1K-01

最大次数が未知の多項式回帰におけるスパース推定

井上　一磨　　須子　統太
早稲田大学社会科学部

1.1 研究背景

近年, 統計学, 機械学習の分野でスパース推定に関する研究が数多く行われている[1]. 他方, 多項式回帰モデルは古くから研究されているモデルで, 様々な応用研究が行われている. 最近の研究では, 多項式回帰モデルとスパース推定を組み合わせることで, 実データに対して有効性がある場合が報告されている[2]. しかし, 従来の多項式回帰モデルに対するスパース推定では, 最大次数をあらかじめ固定する必要があった.

そこで本研究では, 最大次数が未知の多項式回帰におけるスパース推定法を提案する. また, 人工データと実データに対する数値実験により, 提案手法の有効性を示す.

1.2 問題設定

まず, 本研究の問題設定を説明する. i 番目の説明変数を$\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{ip})^T$, 目的変数を$y_i$とする. ただし, Tは転置を表すとする. また, $x_{ij}, y_i \in \mathbb{R}$とする. $f_d(\boldsymbol{x}, \boldsymbol{\beta})$を d 次の多項式関数とし, $\boldsymbol{\beta}$を係数ベクトルとする.

例えば, $\mathrm{p} = 2, \mathrm{d} = 3$のとき,

$$f_3(\boldsymbol{x}, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \beta_{112} x_1^2 x_2 + \beta_{122} x_1 x_2^2 + \beta_{111} x_1^3 + \beta_{222} x_2^3$$

となる.

本研究では,

$$\mathrm{y} = f_d(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon \tag{1}$$

で表されるd次多項式回帰モデルの回帰係数$\boldsymbol{\beta}$を推定することを目的とする. ただし, 本研究ではdは未知で, $\boldsymbol{\beta}$は非ゼロの係数が少ない疎なベクトルであると仮定する. ここで, 説明変数の行列Xを

$$X = \begin{bmatrix} \boldsymbol{x_1} \\ \vdots \\ \boldsymbol{x_n} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{2}$$

と定義する. また, 目的変数のベクトル$\mathbf{y}$を

$$\mathbf{y} = (y_1, \cdots, y_i, \cdots, y_n)^T \tag{3}$$

と定義する.

A Sparse Estimation Method for Polynomial Regression Model with Unknown Maximum Degree

Kazuma Inoue (lotte01433@fuji.waseda.jp)
Tota Suko (suko@waseda.jp)
School of Social Sciences, Waseda University

1.3 従来研究

Huang[1]らは, 多項式回帰モデルに対するスパース推定を行うために, SPORE-LASSO を提案した. まず, LASSO[2]を用いて, 少数の変数をXから選択する. 次に残った変数をd次展開する. 最後にAdaptive-LASSO[3] を用いて係数を推定する. Adaptive-LASSO は変数間に相関がある場合に, 用いられるスパース推定法である.

この手法では, d をあらかじめ固定しておく必要がある. また, d を大きくとると d 次展開により指数的にパラメータ数が増えるという問題がある.

2　提案手法

従来の手法では, 最大次数を固定したもとでのみ多項式回帰モデルのパラメータの推定が行われていた. 一方, 提案手法では, 最大次数が未知の場合におけるパラメータの推定手法を提案する.

ここで, 2 つ関数を定義する. Sはインデックスを表し,

$$\Omega(X, \mathrm{S}) \tag{4}$$

は元の行列Xからインデックスのある列のみ抽出することを表す. 例えば,

$$\mathrm{S} = \{1,2,3\}$$

のときには,

$$\Omega(X, \mathrm{S}) = \begin{bmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} \\ \vdots & \vdots & \vdots \\ 1 & \mathrm{x}_{\mathrm{n}1} & \mathrm{x}_{\mathrm{n}2} \end{bmatrix}$$

となる.

次に,

$$\mathrm{expan}(X', X) \tag{5}$$

を定義する. 後ろの行列の各列を前の行列の各列に掛けて, 列ベクトルを得た後, それらを結合して新しい行列を得ることを表す. 例えば,

$$X' = \begin{bmatrix} x_{11}^2 & x_{11}x_{12} \\ \vdots & \vdots \\ x_{n1}^2 & x_{n1}x_{n2} \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

とすると,

$$\mathrm{expan}(X', X) = \begin{bmatrix} x_{11}^2 & x_{11}x_{12} & x_{11}^3 & x_{11}x_{12}^2 & x_{11}^2x_{12} & x_{11}^2x_{13} & x_{11}x_{12}x_{13} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1}^2 & x_{n1}x_{n2} & x_{n1}^3 & x_{n1}x_{n2}^2 & x_{n1}^2x_{n2} & x_{n1}^2x_{n3} & x_{n1}x_{n2}x_{n3} \end{bmatrix} \tag{6}$$

となる.

