クラスタワイズ-ノンパラメトリック回帰分析

大阪大学　　西田 豊

1　はじめに

観測値が複数の母集団から発生しており，層になっているデータに対して回帰分析を行うことを考える．このようなデータに対しては層別に回帰を行うのが適切であるが，各層のラベルが与えられていない状況であるとする．対応策としてクラスタリングしてからクラスターごとに回帰するという考え方があるだろう．しかし，クラスタリングしてから回帰してもデータのクラスター構造をとらえた回帰を行えない場合がある．このようなときは，クラスターラベルと，回帰係数の推定を同時の行うクラスタワイズ回帰分析 (Späth, 1979) を使うことで問題を回避できることがある．

しかしながら，クラスター構造が常に線形関係を持っているとは限らない．例えば非線形なクラスター構造を捉えるためには，従来の線形クラスタワイズ回帰では対応できず，クラスターラベルと，回帰係数の推定に失敗してしまう．本研究では，B-スプライン基底を用いた，ノンパラメトリック回帰モデル (Green & Silverman, 1994) を導入し，非線形なクラスター構造をもつ層別データに対しても適用可能な，クラスタワイズ-ノンパラメトリック回帰分析を提案する．

2　目的関数

ここでは簡単のため単回帰モデルを考える．得られたデータ $\{(y_i, x_i)\}(i = 1, \ldots, n)$ に対して，クラスター数 k の数だけ回帰曲線を当てはめることを考える．個々の回帰曲線には B-スプライン基底関数を用いた，ノンパラメトリック回帰モデルを考える．

$$y_i = f(x_i) + \epsilon_i$$
$$f(x) = \sum_{j=1}^{m} \alpha_j B_j(x)$$

$B_j(x)$ は B-スプライン基底関数，α_j をスプライン回帰における回帰係数とする．提案手法の目的関数は以下のようにかける．

$$F(\boldsymbol{\alpha}_k, u_{ik}) = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik}(y_i - \mathbf{b}_i'\boldsymbol{\alpha}_k)^2 + \lambda \boldsymbol{\alpha}_k' \mathbf{G} \boldsymbol{\alpha}_k$$

ここで，$\mathbf{b}_i$ はスプライン基底関数のベクトル，$\boldsymbol{\alpha}_k$ はクラスターごとの回帰係数ベクトルとする．また，$\mathbf{G}$ は $G_{jj'} = \int \frac{d^2 B_j(x)}{dx^2} \frac{d^2 B_{j'}(x)}{dx^2} dx$ であり，λ は $\lambda > 0$ で推定する関数の滑らかさを調節する平滑化パラメータである．u_{ik} は i 番目の観測値がクラスター k に所属するか否かを示すインデックスとする．ただし $u_{ik} = \{0, 1\}$ かつ $\sum_{k=1} u_{ik} = 1$ を満たす．目的関数を最小にする $\boldsymbol{\alpha}_k$ と u_{ik} を求める．アルゴリズムには，パラメータ $\boldsymbol{\alpha}_k$ と u_{ik} を交互に最適化する，交互最小二乗法を用いる．

引用文献

Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. CRC Press.

Späth, H. (1979). Algorithm 39: clusterwise linear regression. *Computing*, **22**, 367–373.