ランダム行列のリッジ回帰への応用

大谷 有美　(指導教員：吉田 裕亮)

1　はじめに

線形回帰モデルにおいて独立変量 (説明変数) の数が多くなると, 過学習と呼ばれる訓練データ (与えられたデータ) に過度にフィットする状況がしばしば発生する. 与えられたデータへの過度なフィッティングは汎化性が悪く予測モデルとしての精度は良くない.

回帰モデルにおいて過学習を抑える手法のひとつにとして, 回帰係数の L^2 ノルムによる罰則項を付けて, 正則化を行うリッジ回帰がある. リッジ回帰においては, L^2 罰則項の最適な重み λ を定めることは重要である.

本研究では, このパラメータ λ の最適な値の選択を行う手法へのランダム行列の応用を試みる. すなわち, ランダム行列の中でも Wishart 行列に着目し, 1 次の漸近挙動であるモーメントと 2 次の漸近挙動であるモーメントのゆらぎ用いてパラメータ λ の選択に利用する手法の検討を行う.

2　線形回帰モデルとリッジ回帰

線形回帰は, 応答変数 Y を予測変数 X_j　$(j = 1, 2, \ldots, p)$ に基づいて予測する手法で, 下記のようにモデル化される.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

各 β_i は各予測変数 X_i の回帰係数である. また ε は誤差項で独立に Gauss 分布に従うものと仮定される.

通常の最小 2 乗法では, 観測データ $\{y_i, (x_{ij})_{j=1}^p\}_{i=1}^N$ に基づき回帰係数は,

$$S = \sum_{i=1}^N \Bigl(y_i - \sum_{j=1}^p \beta_j x_{ij}\Bigr)^2$$

を最小化をする $\{\beta_j\}_{j=1}^p$ として推定される.

残差平方和 S の最小化のみを評価関数とした場合, 過学習が発生する. そこで回帰係数 $\{\beta_j\}$ の L^2 ノルムを罰則項として加えた

$$\sum_{i=1}^N \Bigl(y_i - \sum_{j=1}^p \beta_j x_{ij}\Bigr)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (\lambda \geq 0)$$

を最小化することにより, 過学習を抑える手法がリッジ (Ridge) 回帰である. ここで $\lambda > 0$ は罰則項を制御するパラメータである.　なお リッジ回帰では, $\{\beta_j\}$ が解析的に求められるという, メリットもある。すなわち, $\boldsymbol{X} = \bigl(x_{ij}\bigr)_{N\times p}$, $\boldsymbol{y} = \bigl(y_j\bigr)_N$ として, 最適解 $\boldsymbol{\beta}$ は,

$$\boldsymbol{\beta} = \bigl(\boldsymbol{X}^T \boldsymbol{X} + \lambda \mathbf{1}\bigr)^{-1} \boldsymbol{X}\boldsymbol{y}$$

と与えられる.

3　Wishart 行列

確率変数を要素とする行列は、一般にランダム行列と呼ばれる。特に、$G_{N,M}$ が $N \times M$ で, 独立に平均 0 分散 1 の標準 Gauss 分布に従う確率変数を要素に持つランダム行列のとき Ginibre 行列と呼ばれる. また, このとき $N \times N$ の共分散行列

$$W_N = \frac{1}{N} G_{N,M}\, {}^t G_{N,M}$$

は Wishart 行列と呼ばれる.

W_N は N 個の固有値を持つが, 固有値漸近分布を考える. ただし, 極限 $N \to \infty$ においては $N \times M$ の Ginibre 行列の漸近縦横比を $M/N \to \gamma$ とする. この極限における固有値漸近分布は, パラメータ γ の Marchenko-Pastur 分布と呼ばれ

$$d\pi_\gamma(x) = \frac{\sqrt{-(x-\gamma_-)(x-\gamma_+)}}{2\pi\gamma x} 1_{[\gamma_-,\gamma_+]}(x)dx + \max\{0, 1-\gamma\}\delta_0(x)$$

で与えられる. ただし, $\gamma_\pm(\gamma) = (1 \pm \sqrt{\gamma})^2$.

