L_1 正則化法に基づく関数回帰モデリング

九州大学大学院数理学府　松 井 秀 俊
九州大学大学院数理学研究院　小 西 貞 則

1. はじめに

縮小推定は，モデル推定に伴う最適化問題において，パラメータに制約を課す推定法で，これにより安定した推定量を得ることができる．特に，SCAD ペナルティ (Fan and Li, 2001) に基づく縮小推定は，その制約の性質から変数選択の役割を担うなど優れた性質を有しており，線形モデルをはじめとした様々なモデルに対する適用例が報告されている．本報告では，関数データとして与えられた説明変数と，スカラーとして与えられた目的変数との関係をモデル化する関数回帰モデルの推定問題に対して，制約の一つである group SCAD (Wang *et al.*, 2007) を用いることにより，モデルの推定および変数選択を同時に行う方法を提案する．さらに，縮小推定に伴う正則化パラメータを選択するためのモデル評価基準を導出する．

2. 関数回帰モデル

いま，スカラーとして与えられた目的変数と，関数データとして与えられた M 変量の説明変数に対して，n 組のデータ $\{(y_\alpha, x_{\alpha m}(t));\, t \in \mathcal{T},\, \alpha = 1, \cdots, n,\, m = 1, \cdots, M\}$ が観測されたとする．このとき，説明変数と目的変数との関係を次のようにモデル化する (Ramsay and Silverman, 2005).

$$y_\alpha = \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}} x_{\alpha m}(t)\beta_m(t)dt + \varepsilon_\alpha, \quad \alpha = 1, \cdots, n. \tag{1}$$

ただし，β_0 は定数項，$\beta_m(t)$ は係数関数を表し，ノイズ ε_α は互いに独立で平均 0，分散 σ^2 の正規分布に従うものとする．また，$x_{\alpha m}(t)$ および $\beta_m(t)$ は共通の基底関数 $\boldsymbol{\phi}_m(t) = (\phi_{m1}, \cdots, \phi_{mp_m})'$ によって $x_{\alpha m}(t) = \boldsymbol{w}'_{\alpha m}\boldsymbol{\phi}_m(t)$，$\beta(t) = \boldsymbol{b}^{*'}_m\boldsymbol{\phi}_m(t)$ と表されるとする．ここで，$\boldsymbol{w}_\alpha = (w_{\alpha 0}, w_{\alpha 1}, \cdots, w_{\alpha p_m})'$ は平滑化によって推定された p_m 次元係数ベクトルを表し，$\boldsymbol{b}^*_m = (b^*_{m1}, \cdots, b^*_{mp_m})'$ は係数パラメータベクトルとする．これより，(1) 式は次のように書き換えることができる．

$$y_\alpha = \beta_0 + \sum_{m=1}^{M} \boldsymbol{w}'_{\alpha m}\boldsymbol{J}_{\phi_m}\boldsymbol{b}^*_m + \varepsilon_\alpha = \boldsymbol{z}'_\alpha\boldsymbol{b} + \varepsilon_\alpha. \tag{2}$$

ここで $\boldsymbol{z}_\alpha = (1, \boldsymbol{w}'_{\alpha 1}\boldsymbol{J}_{\phi_1}, \cdots, \boldsymbol{w}'_{\alpha M}\boldsymbol{J}_{\phi_M})'$，$\boldsymbol{b} = (\beta_0, \boldsymbol{b}^{*'}_1, \cdots, \boldsymbol{b}^{*'}_M)'$，$\boldsymbol{J}_{\phi_m} = \int_{\mathcal{T}} \boldsymbol{\phi}_m(t)\ \boldsymbol{\phi}'_m(t)dt$ とする．従って，関数回帰モデル (1) は確率密度関数

$$f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{b}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_\alpha - \boldsymbol{z}'_\alpha\boldsymbol{b})^2}{2\sigma^2}\right\}$$

