Ia 型超新星の極大等級の変数選択

植村誠
広島大学　宇宙科学センター

概要
Ia 型超新星の極大等級の標準化に適した変数をスパースモデリングの一種である LASSO と交差検定を用いて行った。データは Berkeley グループのデータベースから 78 サンプルを用い、それぞれの色や減光率、スペクトルデータ等を変数候補とした。その結果、従来使われてきた色と減光率が選択され、他の変数はモデルを改善しないことがわかった。

１．はじめに

超新星は宇宙の中でも最大規模の爆発現象で、銀河全体より明るく観測されることもある。超新星のなかでも「Ia 型」と呼ばれるものは爆発極大時の光度が天体によらず一定であることが知られている。そのため、観測からみかけの等級を得ることで、超新星までの距離がわかる。銀河は暗くて見えない場合でも超新星は明るく観測されることがあり、超新星は銀河までの距離を知る手段を与えてくれる。

ただし、Ia 型超新星の極大等級が観測できれば即座に距離が推定できるわけではない。まず、観測される等級は母銀河・天の川銀河の両方で星間ダストの吸収・散乱による減光を受けている。この減光効果は波長の短い、青い光ほど強く受けるため、星間吸収によって天体の色は赤くなる（星間赤化）。実際、観測される超新星の極大等級と色はよく相関し、星間赤化の影響だと考えられている。色で補正された等級は、さらに超新星の減光速度とよく相関することも知られている。[1]　Bバンドでの絶対等級 M_B を目的変数に、「色」（c）と「減光速度」（もしくは光度曲線の幅, x_1）を説明変数にした線形モデル：

$$M_B = M_{B0} + a_1 \cdot c + a_2 \cdot x_1$$

から係数 a_1, a_2 を求め、絶対等級を補正することによって、ようやく距離指標として使うことができる。

しかし、色と減光速度で補正された絶対等級はまだ有意にばらついており、別の変数を加えることでより精度の高い距離指標が得られる可能性がある。そこで、

$$M_B = M_{B0} + a_1 \cdot c + a_2 \cdot x_1 + a_3 \cdot X$$

において、未知の説明変数“X”を探す研究がこれまで多く報告されてきた。候補として

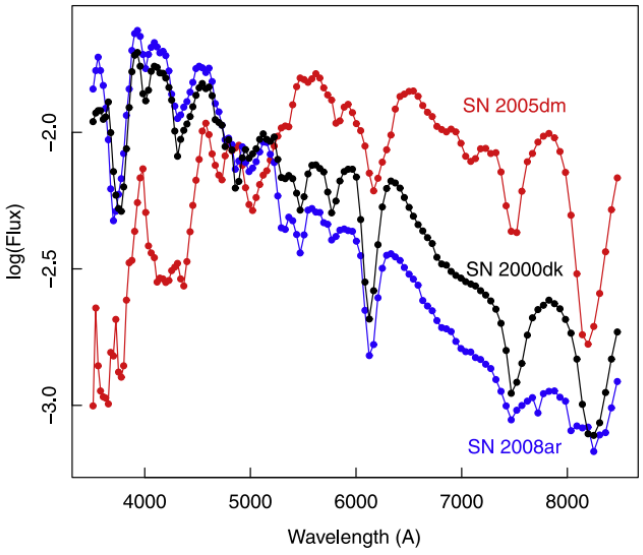

図１．典型的な Ia 型の超新星

は、近年スペクトルデータが充実してきたこともあり、吸収線の強度や速度、さらには2 つの吸収線の強度比、深さの比、そして最近ではスペクトルデータを使い、任意のフラックス比の全てを説明変数の候補とし、有効な変数を探す研究も報告されている。[2]　図1 に典型的なIa型超新星のスペクトルを示す。波長350 nm から850 nm までのスペクトルを134 分割してビンごとに平均したものが丸印である。先行研究で説明変数の候補として使われている「任意のフラックス比」とは、この134 点の任意の組み合わせについてその比をとったもので、134×133 = 17822 個の候補を考えている。なお、データの数は＜100 程度である。

さて、そのような先行研究のモデルは全て目的変数である絶対等級 M_B を N 個の変数 $\{x_i\}$ ($i=1\sim N$)の線形結合モデル

$$M_B = M_{B0} + a_1 x_1 + a_2 x_2 + \cdots + a_N x_N$$

である。$\{a_i\}$ ($i=1\sim N$) は係数である。スペクトルの任意のフラックス比を説明変数候補にすると、変数の数は＞10000 になる。しかし、本当に必要な変数はそのうちの数個だと期待される。「数個」が具体的に何個なのか、どの組み合わせが最も良いモデルなのかが問題となる。

要素のほとんどがゼロで、非ゼロ要素がわずかしかないようなベクトルは「スパース」である、と呼ばれる。このようなスパースなベクトルを高精度に再構成する解析手法はスパースモデリングと呼ばれる。Uemura, et al. (2015) [3] ではスパースモデリングの一種である “LASSO” を用いて、サンプル数よりも変数の数の方が多い状況で、主観によって変数候補を絞ることなく、データ駆動型で有効な変数を選択する手法をIa型超新星の極大等級に応用した。本稿ではその内容を紹介する。次章では手法を説明する。3章ではデータについて、4章では結果、5章では結果に関する議論とまとめを記す。

