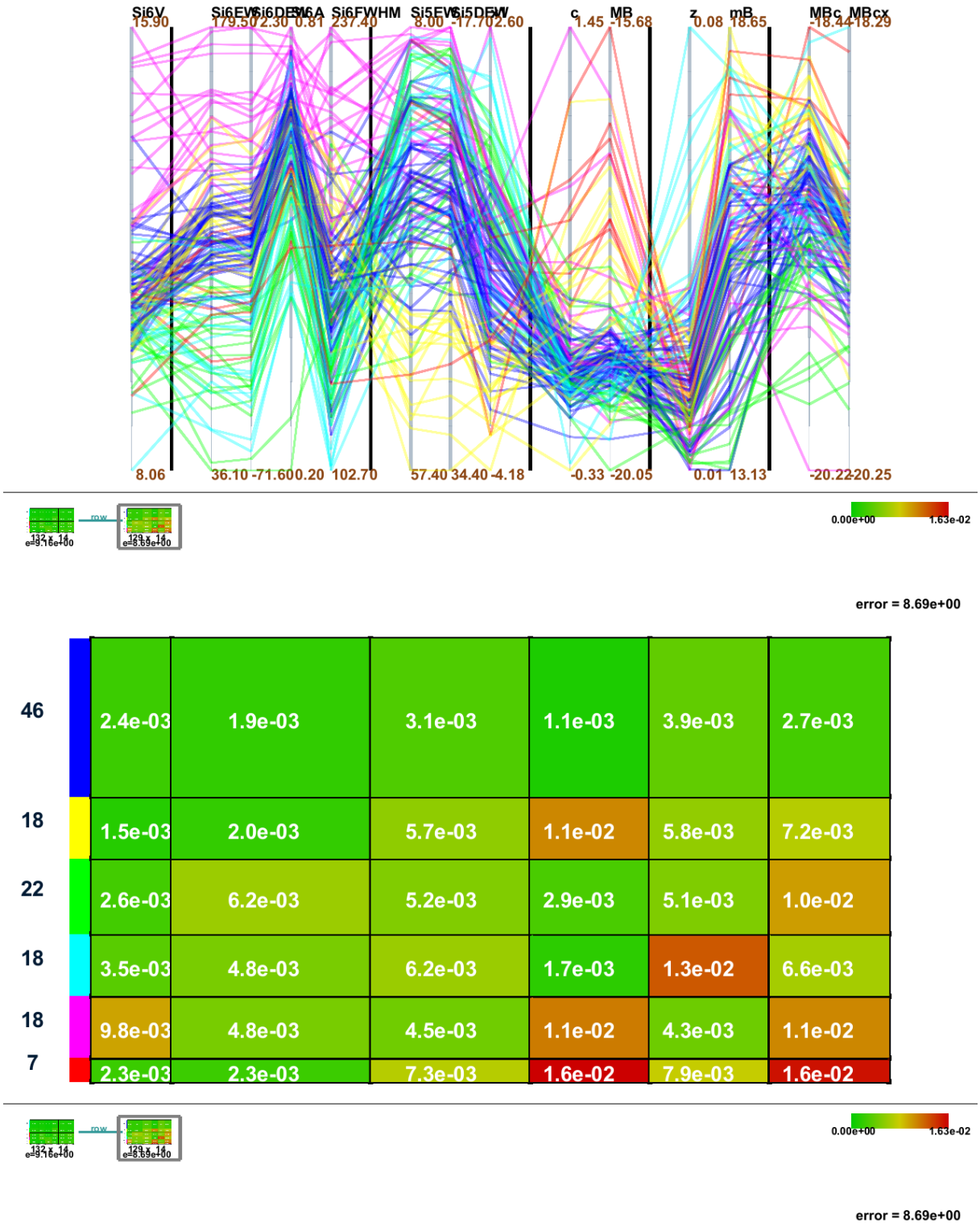

Figure 4. Biclustering 14 axis and 129 samples to 6 axis and 6 sample clusters. Upper panel: Clustered PCP. Lower panel: Block matrix diagram.