表 1:実データに対する分析結果

手法	Auto MPG	Concrete Slump Test(slump)	Concrete Slump Test(flow)	Concrete Slump Test(strength)	Airfoil	Yacht Hydrodynamics
Lasso(重回帰)	0.2306	0.8258	0.5145	0.1570	25.1340	0.4017
最小二乗法(重回帰)	0.2022	0.8699	0.5863	0.1132	22.9356	0.3881
ランダムフォレスト	**0.1634**	**0.7156**	0.5617	0.3659	15.4423	0.1128
SPORE-Lasso(d=2)	0.1652	0.8306	0.5417	**0.0400**	17.5981	0.0918
SPORE-Lasso(d=3)	0.1747	1.1076	0.5331	0.1985	**15.4192**	0.0164
提案 1	0.1648	0.7424	**0.4876**	0.0822	19.5461	**0.0041**
提案 2	0.3030	0.9283	0.8556	0.2472	17.3880	0.1259

[提案 1]

1. $X_{old} = X$
2. $\widetilde{\boldsymbol{\beta}} = arg\min_{\beta} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$
3. $\mathrm{S} = \{j : \widetilde{\boldsymbol{\beta}} \neq 0\}$
4. $X' = \Omega(X, \mathrm{S})$
5. $X_{new} = expan(X', X)$
6. $while\ X_{new} \neq X_{old}$
   - I. $X_{old} = X_{new}$
   - II. $\widetilde{\boldsymbol{\beta}'} = arg\min_{\beta} \|\boldsymbol{y} - X_{new}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$
   - III. $\mathrm{S}' = \{j : \widetilde{\boldsymbol{\beta}'} \neq 0\}$
   - IV. $X' = \Omega(X_{new}, \mathrm{S}')$
   - V. $X_{new} = expan(X', X)$

収束したときの$\widetilde{\boldsymbol{\beta}'} = \widehat{\boldsymbol{\beta}}$とする.
ただし, 収束しない場合は任意の繰り返し回数で打ち切る.

[提案 2]

提案 2 は 6.内の II を(7)式に置き換えたものとする.

$$\widetilde{\boldsymbol{\beta}'} = arg\min_{\beta} \|\boldsymbol{y} - X_{new}\boldsymbol{\beta}\|_2^2 + \lambda_2 \sum_j \left|\frac{\beta_j}{w_j}\right| \quad (7)$$

重みベクトルw_jは,

$$w_j = \left({X_{new}}^T X_{new} + 0.001I\right)^{-1} {X_{new}}^T \boldsymbol{y} \quad (8)$$

で求める. また, λ_1, λ_2はハイパーパラメータである.

このように繰り返し行列の更新を行うことで最大次数が未知の場合にも多項式回帰モデルに対するパラメータの推定が可能になる.

3 有効性の検証

3.1 人工データによる実験

提案手法の有効性を検証するためシミュレーション実験を行う. あらかじめパラメータを設定した多項式回帰モデルに対して, 人工的に発生させたデータによってパラメータの推定を行った. Xとεは正規分布に従って発生させた. サンプル数$\mathrm{n} = 500$, 変数の数$\mathrm{p} = 3$とした. また, 本研究では, λ_1は誤差が最小の 1 標準誤差以内になるように設定し, λ_2は BIC を最小にするように設定した. そのもとで, 2 つの提案手法によって推定された値と真のパラメータ$\boldsymbol{\beta}$の値をプロットした図を, 図 1 に示す.

図 1 より, 提案 1,2 共に非ゼロの係数を良く推定できていることが分かる. また, 提案 2 の方が提案 1 に比べて比較的推定値が真の値に近いことが分かる.

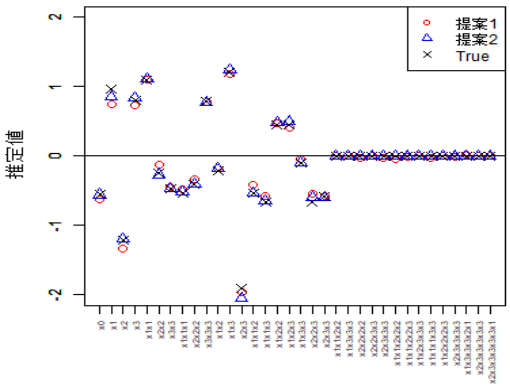

図 1:多項式回帰モデルに対する実験結果(p=3,n=500)

3.2 実データに対する実験

次に, UCI Machine Learning Repository のベンチマークデータに対し, 提案手法を用いた実験を行う.

まず, 実験の手順について説明する. データを学習データとテストデータに分け, 学習データでパラメータを推定し, そのもとでテストデータの目的変数を予測する. 最小二乗誤差(MSE)によって提案手法の有効性の評価を行う. 実験結果を表 1 に示す.

表 1 より, 特定のデータに対して提案手法の予測性能が良いことが分かる. また, 提案 1 の方がシミュレーションと異なり優れていることが多い. さらに, 提案 1 は収束することが多いのに対し, 提案 2 は収束することが少なかった.

4 まとめと今後の課題

本研究では, 最大次数が未知の多項式回帰に対するスパース推定法を提案し, 提案手法に対するシミュレーションと実データにおける検証結果を行った. 今後の課題として, 収束基準の検証や, 変数間の相関の強さが提案手法に及ぼす影響の検討などが挙げられる.

参考文献

[1]冨岡亮太"スパース性に基づく機械学習" 講談社サイエンティフィック,2015
[2] L. Huang, J. Jia, B. Yu, B. G. Chun, P. Maniatis, and M. Naik. "Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression." In Proc. NIPS, 2010.
[3] R. Tibshirani. "Regression shrinkage and selection via the lasso." J. Royal. Statist. Soc B., 1996.
[4] H. Zou. "The adaptive lasso and its oracle properties." Journal of the American Statistical Association, 101(476):1418–1429, 2006.