回帰分析における Lasso 正則化を用いた過学習への対応と変数選択

寺坂 知佳　(指導教員：吉田 裕亮)

1　はじめに

一般に, 回帰分析とは説明変数からなるモデル式に基づいて, その応答とされる目的変数の振る舞いを予測する方法である．モデル式が説明変数の線形和の場合には線形回帰モデルと呼ばれている．単純多項式回帰とは, ひとつの説明変数の冪たちからなる多項式をモデル式とする回帰手法であり, 非線形回帰モデルの代表的例でもある．

回帰分析を行う際に説明変数の数が増すと, モデル式のデータへのフィティング精度は上がると同時に過剰なフィッティング現象も発生する．これは統計的機械学習では過学習 (オーバ・フィティング) とよばれ, 最適な予測がされない．回帰分析において過学習を抑える手法の一つとして正則化が知られている．

本研究では正則化手法のうち Lasso 正則化とモデル選択手法での情報量基準を用いて, 多項式回帰モデルにおいて過学習を抑えながら最適な多項式モデルの最大次数とモデルに必要な最適な単項式の効率的な推定手法の提案する．

2　正則化

正則化とは過学習を抑えるために罰則項を加え, 最適化することである．回帰分析における正則化には Ridge 正則化と Lasso 正則化が良く知られている．Ridge 正則化では回帰係数の 2 乗和を罰則項とし, Lasso 正則化では回帰係数の絶対値和が罰則項となる．なおこれらは罰則項の状況に応じて, それぞれ L^2-正則化, L^1-正則化とも呼ばれている．

本研究では外れ値の大きさに応じて幾つかの回帰係数が 0 に近づくように振る舞い, モデルの予測応答が滑らかとなる Lasso 正則化を用いる．

多項式回帰モデル

$$Y = \sum_{j=0}^{d} b_j X^j + \varepsilon$$

における Lasso 正則化 (L^1-正則化) は, 与えられたデータ点を $\{(x_i, y_i)\}_{i=1}^{n}$ としたとき

$$\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{d} b_j x_i^j \right)^2 + \lambda \sum_{j=1}^{d} |b_j| \right\}$$

で与えられる．ただし, $\lambda > 0$ は罰則項の重みとなる正則化パラメータである．

3　変数選択

モデル選択を応用して, 変数選択が可能である．本研究で用いる AIC, BIC はともに分布に基づいて尤度を援用して, パラメタライズされた統計モデルにおいて予測誤差とパラメータ数のトレードオフの関係を調整することで, 最適なモデルを選択する基準である．

AIC, BIC は以下の式である．

$$\mathrm{AIC} = -2\log L + 2k$$
$$\mathrm{BIC} = -2\log L + k\log n$$

ただし, $\log L$ は最大対数尤度, k は自由パラメータ数, n はデータ数である．

BIC は AIC と比べて一般的に多くの変数を持つモデルに重いペナルティを課すことになるため AIC より小さなモデルが選択される．そのため, AIC と BIC で最適モデルが異なる場合もある．

4　提案手法

本研究では多項式回帰のモデル決定を以下のように行うことを提案する．

1. 多項式モデルの最高次数を変えて最小 2 乗法を行い, AIC(または BIC) を用いて最高次数を推定する．
2. 1. で推定した次数の多項式回帰で Lasso 正則化を行う．このとき正則化パラメータ $\lambda > 0$ を大きくしていきながら 0 に近づきモデルにおいて不要となる係数の候補を見る．
3. 候補となった係数の変数を加えたり除いたりして再び AIC(または BIC) に基づいて変数選択を行い, 最適な多項式モデルを推定する．

5　シミュレーションデータによる数値実験

以下の多項式にノイズをのせてシミュレーションデータとして使用した．

$f(x) = x^5 - 1.25x^3 + 0.25x$

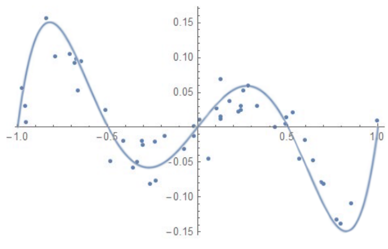


図 1: 実線: 元のモデル / 点: ノイズをのせた図

このデータに 2 節の多項式モデルを考え, 最高次数 d を変化させ回帰を行い, AIC, BIC を見ると次のような結果になった．

d	AIC	BIC
1	−292.575	−290.663
2	−290.578	−286.754
3	−293.505	−287.769
4	−291.577	−283.928
5	**−370.291**	**−360.731**
6	−368.509	−357.037
7	−367.896	−354.511

AIC, BIC ともに $d=5$ が最適と推定される．

$d=5$ とし Lasso 正則化を行い，回帰係数の変化を見ると以下のようになり，次数 0, 2, 4 のいずれかは影響のない変数であると考えられる．

λ	b_0	b_1	b_2
0.00001	−0.00613460	0.265741	0.0457848
0.00010	−0.00610075	0.265380	0.0454933
0.00100	−0.00576231	0.261768	0.0425788
0.01000	−0.00237777	0.225650	0.0134320

