

予測制御研究系

ブートストラップと情報量規準 —— 回帰モデルの予測評価 ——

北川 源四郎

1. EIC

真の分布関数を $G(x)$, モデルの密度関数を $f(x|\theta)$ とする. このとき, $G(x)$ からの標本 $x=(x_1, \dots, x_n)$ にもとづいて, 何らかの方法で定義された $y=(y_1, \dots, y_m)$ の密度関数を $h(y|x)$ と表し, y の予測分布と呼ぶ. $h(y|x)$ は多くの推定方式を含んでおり, $h(y|x)=f(y|\hat{\theta}(x))$ とすると最尤推定量などの点推定, $h(y|x)=\int f(y|\theta)p(\theta|x)d\theta$ とするとベイズモデルによる予測分布が得られる.

ここで $-2 \log h(x|x)+2C$ によって AIC と同様に情報量規準が定義できる. ただし, C は $\log h(x|x)$ を平均対数尤度の推定値とみなしたときのバイアスである. 最尤法以外の方法によって求められたモデルに対して C を解析的に評価することは一般に困難であるが, ブートストラップ法を用いると

$$C^* = E_{X^*} \{ \log h(X^* | X^*) - E_{Y^*} \log h(Y^* | X^*) \}$$

によりかなり一般的な仮定のもとで C の近似値が得られる. ただし, X^* と Y^* は経験分布関数 $G_n(x)$ から得られるブートストラップ標本, E_{X^*} と E_{Y^*} は $G_n(x)$ に関する期待値を表す. このとき,

$$\text{EIC} = -2 \log h(x|x) + 2C^*$$

によりブートストラップ法にもとづく情報量規準が得られる. EIC は (i) 最尤推定法以外にも適用できる, (ii) y と x が同一でなくてもよい, (iii) 有限修正などが自動的に行われる, などの特長がある. C^* を実際に求める場合には, 誤差分散を小さくするように工夫することによりブートストラップの反復回数を少なくすることができる.

2. 回帰モデルの予測評価

(多項式) 回帰モデル $x_i = \sum_{j=0}^m a_j x_i^j + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ の場合には以下の手順により EIC を求めることができる.

- (i) パラメータ $\theta = (a_0, \dots, a_m, \sigma^2)$ を推定し, 残差 $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ を計算する.
- (ii) 残差のブートストラップ標本 $\varepsilon_1^*, \dots, \varepsilon_n^*$ を生成する.
- (iii) ブートストラップ標本 $\{y_i^*, x_i^*\}$ を $y_i^* = \sum_{j=0}^{m'} \hat{a}_j(x_i^*)^j + \varepsilon_i^*$ により生成する.
- (iv) (ii)-(iii) を繰り返して C^* を求める.

ここで注意すべきことは, ブートストラップ標本数 m および説明変数 x_i^* は必ずしも実際のデータの n, x_i と同じでなくてもよいということである. したがって, EIC は推定されたモデルを欠測値の補間や外挿に用いたときの推定精度の評価に用いることができる.