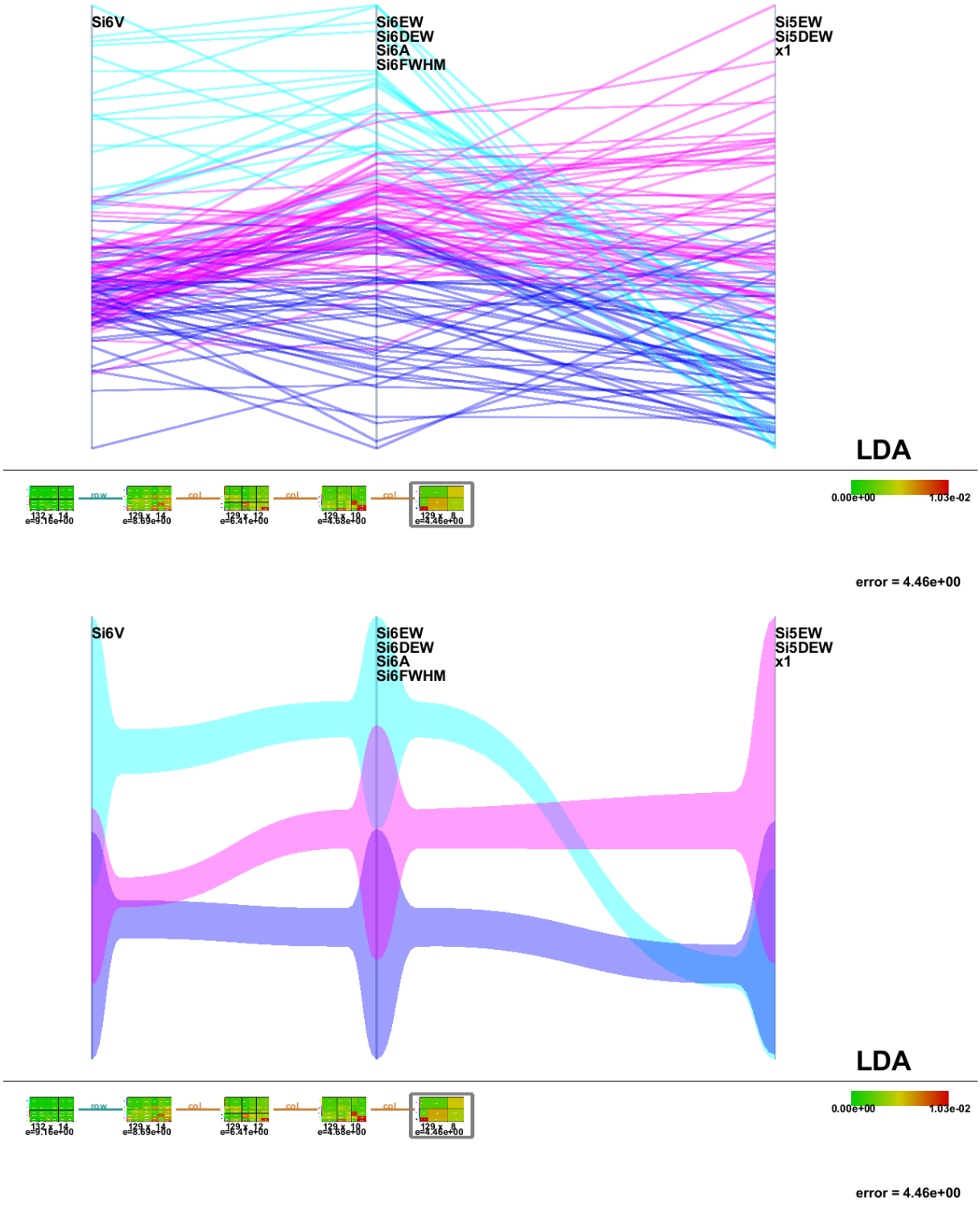

Figure 5. Biclustering 8 axis and 129 samples to 3 axis and 3 sample clusters. Upper panel: Contracted PCP depicted in polylines. Lower panel: Contracted PCP depicted in strip rendering.

is proposed to have a different color behavior [10]. Our analysis also suggests the presence of a high velocity cluster. Furthermore, its behavior of the line strengths corresponds to the "Broad line" type in the Si II 6355—5972Å classification scheme. The axis of the peak magnitude was eliminated because of large errors in our analysis process. This indicates that the diversity of the peak magnitude is not significant between each sample classes in our analysis. Thus, our result obtained by data-driven approach is successfully consistent with those previously proposed classifications based on the experience of domain experts.

