High-Dimensional Data-Driven Approach to Type Ia Supernovae

Makoto Uemura[1]

K. S. Kawabata[1], S. Ikeda[2], K. Maeda[3], K. Watanabe[4], H.-Y. Wu[5], S. Takahashi[6], I. Fujishiro[5]

1. Hiroshima astrophysical science center, Hiroshima Univ. 2. The institute of Statistical Mathematics, 3. Kyoto Univ., 4. Toyohashi Univ. of Tech., 5. Keio Univ., 6. Aizu Univ.

Variable Selection for the Peak Luminosity

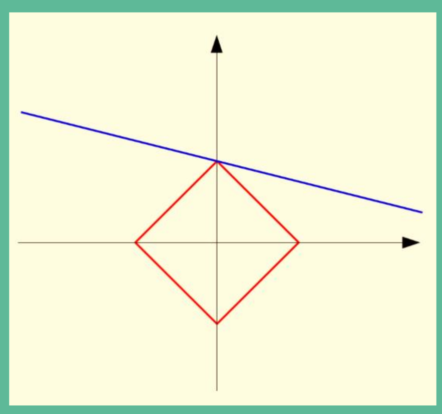
M. Uemura, K.S. Kawabata, S. Ikeda, K. Maeda, PASJ, 2015, 67, 55

We discuss what is an appropriate set of explanatory variables to predict the absolute magnitude at the maximum of Type Ia supernovae. We use cross-validation to control the generalization error and a LASSO-type estimator to choose the set of variables. We studied the Berkeley supernova data base with our approach. As a result, the best set of variables are the color and light-curve width. Our approach does not support adding any spectroscopic variables.

1. LASSO

LASSO (Least Absolute Shrinkage and Selection Operator, [1])
= A kind of sparse regression
→ Selecting an appropriate set of variables by making the coefficient vector sparse.

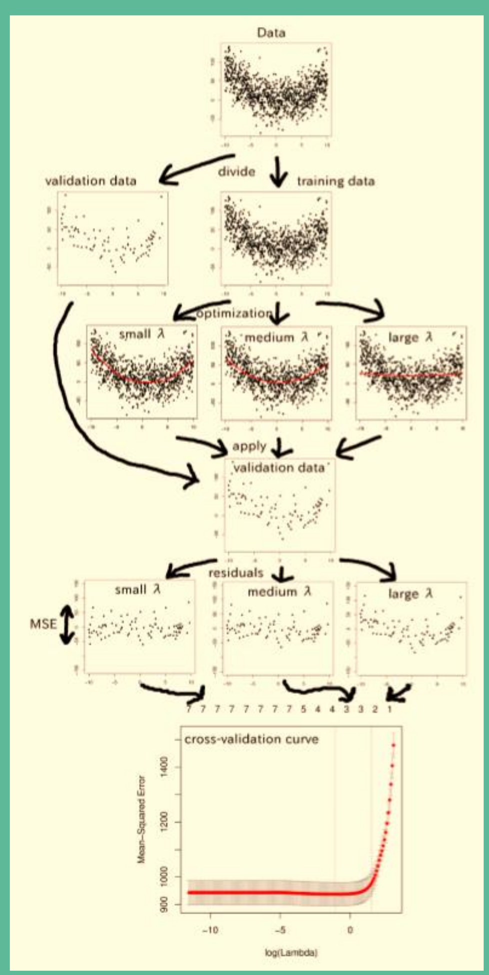
$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \left\{ \| y - X\beta \|_2^2 + \lambda \|\beta\|_1 \right\}$$



2. Cross-validation

= A useful method for model selection
→ Estimating an appropriate value of λ in LASSO

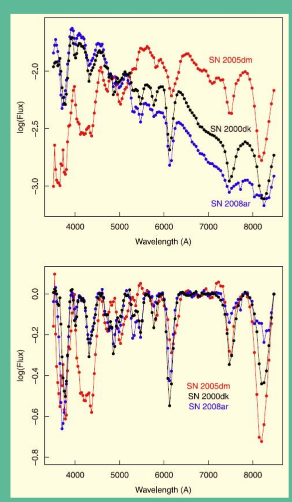
Step 1 : Dividing the data into training and validation data
Step 2 : Optimization of the model to the training data.
Step 3 : Calculating MSE from the model and validation data.