２．LASSO と交差検定

LASSOは Tibshirani (1996)[4] が提案した過適合を避けるための回帰手法で、1次ノ

ルムを制約項にすることによってスパースなベクトルを再構成する。$\boldsymbol{M_B}$ を測定された絶対等級、X を説明変数の行列、$\boldsymbol{a}$ を係数ベクトルとしたとき、LASSO による解は、

$$\hat{\boldsymbol{a}} = \text{argmin } ||\boldsymbol{M_B} - X\boldsymbol{a}||_2^2 + \lambda||\boldsymbol{a}||_1$$

である。評価式中の第一項はデータとモデルの差の2次ノルムの2乗で、最小二乗項であり、第二項が係数ベクトルの1次ノルムである。

式中のλはスパース度を調整するパラメータで、本研究ではこのλを決定するために交差検定（cross-validation: CV）を用いた。CVではデータを k 個のサブグループ分割し、k-1個のサブグループからモデルを決定、残り1個のサブグループによってモデルの予測誤差を平均二乗誤差（mean-squared error: MSE)等で推定する（k-fold CV）。λが大きすぎるとモデルが単純になり過ぎて、データを説明できず、MSEが大きくなる。一方、λが小さすぎるとデータのノイズ成分を説明するために多すぎる説明変数が使用され、結果、モデルの予測性能は低くなり、MSEは大きくなる。したがって、MSE最小をとるλが最適なモデルである。k-fold CVの場合、テストデータを変えることで各λに対して k個のMSEが計算でき、その加算平均と標準誤差が得られる。λに対してMSEの最小値にその標準誤差を加えて最も少ない説明変数のモデルを採用する手法は“one-standard-error rule”と呼ばれ、今回もそれを用いた。

３．使用データ

Uemura, et al. 2015 [3] ではカリフォルニア大学バークレー校のグループが公開しているIa 型超新星データベースから78 サンプルを用いている。[2] 説明変数の候補として、従来から使用されている色(c)と減光率(x_1)に加えて、2種類の規格化スペクトルを用いた。１つは連続光で規格化したスペクトルで、連続光成分に対する吸収線の情報を持つことが期待される。もう１つは350nm～850nmまでの総フラックスで規格化したスペクトルで、広帯域の色指数(c)よりも狭い波長領域での色情報を持つことが期待される(図1)。先行研究のような任意のフラックス比を用いると変数候補の数が1万個を超えるが、78サンプルからそのような多数の候補を考えると偶然良い変数が混入する可能性が高くなる。本研究では上記の規格化スペクトルのフラックスの対数をとることで、任意のフラックス比がもつ情報をなるべく含み、かつ変数候補の数を減少させた。先行研究で提案されているフラックス比も加え、合計 276個の説明変数候補に対して、1次ノルムの制約項を使うことで、変数の選択を試みた。

４．結果と考察

解析の結果を表１にまとめる。モデル1 は276個の変数候補全てを含めたモデルの結果で、従来から提案されている「色」と「減光率」の他に、いくつかのスペクトルデータが変数として選択された。これらのいくつかは「色」と強く相関するため、次に「色」であらかじめ補正した M_B を目的変数とするモデル２で解析を行ったところ、それらの変数は選択されなかった。さらに、「色」と「減光率」の両方で補正したM_B を目的変数とするモデル3では何も選択されない。一方で、減光率 x_1 を目的変数としたモデル4 では Si II 吸収線の強度に関するフラックスが選択された。この結果は先行研究と一致する。[5] 結論としては、1）絶対等級の説明変数としては色と減光速度の組み合わせが最適で、それ以外の変数はモデルを改善しない、2）減光速度がSi II 吸収線の強度と相関する等、従来の解析結果を再現、という結果が得られた。

表１．解析結果。c, x_1 は色と減光速度、ft は総フラックスで規格化されたスペクトル、fcは連続光で規格化されたスペクトル、R は先行研究で提案されたフラックス比。

モデル	目的変数	選択された非ゼロ係数をもつ説明変数
1	M_B	*c, x1, ft(6373), ft(3752),fc(6084),fc(6289),fc(6631), R(3780/4580)*
2	M_B-a_1 c	x_1
3	M_B-a_1 c-a_2 x_1	(none)
4	x_1	*EW(Si II 4000), fc(5770), fc(3982), fc(7038), fc(6084), ft(4988)*

このようにデータから変数を選択する問題は「変数選択」と呼ばれ、統計学や機械学習の分野で盛んに研究されている。興味のある観測量に関わりそうな要因が多数ある場合、これまではサンプル数を越えないようにあらかじめ変数を主観的に選ぶことがあったかもしれない。しかし、そのような問題はLASSOを使うと事情が全く変わる。サンプル数よりも変数の数が多い場合でも、係数ベクトルがスパースなら解は決まるので、事前に変数を主観で絞らなくてもデータ自身が変数を選んでくれる。ただし、LASSOはλごとに変数の組を決定するのが役割だが、λを決めるのは今回の場合はCV であり、"one standard error rule"である。λが変われば結論も変わるため、λの選択は重要である。この点においてはさらなる研究が必要となるだろう。

参考文献

[1] Phillips, M. M. 1993, ApJ, 413, L105
[2] Silverman, J. M., Ganeshalingam, M., Li, W., & Filippenko, A. V., 2012, MNRAS, 425, 1889
[3] Uemura, M., Kawabata, K. S., Ikeda, S., & Maeda, K., 2015, PASJ, 67, 55
[4] Tibshirani, R. 1996, J. R. Statistical Soc., Ser. B (Methodological), 58, 267
[5] Hachinger, S., Mazzali, P. A., & Benetti, S. 2006, MNRAS, 370, 299