ノンパラメトリック重回帰分析におけるLasso型推定

島根大　自然科学研究科　松島 佑樹
千葉大　理学研究院　内藤 貫太

目的変数 $Y \in \mathbb{R}$，d 次元の説明変数 $X \in \mathbb{R}^d$ のデータ $(Y_1, X_1), \ldots, (Y_n, X_n)$ にノンパラメトリック重回帰モデル

$$Y_i = f(X_i) + e_i,\ i = 1, \ldots, n$$

を想定し，回帰関数 f の推定を考える．ここで，$e_i \overset{i.i.d}{\sim} N(0, \sigma^2)$ は誤差であり，$n \ll d$ を仮定しておく．f の推定を行うため，カーネル関数を用いた局所重み付き線形回帰の手法を用いる．推定を行う点を $x \in \mathbb{R}^d$ とし，モデル式を線形式

$$Z = A\theta^* + \varepsilon \tag{1}$$

と変形する．ここで，$Z = (Z_1 \cdots Z_n)^T$，$A = \left(A_1^T \cdots A_n^T\right)^T$，$\varepsilon = (\varepsilon_1 \cdots \varepsilon_n)^T$ であり，$i = 1, \ldots, n$ に対して，

$$Z_i = \alpha_i Y_i,\ \alpha_i = \frac{1}{\sqrt{nh^d}} K\left(\frac{X_i - x}{h}\right),\ A_i = \alpha_i \left(\begin{array}{cc} 1 & \left(\dfrac{X_i - x}{h}\right)^T \end{array}\right),\ \varepsilon_i = \alpha_i e_i + \alpha_i f(X_i) - A_i\theta^*$$

である．また K はカーネル関数で原点対称な d 次元密度関数，$h > 0$ はバンド幅である．$\theta^* = (\theta_0^* \cdots \theta_d^*)^T = (f(x)\ h\partial_1 f(x) \cdots h\partial_d f(x))^T$ がこのモデルでの回帰パラメータベクトルとなり，∂_i は第 i 変数による偏微分作用素を表す．(1) における回帰係数ベクトル θ^* の推定を行う上で，特に $f(x)$ の推定を考えているため，$\theta_0^* = f(x)$ の推定が重要となることに注意しておく．

Bertin and Lecué (2008) では，$f(x)$ において本質的となる変数の選択のために Lasso が用いられ，変数選択の一致性が示されている．しかし，この手法では，θ_0^* にも罰則を付けているため，$f(x)$ の値にも制約を課した上で推定量が求められていることから，$\theta_0^* = f(x)$ の推定精度が悪くなる可能性が考えられる．この問題を改善するため，θ_0^* には罰則を付けない推定方法

$$\min_{\theta \in \mathbb{R}^{d+1}} \left\{ \|Z - A\theta\|_2^2 + 2\lambda \|\theta_{-0}\|_1 \right\} \tag{2}$$

を提案する．ここで $\theta_{-0} \in \mathbb{R}^d$ は θ から θ_0 を取り除いた部分ベクトルであり，λ は正則化パラメータである．

本発表では，理論的結果として，(2) の方法から得られる推定量の変数選択の一致性と，$n \to \infty$，$h \to 0$ における推定量のバイアス，分散の漸近評価について報告し，通常の Lasso を用いて推定した場合との違いについて理論的考察を与える．

参考文献

[1] Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*. **2**, 1224–1241.