3. Data & Model

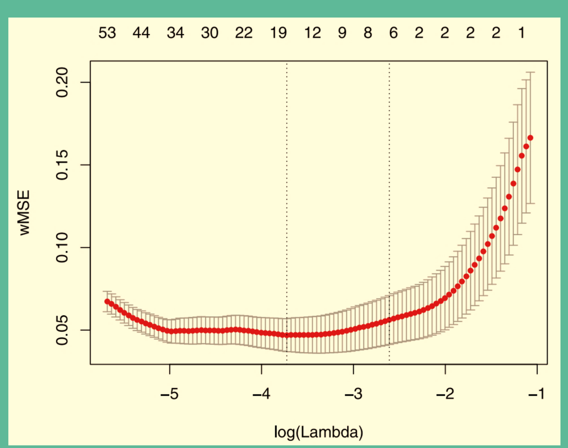
Data:
The peak magnitude of SNeIa from the Berkeley supernova database[2]
(the same as those in [3]) →78 samples

Variables:
color (c), light-curve width (x1),
Total flux normalized spectra (f_{tot}),
Continuum normalized spectra (f_{cnt}),
Previously proposed flux ratios[3,4,5]
→ 276 variables



→ Estimating 276 coefficients from 78 data points.

4. Result



Cross-validation curve for Model 1

Model	Target variable y (N)	Explanatory variables X (L)	Non-zero elements	Coefficients β	Probability p
1	M_B (78)	$x_1, c, f_{tot}, f_{cnt}, \mathcal{R}$ (276)	c	0.376	1.00
			$f_{tot}(6373)$	0.100	1.00
			x_1	−0.050	0.98
			$f_{cnt}(6084)$	−0.034	0.98
			$f_{cnt}(6289)$	−0.045	0.95
			$f_{cnt}(6631)$	−0.061	0.80
			$\mathcal{R}(3780/4580)$	−0.050	0.74
			$f_{tot}(3752)$	0.063	0.73
2	$M_B - \beta_1 c$ (78)	$x_1, f_{tot}, f_{cnt}, \mathcal{R}$ (275)	x_1	−0.020	0.99
3	$M_B - \beta_1 c$ (78)	$x_1, f^c_{tot}, f_{cnt}, \mathcal{R}^c$ (275)	x_1	−0.014	0.85
4a	x_1 (76)	$c, f^c_{tot}, f_{cnt}, \mathcal{R}^c, \mathcal{L}_{Si II 4000}$ (280)	$DpEW_{Si II 4000}$	−0.455	1.00
			$f_{cnt}(5770)$	0.518	1.00
			$f_{cnt}(3982)$	−0.262	1.00
			$f_{cnt}(7038)$	−0.485	0.96
			$f^c_{tot}(4988)$	−0.238	0.77
			$f_{cnt}(6084)$	0.281	0.62
4b	x_1 (74)	$c, f^c_{tot}, f_{cnt}, \mathcal{R}^c, \mathcal{L}_{S II “W”}$ (280)	$f_{cnt}(5770)$	1.034	1.00
			$f_{cnt}(6084)$	0.440	1.00
			$f^c_{tot}(6458)$	0.300	1.00
			$f_{cnt}(3982)$	0.041	1.00
			$f_{cnt}(7179)$	0.289	0.99
			$f_{cnt}(6458)$	−0.236	0.94
			$f_{cnt}(6331)$	0.612	0.92
5	$M_B - (\beta_1 c + \beta_2 x_1)$ (78)	$f^c_{tot}, f_{cnt}, \mathcal{R}^c$ (273)	–	–	–

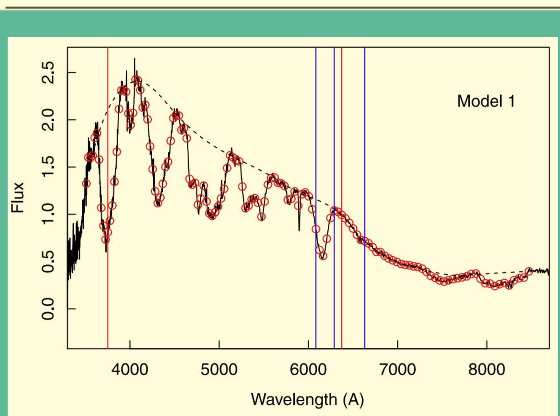
Model 1 (=*MB* and all variables)
→ ftot(6373) = local color, independent of the broadband color, *c* ?

Model 2 (= *c* corrected *MB*)
→ Only the light-curve width, *x1*
→ ftot(6373) in Model 1 is just due to its high correlation with *c*.

