モデル選択手法と漸近的性質について

広島大学・大学院理学研究科・伊森 晋平

Shinpei Imori
Graduate School of Science
Hiroshima University

1　はじめに

統計的モデリングにおいて，複数の候補モデルから最も良いモデルを選択することは重要な問題の一つである．このようなモデル選択問題において，モデルの良さは損失関数やその期待値である期待損失 (リスク関数) でしばしば測られる．代表的なリスク関数の一つに，Kullback-Leibler 情報量 (KL 情報量) [6] に基づくリスク関数がある．このリスク関数の値は真の分布に依存するため，実際の解析時に使用するためには推定が必要である．推定量としては，Akaike information criterion (AIC) [1] がよく用いられており，適当な正則条件のもとで AIC はリスク関数の漸近不偏推定量となることが示される．

AIC はあくまで漸近不偏であるので，サンプルサイズが小さい場合には無視できないバイアスが生じることがある．そこで，AIC のバイアス補正を行った規準量が様々な枠組みで導出されている [例えば 3, 5]．他にも事後確率最大化の観点から導出された Bayesian information criterion (BIC) [8] などがモデル選択にしばしば用いられている．

KL 情報量に基づくリスク関数は将来のデータに対する当てはまりの良さを表す尺度として捉えることができるが，将来のデータはたいてい，現在のデータと同じ分布に従い，同じ変数を持つことが仮定されている．この仮定が成り立たない状況である，共変量シフト [11] や補助変数がある場合 [4] に対して AIC を拡張する研究もこれまでに行われている．

このように多種多様なモデル選択規準が提案されているが，規準の良さは選択されるモデルの漸近的な性質によって定めることができる．本稿では，モデル選択の考え方とその漸近的性質に関して主に他分野の研究者に向けて概略を解説する．近年は Lasso [12] などの L_1 正則化を用いたモデル選択もよく研究されているが，本稿では扱わない．

本稿の構成は以下の通りである．まず第 2 章で本稿におけるモデル選択の枠組みを大まかに説明する．次に第 3 章では，モデル選択規準の漸近的性質としてよく知られる，漸近有効性と一致性を紹介する．最後に第 4 章でまとめを述べる．本稿は広範なモデル選択分野のごく一部に焦点を当てたものであり，決して網羅的ではない．モデル選択問題，特に情報量規準に関して詳しい参考書としては [14] があげられる．

2　モデル選択の枠組み

本章ではモデル選択の枠組みを与える．観測データとして $x_1, \ldots, x_n$ が未知の確率密度関数 $q(x)$ を持つ母集団から独立同一に得られているとする．このとき，$q(x)$ を推定する問題を考える．

候補モデル M に対応する確率密度関数を $p_M(x;\theta_M)$ とする．ただし θ_M は未知パラメータであり，この候補モデルは（解析者自身によって）与えられるものである．候補モデルは一つとは限らず，候補モデルの集合を $\mathcal{M}$ とおく．なお $\mathcal{M}$ の要素数はサンプルサイズ n に依存しても構わない．例えば回帰分析の場合は，M を回帰モデルに用いる説明変数の組み合わせとし，$\mathcal{M}$ を説明変数の組み合わせ全体の部分集合として考えることができる．

各モデル $M \in \mathcal{M}$ におけるパラメータ θ_M を最尤推定で求めるとする．θ_M の最尤推定量 $\hat{\theta}_M$ は次のように与えられる．

$$\hat{\theta}_M = \underset{\theta_M}{\operatorname{argmin}} \sum_{i=1}^{n} \log p_M(x_i;\theta_M).$$

この最尤推定量 $\hat{\theta}_M$ を代入した $p_M(x;\hat{\theta}_M)$ によって $q(x)$ を推定する．この $p_M(x;\hat{\theta}_M)$ の良さの尺度として，KL 情報量を基にした損失関数

$$L(M) = -\int q(x) \log p_M(x;\theta_M) dx \Big|_{\theta_M=\hat{\theta}_M}$$

がしばしば用いられる．また，損失関数のデータに関する期待値 $R(M) = E[L(M)]$ をリスク関数とよぶ．損失関数やリスク関数は候補モデル M の良さの尺度として用いられ，これらの値が小さいほどより良いモデルと捉えられる．

実際には $L(M)$ や $R(M)$ は母集団分布に依存するため未知であり，推定する必要がある．そこで，KL 情報量を基にしたリスク関数（の定数倍）の推定量である AIC がモデル選択によく用いられる．ここで M における AIC は次で与えられる:

$$\mathrm{AIC}(M) = -2 \sum_{i=1}^{n} \log p_M(x_i;\hat{\theta}_M) + 2k_M.$$

ただし，k_M は M における未知パラメータ数である．なお，第二項は罰則項と呼ばれ，$2k_M$ の代わりに $k_M \log n$ を用いたものは BIC と一致する．AIC や BIC などを候補モデルごとに計算し，その値が最も小さくなった候補モデルをベストモデルとして選択する．注意として，AIC によるモデル選択結果と BIC によるモデル選択結果は常に同じになるとは限らない．

例として，二つの候補モデル $\mathcal{M} = \{M_1, M_2\}$ $(k_{M_1} > k_{M_2})$ から AIC でモデルを選択することを考える．このとき，

$$\mathrm{AIC}(M_1) - \mathrm{AIC}(M_2) = -2 \sum_{i=1}^{n} \log \frac{p_{M_1}(x_i;\hat{\theta}_{M_1})}{p_{M_2}(x_i;\hat{\theta}_{M_2})} + 2(k_{M_1} - k_{M_2})$$

