2 次導関数を正則化項に持つルジャンドル多項式回帰について

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.

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.

本研究では回帰関数の推定において, 2 次導関数を罰則化項にもつ正則化により, 過学習が抑える手法について考察する.

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.

本研究では 1 変数の回帰手法による近似関数 f を, 関数の 2 次導関数の 2 乗積分値を罰則項とする正則化により推定する. すなわち,

$$\min\left\{\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda\int\{f''(t)\}^2\right\}$$

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　クロスバリデーション

クロスバリデーションは予測誤差を推定する方法で最もよく知られている. 標本データを分割し, 一部をテストデータとして取り置き, 残りを学習用トレーニングデータとして扱う.

本研究では, よく用いられる 5-fold クロスバリデーションを援用して正則化パラメータ λ の推定を行う. 5-fold クロスバリデーションは, 以下のように行われる.

1. データを 5 つにランダムに分割し, 1 つをテストデータ, 残り 4 つをトレーニングデータとして学習する.
2. テストデータによる平均 2 乗誤差を計算する.
3. 上の 1. 2. の手順をテストデータを変えながら 5 回繰り返す.

これにより, 平均 2 乗誤差の 5 つの推定値が得られ, 5-fold クロスバリデーションの推定値はこれらを平均値により得られる.

4　ルジャンドル多項式

本研究では, 関数近似の基底として, ルジャンドル (Legendre) 多項式を考えることにした.

ルジャンドル多項式は, 重みがつかない (一様分布) の有界区間 (コンパクト台) 上の直交多項式である. すなわち, 直交関係は 閉区間 $[-1, 1]$ の一様分布 による L^2-内積に関して直交する.

$$\int_{-1}^{1} P_m(x)P_n(x)\,\mathrm{d}x = \frac{2}{2n+1}\delta_{mn}$$

ただし, δ_{mn} はクロネッカーのデルタである.

ルジャンドル多項式を用いる利点は, 以下のことが考えられる. 重み分布が一様であることから, 1 変数のランダム点での観測値や等間隔な時系列データの関数近似に適していると考えられる.

また, 等間隔観測の時系列データから 5-fold クロスバリデーションのように, ランダムにテストデータを取り置き, 学習を行う場合にも適している.

単項式を基底として用いた多項式回帰では, 単項式間の直交性は低く, また区間の両端での変動が大きいことが, しばしば問題となる.

なお, ルジャンドル多項式は, 直交性より低次から帰納的に求めることも可能であるが, 以下の式としても、与えられる.

$$P_n(x) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n$$

5　多項式での数値実験例

本研究では多項式回帰において, 比較的低次数のモデルにおいて次数決定が可能かを確認するために, 以下のような実験を行った.

まず, 高い次数までのルジャンドル多項式を行い, 回帰モデルにおいて過学習が発生する状態を設定する. 2 次導関数を罰則項に加え, 正則化パラメータ λ を変化させながら 5-fold クロスバリデーションにより適切な λ を定める.

ここでは多項式 $x^3 - 0.81x$ に正規ノイズを加え, ランダムな 50 個の点を生成し, シミュレーションデータとした.

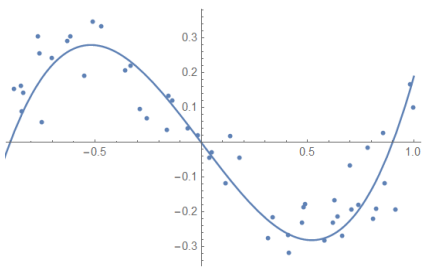


図 1: 実線: 元のモデル / 点: ノイズをのせた図

$d = 12$, $\lambda = 0$ でルジャンドル多項式回帰を行うと，推定されたモデルは，以下のようになった．これは，高次数で正則化項が無い状態であり，過学習が発生する状況になっている．

b_0	0.0473116	b_7	−0.2354310
b_1	−0.3505270	b_8	0.2553130
b_2	0.1864690	b_9	−0.0688046
b_3	0.0735535	b_{10}	0.1008480
b_4	0.2132620	b_{11}	−0.0829278
b_5	−0.2653070	b_{12}	0.0924652
b_6	0.2520580		