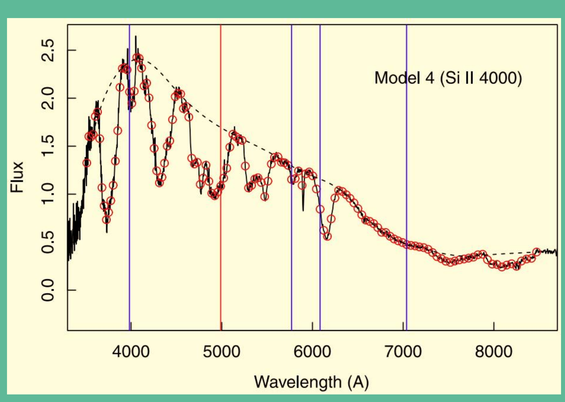
Model 3
(= similar to Model2, but broadband-color corrected variables)
→ Again, only x1

Model 4a, b
(= variable selection for x1 with Si II 4000 (a), or S II "W" (b))
→ Si II 4000 and 5765 is selected. Consistent with past studies.

Model 5 (= *c* and *x1* corrected *MB*)
→ No variable is selected.



Selected variables in Model 1



Selected variables in Model 4a

→ Best model: color (*c*) and light-curve width (*x1*), without any spectroscopic variables.

Classification of SNeIa via Visual Analytics

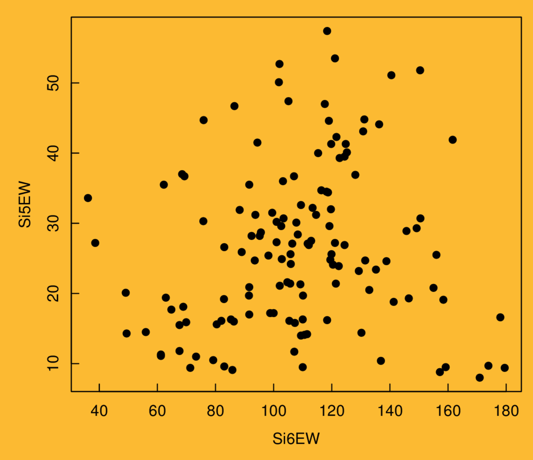
M. Uemura, K. Watanabe, H.-Y. Wu, S. Takahashi, I. Fujishiro, JPhCS, 2016, 699, 2009

The classification scheme of SNeIa is revisited. Recently, the data size has increased both in the numbers of samples and variables, while it is hard to find a hidden structure in the high-dimensional space. We used a visual analytics tool to find both a good set of variables and samples at the same time. Using 14 variables and 132 samples from the Berkeley supernova database, we found that SNeIa can be divided into two categories by the velocity of Si II 6355.

1. Classification of SNeIa

Examples of classification:
- Based on lines (especially Si II 6355 EW v.s. 5972 EW) [6]
  → Normal, Cool, Broad line, shallow silicon groups
- Based on the expansion velocity of Si II 6355 [7]
  → High velocity group

Recently, we have more samples and more variables.
→ Which variables should be used for appropriate classification?

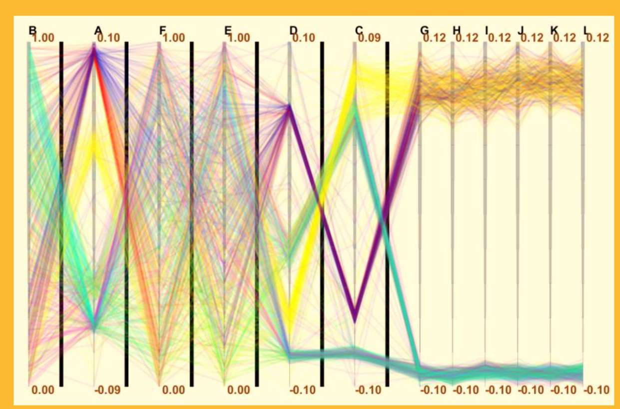


EWs of Si II 6355 and 5972. The data from [2]

2. Visual Analytics for asymmetric biclustering

Asymmetric biclustering
= clustering the samples and variables at the same time.
(using K-means [8], probabilistic method [9])

Visual analytics tool (figure →)
= interactively deleting the clusters of variables and samples, and finding the structure hidden in high-dimensional data.