が正ならば M_2 が，負ならば M_1 が選択される．BIC で選択する場合には右辺第二項が $\log n(k_{M_1} - k_{M_2})$ に変わり，$n \geq 8$ であれば $\log n > 2$ であるから，このときは AIC より BIC の方がモデル M_2 を選択する確率が増加する．一般に罰則項が大きいほど，よりパラメータ数の小さなモデルが選択されやすくなる．

KL 情報量を基にしたリスク関数の他にも，様々な候補モデルの良さを測る尺度が知られており，またそれに対応するモデル選択規準も存在するが，ここでは説明を割愛する．

3　モデル選択結果の漸近的性質

上述のようにリスク関数などのモデルの良さに依存して，様々なモデル選択規準がこれまでに提案されている．選択されたモデルの漸近的性質を調べることで，モデル選択規準の特徴を捉えることができる．本稿では特に漸近有効性 (asymptotic efficiency) と一致性 (consistency) を紹介する．

モデル選択規準によって選ばれたモデル $\hat{M}$ が

$$\frac{L(\hat{M})}{\min_{M\in\mathcal{M}} L(M)} \xrightarrow{p} 1, \quad n \to \infty$$

を満たすとき漸近損失有効性 (asymptotic loss efficiency) を持つといい，

$$\frac{E[L(\hat{M})]}{\min_{M\in\mathcal{M}} R(M)} \to 1, \quad n \to \infty$$

であれば漸近平均有効性 (asymptotic mean efficiency) を持つという．これは選択されたモデルの損失（あるいはリスク）がその最小値と同程度であることを意味しており，選択されたモデルの良さを保証する性質であると言える．線形回帰モデルにおいて二乗損失関数を用いた場合に，適当な条件下で AIC が漸近損失有効性と漸近平均有効性を持つことが Shao [9] と Shibata [10] でそれぞれ示されている．ここで，二乗損失関数は，

$$L(M) = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \hat{\mu}_{i,M})^2$$

で与えられ，μ_i は i 番目の目的変数の期待値で，$\hat{\mu}_{i,M}$ は候補モデル M における μ_i の推定量である．この結果により，AIC は予測の観点から優れたモデル選択規準であると捉えることができる．

次に，候補モデルの集合 $\mathcal{M}$ の中に真のモデル M_0 が含まれているとする．このとき，モデル選択規準によって選ばれたモデル $\hat{M}$ について

$$Pr(\hat{M} = M_0) \to 1, \quad n \to \infty$$

が成立するとき，そのモデル選択規準は一致性を持つという．真のモデルを選択する確率 $Pr(\hat{M} = M_0)$ は定義関数 $1\{\cdot\}$ を用いて $E[1\{\hat{M} = M_0\}]$ と表されるため，これもリスク関数の性質とみなすことができる．Nishii [7] では，線形回帰モデルにおいて，BIC などのモデル選択規準が一致性を持つための十分条件が示されている．一方で，同じ条件下で AIC は一般に一致性を持たないことが知られている．

しかしながら，近年盛んに研究されている高次元多変量線形回帰モデルにおいては，AIC による変数選択は一致性を持つことが示されている [2, 13]．ただし，ここでの高次元とは，目的変数の次元がサンプルサイズと同程度に大きいことを意味している．一方で，同じ条件下でも BIC は一致性を持たないことも示されており，単変量の場合とは異なる結果が得られている．これは各モデル間のリスク関数がサンプルサイズとともに発散することに起因すると考えられる．

4　まとめ

本稿ではモデル選択の考え方や関連研究の概要を説明し，モデル選択の枠組みを大まかに定式化した．モデル選択規準の特徴はモデル選択結果の漸近的なふるまいによって捉えることができる．そこで，モデル選択規準の漸近的性質の例として漸近有効性と一致性を説明し，AIC や BIC の持つ漸近的性質について紹介した．近年は目的変数の次元が大きい状況での変数選択に対する一致性に関する研究が盛んに行われている．この状況下では AIC と BIC の一致性に関して，単変量の場合と異なる興味深い結果が得られており，さらなる研究の発展が期待される．

参考文献

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Y. Fujikoshi, T. Sakurai, and H. Yanagihara. Consistency of high-dimensional aic-type and cp-type criteria in multivariate linear regression. *Journal of Multivariate Analysis*, 123:184–200, 2014.

[3] Y. Fujikoshi and K. Satoh. Modified AIC and Cp in multivariate linear regression. *Biometrika*, 84(3):707–716, 1997.

[4] S. Imori and H. Shimodaira. An information criterion for auxiliary variable selection in incomplete data analysis. *Entropy*, 21(3), 2019.

[5] S. Imori, H. Yanagihara, and H. Wakaki. Simple formula for calculating bias-corrected AIC in generalized linear models. *Scandinavian Journal of Statistics*, 41(2):535–555, 2014.

[6] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[7] R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765, 1984.

[8] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[9] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242, 1997.

[10] R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics*, 35(3):415–423, 1983.

[11] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[12] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[13] H. Yanagihara, H. Wakaki, and Y. Fujikoshi. A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electronic Journal of Statistics*, 9(1):869–897, 2015.

[14] 小西, 北川. 情報量規準. 朝倉書店, 2004.