この極限分布は Wishart 行列のモーメントの極限から定まるものである。すなわち, Wishart 行列の k 次モーメント m_k

$$m_k = \lim_{N\to\infty} \frac{1}{N} \mathrm{Tr}\bigl(W_N^k\bigr) = \lim_{N\to\infty} \mathrm{tr}\bigl(W_N^k\bigr)$$

が分布 π_γ の k 次モーメントで与えられることである。

3.1　Catalan 数

Marchenko-Pastur 分布において, $\gamma = 1$ の時, つまり Wishart 行列が正方の Ginibre 行列のとき, そのモーメント列は Catalan 数 C_k で与えられる. ここで C_k は,

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)!n!} \quad (n \geq 0)$$
$$= 1, 1, 2, 5, 14, 42, 132, 429, \cdots$$

となる自然数列である.

3.2　Wishart 行列のモーメントのゆらぎ

Wishart 行列のゆらぎモーメントとは, m_p と m_q の共分散の情報で

$$\alpha_{p,q} = \lim_{N\to\infty} \mathrm{Cov}\bigl[\mathrm{Tr}\bigl(W_N^p\bigr), \mathrm{Tr}\bigl(W_N^q\bigr)\bigr]$$

と定義される. この $\alpha_{p,q}$ は 2 次の漸近挙動でモーメントより, 詳細な漸近情報を含む. 例えば $\alpha_{k,k}$ は 極限 $N \to \infty$ における $\mathrm{tr}\bigl(W_N^k - m_k \cdot I\bigr)$ の収束のゆらぎを与えることになる. Wishart 行列の漸近的縦横比 $\gamma = 1$ のときの、ゆらぎモーメントは、以下の式で与えられる.

$$\alpha_{p,q} = 2\frac{2pq}{p+q}\binom{2p-1}{p}\binom{2q-1}{q}$$

6 次までの場合を挙げると、以下のようになる.

$$\begin{pmatrix} 2 & 8 & 30 & 112 & 420 & 1584 \\ 8 & 36 & 144 & 560 & 2160 & 8316 \\ 30 & 144 & 600 & 2400 & 9450 & 36960 \\ 112 & 560 & 2400 & 9800 & 39200 & 155232 \\ 420 & 2160 & 9450 & 39200 & 158760 & 635040 \\ 1584 & 8316 & 36960 & 155232 & 635040 & 2561328 \end{pmatrix}$$

4 本研究での手法

本研究では以下の手順で行う.

1. ある関数にノイズを加えたシュミレーションデータを用いて, 多項式回帰モデルを用い様々なパラメータ λ でリッジ回帰を行う
2. 各 λ においてデータの残差 (各データと推定値の誤差) を求める
3. 残差からランダムリサンプリングをして 50×50 の Ginibre 行列を構成し, $\gamma = 1$ の Wishart 行列を作る
4. この Wishart 行列の 6 次までのモーメント列を求める
5. 手順 3.4 を多数 (N=100,000) 回繰り返す
6. N サンプルの平均によりモーメントを求め, Catalan 数と比較
7. N サンプルの共分散行列を求め, 6 次までのゆらぎのモーメントと比較する

1〜7 までの手順で最適な λ の値を選択することが可能かを実験する. また, モーメント (1 次漸近挙動) に大差がない場合でもゆらぎモーメント (2 次漸近挙動) により判定可能であることをみる.

5 実験例

関数 $\sin 4\pi x (-0.5 \leq x \leq 0.5)$ の標本点にノイズを加えた 31 個のシミュレーションデータを用いる. それらのデータに対して 31 次の多項式回帰モデルを用いてリッジ回帰を行う.

$\lambda = 10^{-27}$ の場合, 罰則項がほとんど 0 になり通常の回帰を行うことになり図 1 のように過学習を起こす.