λ	b_3	b_4	b_5
0.00001	−1.35451	−0.0472083	1.104700
0.00010	−1.35294	−0.0468904	1.103370
0.00100	−1.33721	−0.0437116	1.090100
0.01000	−1.17992	−0.0119226	0.957364

次に単項次数 $0,2,4$ の変数を残したり抜いたりして再び変数選択を行う．

選択した次数	AIC	BIC
$0,1,2,3,4,5$	−364.111	−352.639
$1,2,3,4,5$	−365.377	−355.817
$0,1,3,4,5$	−365.328	−355.768
$1,3,4,5$	−367.279	−359.631
$0,1,2,3,5$	−365.556	−355.996
$1,2,3,5$	−367.312	−359.664
$0,1,3,5$	−367.328	−359.680
$1,3,5$	**−369.279**	**−363.543**

次数 $1,3,5$ を選択したときに最適な多項式モデルになると推定され，最初に設定した多項式と同じとなり正しい多項式モデルが推定されたと考えられる．

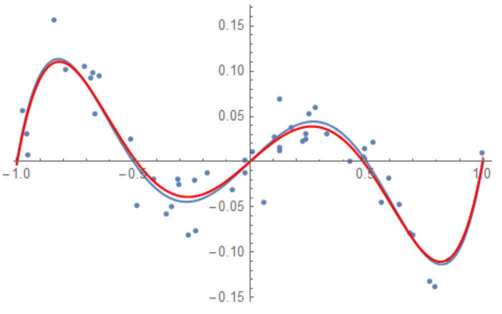


図 2: 青線: 元のモデル / 赤線: 推定されたトレンド

6 実データへの応用

図 3 は東京の 2020 年の日平均湿度のデータである．$x=0.0$ を 1 月 1 日，$x=1.0$ を 12 月 31 日とする．

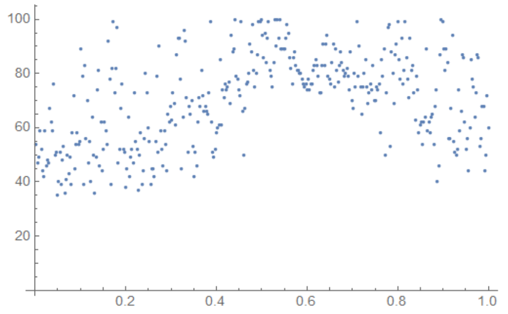


図 3: 東京の日平均湿度のデータ

シミュレーションデータを用いたときと同様に最高次数 d を変化させながら多項式回帰を行い，最高次数を推定するために AIC, BIC を用いると次のような結果になった．

d	AIC	BIC
1	1998.50	2002.40
2	1906.27	1914.07
3	1892.68	1904.38
4	1887.63	**1903.23**
5	1885.74	1905.24
6	**1880.32**	1903.72
7	1881.97	1909.27

AIC, BIC で推定される最高次数は 6, 4 であるが，より簡単なモデルを選択するため最高次数は 4 と推定する．

次に回帰係数を見ると以下のようになる．

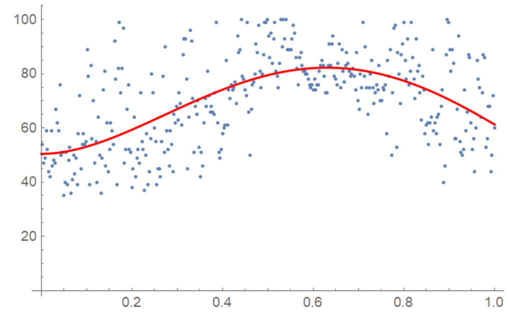
λ	b_0	b_1	b_2
0.01	55.3839	−77.7156	616.336
0.10	54.9516	−69.7650	582.255
0.50	53.0303	−34.4288	430.782
0.90	51.1663	−0.0000005	282.866

λ	b_3	b_4
0.01	−913.216	382.362
0.10	−861.920	357.328
0.50	−633.938	246.066
0.90	−411.002	137.163

以上より，単項次数 1 は影響のない変数といえる．AIC, BIC により推定すると以下のようになる．

選択した次数	AIC	BIC
$0,1,2,3,4$	1891.56	1918.86
$0,2,3,4$	**1889.56**	**1912.96**

次数 1 の項を抜いた時に最適モデルになるため，最適な多項式トレンドは以下のようになる．



7 まとめ

本研究では多項式回帰モデルにおいて Lasso 正則化と情報量基準を用いて過学習を抑え，多項式トレンドを抽出する方法を与えた．削除される変数の候補を先にあげておくことによって全ての組み合わせを調べる必要がなくなり計算量が大幅に削減される．

参考文献

[1] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, New York, NY (2017).