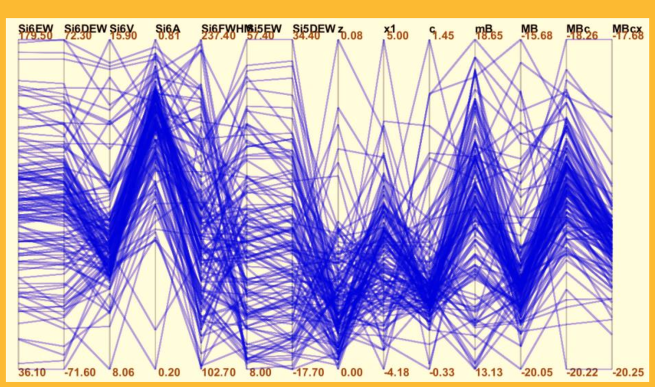


Axes having large variances are not good for classification. *Correlated variables are clustered.*

3. Data

Data: 132 SNeIa samples from the Berkeley supernova database [2]

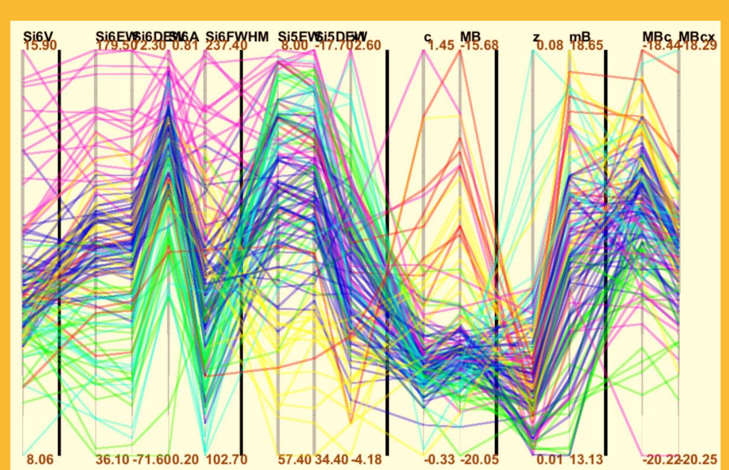
Variables: 14 variables: Si II 6355 EW, DEW, Velocity, Depth, FWHM, Si II 5972 EW, DEW, z, light-curve width (*x1*), broadband color (*c*), apparent magnitude (m_B), absolute magnitude (M_B), color-corrected M_B, color & light-curve corrected M_B

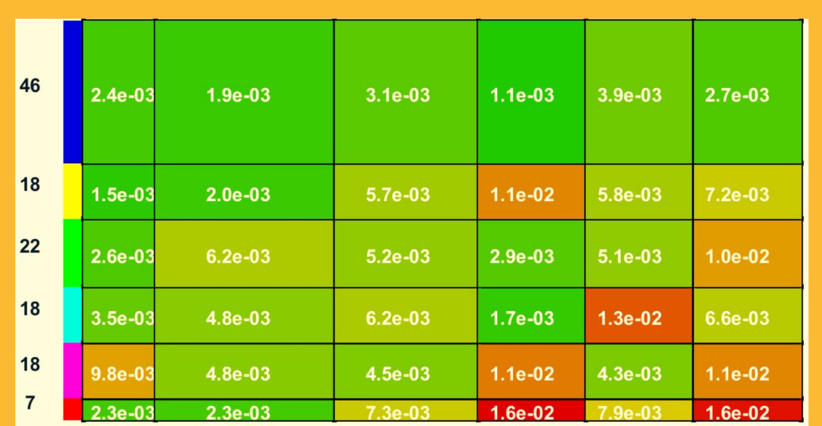


Initial state for 14 variables of 132 SNeIa samples in the parallel coordinate plot.

