LASSO における AIC の揺らぎとブートストラップ法による近似

統計数理研究所　モデリング研究系　坂田綾香

LASSO においては、AIC が予測誤差の不偏推定量となることが知られている。AIC はデータに応じて揺らぐため、モデル選択の信頼性を確かめるうえでも、揺らぎを定量的に把握することが重要である。このような動機下の研究は新しいものではないが、LASSO においては AIC の分布に関する性質が明らかにされていない。そこで① AIC の真の揺らぎの評価　② ブートストラップ法による近似の精度評価について考える。線形モデルの LASSO は次のように定義される。

$$\min_{\boldsymbol{x}} \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2+\lambda\|\boldsymbol{x}\|_1$$

ここで$\boldsymbol{y}\in\mathbb{R}^M$、説明変数$\boldsymbol{A}\in\mathbb{R}^{M\times N}$、スパース表現$\boldsymbol{x}\in\mathbb{R}^N$とし、$M<N$の場合を考える。

① AIC の真の揺らぎの評価

大偏差原理に従うと、AIC 値Naの分布はレート関数$R(a)\le 0$により次のように与えられる。

$$P(a)\propto\exp\big(NR(a)\big)$$

典型的な値は$R(a)=0$に対応し、これは予測誤差に一致する。パラメータnを導入して次のように母関数$\phi(n)$を定義すると、レート関数により次のように与えられる。

$$\phi(n)\equiv\frac{1}{N}\log\int d\boldsymbol{y}\,d\boldsymbol{A}\exp\big(-nNa(\boldsymbol{y},\boldsymbol{A})\big)=\frac{1}{N}\log\int da\exp\{N(R(a)-na)\}\to\max_a R(a)-na$$

ここで、積分を鞍点評価した。つまり、レート関数のルジャンドル変換が母関数となっているため、母関数に逆ルジャンドル変換を適用することで、レート関数を次のように導出できる。

$$R(a)=\phi-n\frac{\partial\phi}{\partial n}$$

このように導出したレート関数を用いて、AIC のデータ揺らぎを明らかにする。

② ブートストラップ法による近似の精度評価

ここではノンパラメトリックなブートストラップ法を考える。サンプル$\{\boldsymbol{y},\boldsymbol{A}\}$の経験分布に従って生成されたブートストラップ標本を$\{\boldsymbol{y}^\dagger,\boldsymbol{A}^\dagger\}$と表記する。ここで$\boldsymbol{y}^\dagger\in\mathbb{R}^{M_B},\boldsymbol{A}^\dagger\in\mathbb{R}^{M_B\times N}$とする。全てのブートストラップ標本は、変数$\boldsymbol{c}\in\big\{\{0,1,\dots,M_B\}^M\big|\sum_\mu c_\mu=M_B\big\}$を用いて表現できる。$c_\mu$は、データの$\mu$番目の成分がリサンプリングによって引かれる回数を意味する。AIC がある値Naをとるブートストラップ標本の数のエントロピーを次のように定義する。

$$\omega(a)\equiv\frac{1}{N}\log\sum_{\boldsymbol{c}}\delta(a(\boldsymbol{c})-a)$$

ブートストラップ標本に関する母関数を次のように定義すると、エントロピーはその逆ルジャンドル変換から導出できる。

$$\phi^\dagger(\nu)\equiv\frac{1}{N}\log\sum_{\boldsymbol{c}}\exp(-\nu Na)=\frac{1}{N}\int da\exp\{N(\omega(a)-\nu a)\}\to\max_a\omega(a)-\nu a$$

これにより導出したエントロピーを用いて①のレート関数と比較し、ブートストラップ法による AIC の揺らぎの近似評価の精度を明らかにする。