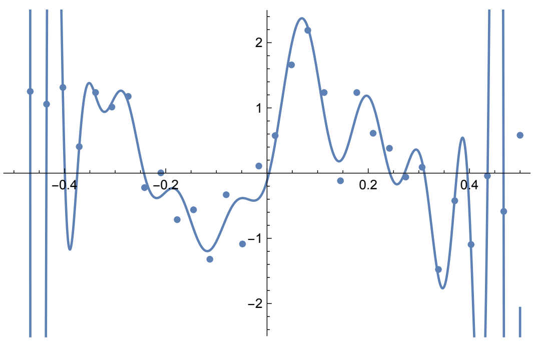

図 1: $\lambda = 10^{-27}$ の場合

6 次のモーメント列は,
$\{1.056, 2.840, 11.743, 63.519, 401.428, 2759.873\}$
になり, Catalan 数とは明らかにかけ離れている.

$\lambda = 0.1$ の場合, 罰則項の影響が強すぎるため図 2 のように過度に滑らかな曲線を描いてしまう.

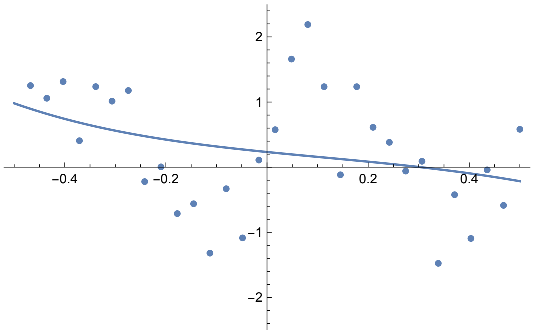

図 2: $\lambda = 0.1$ の場合

この時 6 次までのモーメント列は,
$\{0.968, 1.885, 4.603, 12.624, 37.196, 115.129\}$ になる.

様々な λ で比較し最適と推定されるのは, $\lambda = 10^{-12}$ であり, 6 次までのモーメント列は,
$\{0.968, 1.901, 4.695, 13.046, 38.997, 122.575\}$
ゆらぎモーメントは以下のようになる.

$$\begin{pmatrix} 2.3059 & 9.0594 & 33.529 & 124.15 & 463.71 & 1748.5 \\ 9.0594 & 39.312 & 154.13 & 592.67 & 2274.1 & 8751.8 \\ 33.529 & 154.13 & 628.07 & 2483.1 & 9730.7 & 38080. \\ 124.15 & 592.67 & 2483.1 & 10028. & 39968. & 158602 \\ 463.71 & 2274.1 & 9730.7 & 39968. & 161542 & 648635 \\ 1748.5 & 8751.8 & 38080. & 158602 & 648635 & 2631160 \end{pmatrix}$$

またリッジ回帰曲線は図 3 のようになる.

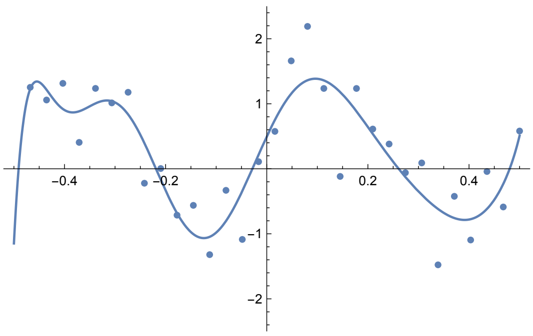

図 3: $\lambda = 10^{-12}$ の場合

6 まとめ

本研究では, リッジ回帰におけるパラメータ λ の最適な値の選択にランダム行列のひとつである Wishart 行列に着目し, 1 次漸近挙動であるモーメント列と 2 次漸近挙動であるゆらぎモーメントを利用する手法を提案した.

今回行った実験より, モーメント列とゆらぎモーメントがパラメータ λ の値の選択に利用可能であること, そしてモーメント列に大差がない場合, ゆらぎモーメントが利用可能であることが分かった.

今後の課題としては, 今回扱った関数以外の他の関数でも更なる実験を行い, また Cross-Validation との比較検討も行う必要がある.