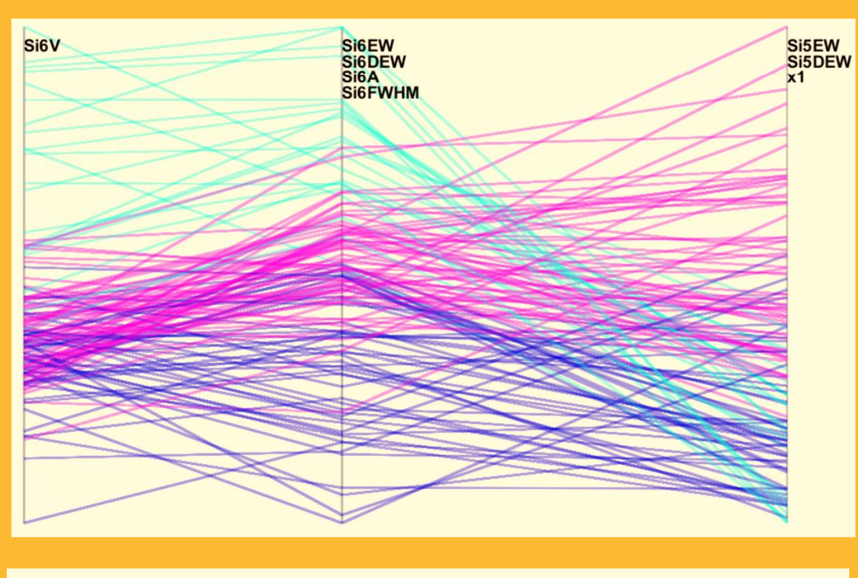
4. Result

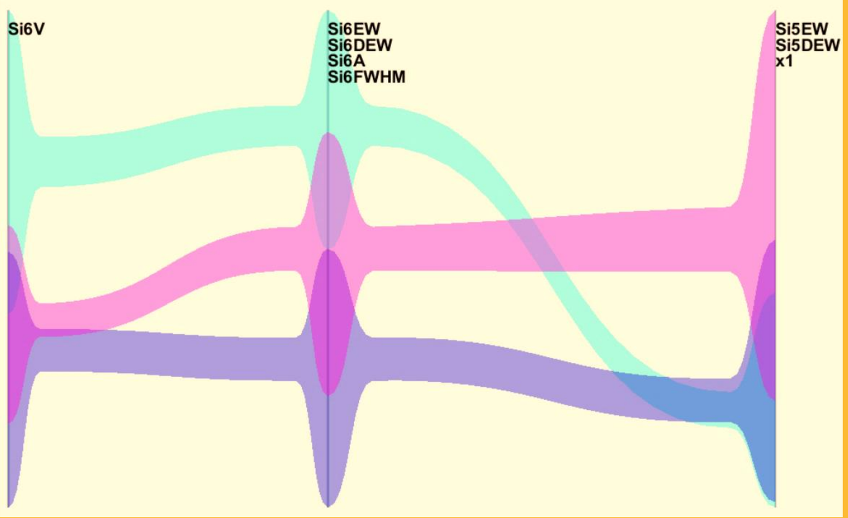
Axis clusters:
- Si II 6355 Velocity
- Si II 6355 EW, DEW, Depth, FWHM
- Si II 5972 EW, DEW, light-curve width
- Broadband color, uncorrected MB
  → Indicating interstellar extinction and reddening
  → Deleted
- z, mB
  → Indicating the distance to the object
  → Deleted
- Corrected MB
  → the intrinsic peak luminosity
  → Deleted

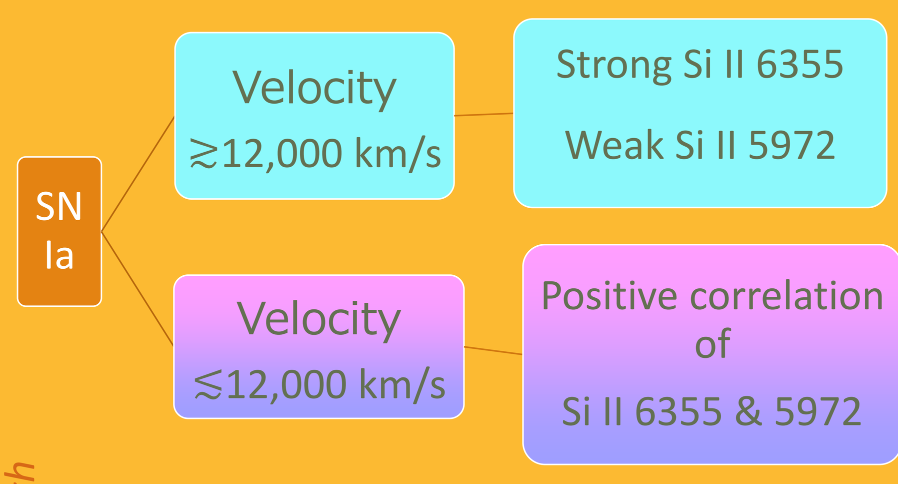




A snap shot of biclustering







A demo can be seen on my laptop!

References: [1] Tibshirani, R. 1996, J. R. Statistical Soc., Ser. B (Methodological), 58, 267, [2] Sullivan, M., et al. 2010, MNRAS, 406, 782, [3] Silverman, J. M., Ganeshalingam, M., Li, W., & Filippenko, A. V. 2012, MNRAS, 425, 1889, [4] Bailey, S., et al. 2009, A&A, 500, L17, [5] Blondin, S., Mandel, K. S., & Kirshner, R. P. 2011, A&A, 526, A81 , [6] Branch D, Dang L C, Hall N, Ketchum W et al 2006 PASP 118 560, [7] Wang X, Filippenko A V, Ganeshalingam M, Li W et al 2009 ApJL 699 L139, [8] Watanabe K, Wu H Y, Niibe Y, Takahashi S and Fujishiro I 2015 Proceedings of IEEE Pacific Visualization Symposium 2015, [9] Watanabe K, Wu H Y, Takahashi S and Fujishiro I 2016 J. Phys. Conf. Ser. 699, 12018