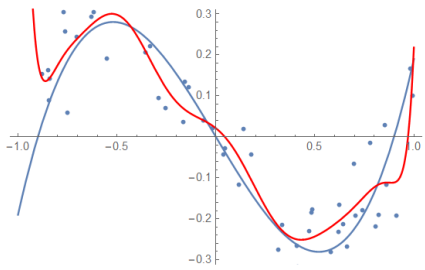


図 2: $d = 12$, $\lambda = 0$ のとき推定されたモデル

ここで λ を変化させながら，クロスバリデーションを用いて平均 2 乗誤差をみると以下のようになった．

λ	平均 2 乗誤差
0.0001	0.0670619
0.0010	0.0641861
0.0100	0.0665166
0.1000	0.0981361

λ の値が 0.0010 前後のとき，平均 2 乗誤差が最小になると考えられる．前後を調べると，$\lambda = 0.0011$ と推定された．

$d = 12$, $\lambda = 0.0011$ でルジャンドル多項式回帰を行うと，回帰係数は以下のようになり，次数 4 以上の係数は小さく影響のないものと考えられる．さらに 1 次と 3 次だけが有効であり，0 次と 2 次の影響も小さいことも係数から読み取れる．

b_0	0.01594500	b_7	−0.02761110
b_1	−0.24076900	b_8	0.03198990
b_2	0.04207410	b_9	0.03163150
b_3	0.28340900	b_{10}	−0.02617450
b_4	−0.00145807	b_{11}	−0.01222250
b_5	−0.01685010	b_{12}	0.00571114
b_6	0.00791396		

推定されるモデルは以下のようになった．

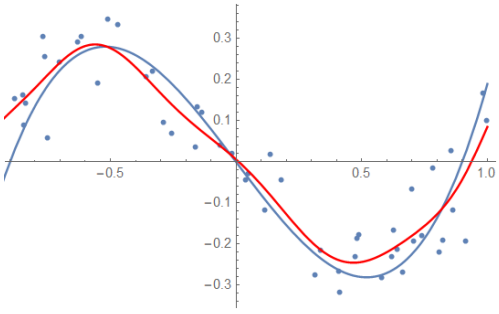


図 3: 青線: 元のモデル / 赤線: 推定されたモデル

以上より，2 次導関数を正則化項に持つルジャンドル多項式回帰の手法を用いることで過学習を抑え，適切な次数のモデルを探索することも可能であると考えられる．

6 実データへの応用

本手法を以下のような実データに応用することを試みた．図 3 は東京の 2017 年 10 月 1 日から 730 日間の日最高気温のデータである．ルジャンドル多項式回帰では，観測点の x 側は $[-1, 1]$ の区間にスケーリングする必要があるため，$x = -1.0$ を 2017 年 10 月 1 日，$x = 1.0$ を 2019 年 9 月 30 日として，区間 $[-1, 1]$ を観測点数 730 で等分割した．

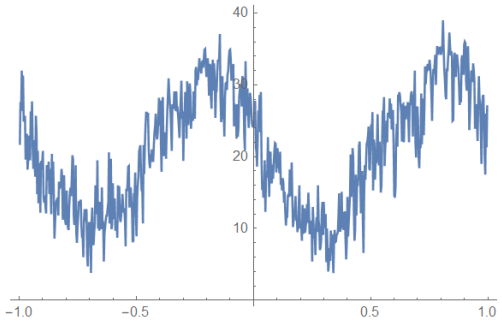


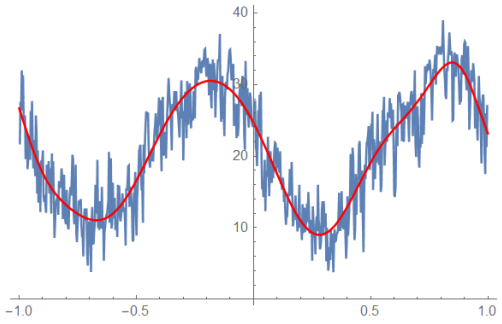
図 4: 東京の日最高気温のデータ

シミュレーションデータを用いたときと同様に λ の推定を行うと次のような結果になった．

λ	平均 2 乗誤差
0.00001	1667.72
0.00010	1658.21
0.00100	1657.41
0.01000	1752.00
0.10000	2766.69

λ の値が 0.00100 前後のとき，平均 2 乗誤差が最小になると考えられる．前後を調べると，$\lambda = 0.00047$ と推定された．

以上より，最適なモデルは以下のようになる．



7 まとめ

本研究では正則化項に 2 次導関数を持つルジャンドル回帰について過学習を抑えられるかの実験を行った．

5-fold クロスバリデーションで適切な正則化パラメータ λ を定めたのち，2 次導関数を正則化項に持つルジャンドル多項式回帰の手法を用いることで過学習を抑え，適切な次数のモデルを探索することも可能であることがわかった．

参考文献

[1] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer (2017).