Our study demonstrates that the asymmetric biclustering tool is useful to find a meaningful classification scheme hidden in high-dimensional data. A potential problem of this tool is the ambiguity of the number of clusters. As mentioned above, we selected it based on the total error, while it may be determined by more proper model selection criteria. The revised version of this tool could overcome this difficulty [15].

5. Summary

In this article, we introduced our study on SNIa using data-driven approach. The LASSO-type estimator enables us to determine the number of variables and select their best set. Our analysis confirmed the classical picture, that is, the model with the color and decay rate (section 3). The ABC tool is useful to find hidden structures in multivariate data. Using this tool, we revisit the classification of SNIa. The result is consistent with those proposed based on the experience of domain experts (section 4). Owing to recent survey projects, we can access large uniform datasets of SNIa. However, old-style analysis might overlook intriguing features in such high-dimensional large dataset. The two examples shown in this article demonstrate that the data-driven approach is useful for modern astronomy.

Acknowledgments

We would like to thank Drs. J. M. Silverman and A. V. Filippenko for providing the Berkeley supernova database. The present study was financially supported by JSPS KAKENHI Grant Number 25120007, 25120008, and 26800100.

References

[1] Schmidt B P, Suntzeff N B, Phillips M M, Schommer R A *et al* 1998 *Astrophysical Journal* **507** (1) 46
[2] Perlmutter S, Turner M S and White M 1999 *Physical Review Letters* **83** (4) 670
[3] Riess A G, Strolger L-G, Tonry J, Casertano S *et al* 2004 *Astrophysical Journal* **607** (2) 665
[4] Law N M, Kulkarni S R, Dekany R G, Ofek E O *et al* 2009 *Publ. of Astron. Soc. of Pacific* **121** (886) 1395
[5] Silverman J M, Foley R J, Filippenko, A V, Ganeshalingam M *et al* 2012 *Monthly Notices of the Royal Astronomical Society* **425** 1789
[6] Phillips M M 1993 *Astrophysical Journal Letters* **413** L105
[7] Bailey S, Aldering G, Antilogus P, Aragon C, Baltay C, Bongard S, Buton C, Childress M *et al* 2009 *Astronomy & Astrophysics* **500** L17
[8] Blondin S, Mandel K S, and Kirshner R P 2011 *Astronomy & Astrophysics* **526** A81
[9] Silverman J M, Ganeshalingam M, Li W and Filippenko A V 2012 *Monthly Notices of the Royal Astronomical Society* **425** 1889
[10] Wang X, Filippenko A V, Ganeshalingam M, Li W *et al* 2009 *Astrophysical Journal Letters* **699** L139
[11] Uemura M, Kawabata K S, Ikeda S and Maeda K 2015 *Publ. of the Astron. Soc. of Japan* **67** 55
[12] Foley R J, Challis P J, Chornock R, Ganeshalingam M, Li W, Marion G H, Morrell N I, Pignata G *et al* 2013 *Astrophysical Journal* **767** (1) 57
[13] Watanabe K, Wu H Y, Niibe Y, Takahashi S and Fujishiro I 2015 *Proceedings of IEEE Pacific Visualization Symposium 2015*
[14] Branch D, Dang L C, Hall N, Ketchum W *et al* 2006 *Publ. of Astron. Soc. of Pacific* **118** (842) 560
[15] Watanabe K, Wu H Y, Takahashi S and Fujishiro I 2016 Asymmetric biclustering with constrained von Mises-Fisher models *J. Phys. Conf. Ser.* (this volume) in press