で表され，モデルの推定問題は従来の回帰モデルの推定問題へ帰着される．

3. Group SCAD ペナルティに基づく推定，評価

パラメータ $\boldsymbol{\theta} = \{\boldsymbol{b},\ \sigma^2\}$ の推定は，モデルの対数尤度関数から係数パラメータに関する制約を差し引いた正則化対数尤度関数

$$l_\lambda(\boldsymbol{\theta}) = \sum_{\alpha=1}^{n} f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{b}, \sigma^2) - n\sum_{m=1}^{M} p_\lambda(\|\boldsymbol{b}^*_m\|_2) \tag{3}$$

の最大化によって行う．ここで，$p_\lambda(\cdot)$ はSCADペナルティで，一階微分が次で与えられる形をもつ．

$$p'_\lambda(|\eta|) = \lambda \left\{ I(|\eta| \le \lambda) + \frac{(a\lambda - |\eta|)_+}{(a-1)\lambda} I(|\eta| > \lambda) \right\}.$$

ここで，λ は制約の度合を調整する正則化パラメータ，a は調整パラメータとする．SCADは，lassoおよびhard thresholding双方のよい性質を取り入れたものと考えることができ，推定量に不偏性を与えるとともに安定した推定を可能にし，"oracle property"とよばれる，真のモデルを適切に選択する性質をもつ．また，$p_m \times p_m$ 行列 $\boldsymbol{G}$ に対して $\|\boldsymbol{b}_m^*\|_2 = (\boldsymbol{b}_m^{*'}\boldsymbol{G}\boldsymbol{b}_m^*)^{1/2}$ とする．このような制約を置くことによって，仮に m 番目の説明変数 $X_m(t)$ が不要な場合，この変数に関わる p_m 個の係数の推定量を同時に0に縮小することができる (Yuan and Lin, 2006)．

SCADをはじめとした L_1 ノルムを含むペナルティは，パラメータに関する絶対値が含まれており微分可能でないため，正則化最尤推定量の解析的な導出が困難となる．そのため，解析的あるいは数値的な近似手法を用いる必要がある．ここでは，Fan and Li (2001) によって提案された，SCADペナルティを2次関数で近似し，反復的に推定量を求める方法を用いる．その結果，パラメータ $\boldsymbol{b}$，σ^2 の更新値はそれぞれ次で与えられる．

$$\tilde{\boldsymbol{b}} = \left(\boldsymbol{Z}'\boldsymbol{Z} + n\sigma^2\Sigma(\boldsymbol{b})\right)^{-1} \boldsymbol{Z}'\boldsymbol{y}, \quad \tilde{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{b}})'(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{b}}). \tag{4}$$

ただし $\Sigma(\boldsymbol{b}) = \mathrm{diag}\{0, p'_\lambda(\|\boldsymbol{b}_1^*\|_2)/\|\boldsymbol{b}_1^*\|_2\boldsymbol{1}_{p_1}, \cdots, p'_\lambda(\|\boldsymbol{b}_M^*\|_2)/\|\boldsymbol{b}_M^*\|_2\boldsymbol{1}_{p_M}\}$ とする．

推定されたモデルは，制約の度合を調整する正則化パラメータに依存しているため，この値の適切な選択が必要となる．そのための選択基準として，一般化情報量規準GIC (Konishi and Kitagawa, 2008) を，group SCADに基づき推定された関数回帰モデルを評価するために導出したものを用いる．

4. 適用例

提案したモデリング手法の有効性を，数値例および実データの解析を通して検証した．実データの解析では，日本のいくつかの地点で月別に観測された年間の平均気温，平均気圧といった6種類のデータのうち，年間の総降水量と関連が強いものを選択することを目的としている．そのために，月別観測データを時間の関数と考えた関数データとして扱い，これらを説明変数，総降水量を目的変数とみなして関数回帰モデルを構築した．そして，group SCADを用いてモデルの推定および関数データの選択を行い，モデル評価基準を用いて正則化パラメータを選択した．その結果，平均気温と平均気圧，最低気温に関わる係数パラメータが全て0と推定され，日照時間，平均湿度，最高気温の3つの変数が降水量と関連が強いという結果を得た．

参考文献

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Konishi, S., and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis (2nd ed.)* Springer.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49–67.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486–1494.