

データ解析入門 18 <スパース推定の適用例>

キーワード：正則化法、スパース推定、連結 Lasso、一般化 Lasso、ADMM

はじめに

前稿¹⁾では、線形回帰モデルにスパース性を仮定することでいくつかの回帰係数が0に推定されることを確認しました。本稿では、系列データ、グラフデータ、および画像データにスパース推定を適用し、各スパースモデリングの特徴を説明します。

連結 Lasso

前稿¹⁾では、Lasso 回帰などのスパース推定法について説明しました。Lasso 回帰の損失関数を再掲します[式(1)]。N、 $\mathbf{y}$ 、 $\mathbf{X}$ は、それぞれデータ数、目的変数ベクトル、データ行列です。

$$L_{\lambda}(\boldsymbol{\beta}) = \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \quad (1)$$

Lasso 回帰では、回帰係数 $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ にスパース性を仮定しています。式(1)の損失関数は、第1項がデータとの整合性、第2項が事前知識を反映した正則化項というように解釈できます。問題に応じて正則化項の定式化などを工夫することで、柔軟なスパース推定法が実行可能になります。

連結 Lasso と呼ばれる手法では、式(1)の正則化項に隣接する回帰係数の差分の絶対値を追加します[式(2)]。

$$L_{\lambda}(\boldsymbol{\beta}) = \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda' \|\boldsymbol{\beta}\|_1 + \lambda \sum_{j=2}^p |\beta_j - \beta_{j-1}| \quad (2)$$

この正則化項により、隣接する回帰係数の値が同じ値に推定されやすくなります。なお、 λ および λ' は、いずれも正則化の強さを制御するハイパーパラメータです。

図1に、比較ゲノムハイブリタイゼーション(CGH)法により計測された、ある病気の患者と健康な人のゲノムのコピー数を対数比で示します(灰点)。CGH データは R 言語の cghFLasso パッケージ²⁾から取得しました。このデータに連結 Lasso を適用すると、ステップ状の推定結果が得られ、データの特徴が強調されます(青線)。連結 Lasso では、式(2)の行列 $\mathbf{X}$ を単位行列としています。連結 Lasso により、対数比が大きくなる特徴的な遺伝子

領域が明瞭になっています。 λ を大きくすると、平滑化の効果が大きくなります。連結 Lasso は、隣接するデータ間の相関が強い時系列データなどに適用されます。

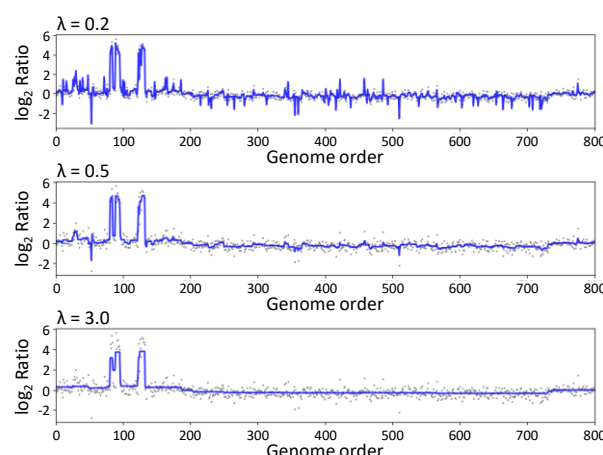


図1 連結 Lasso の適用結果

一般化 Lasso

連結 Lasso の損失関数[式(2)]の正則化項は、行列 $\tilde{D}$ を用いて以下のように変形できます³⁾。

$$\begin{aligned} & \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left(\frac{\lambda'}{\lambda} \|\boldsymbol{\beta}\|_1 + \sum_{j=2}^p |\beta_j - \beta_{j-1}| \right) \\ &= \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\tilde{D}\boldsymbol{\beta}\|_1 \end{aligned} \quad (3)$$

$$\tilde{D} = ((\lambda'/\lambda)I, D_{1d})^T$$

$$D_{1d} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

ここで、 I は $p \times p$ 型の単位行列、 D_{1d} は $(p-1) \times p$ 型の1次差分行列です。

この最小化問題はさらに一般化でき、式(3)第2項の行列 $\tilde{D}$ を所与の行列 D に置き換えた最小化問題は一般化 Lasso と呼ばれます[式(4)]。

$$L_{\lambda}(\boldsymbol{\beta}) = \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|D\boldsymbol{\beta}\|_1 \quad (4)$$

行列 D を単位行列とすると、Lasso 回帰の損失関

数[式(1)]となることがわかります。

データ点のつながりを表す隣接行列を用いることで、グラフデータに連結 Lasso を適用することができます(図 2)。具体的には、各都道府県をグラフのノードとし、隣接関係をエッジで表現します。本稿では、都道府県ごとの人口 10 万人当たりのコンビニ店舗数⁴⁾に連結 Lasso を適用しました。特に、 $\lambda = 2$ とした場合、同じ色の隣接した領域が拡大しており、比較的シンプルな色分けが得られます。

続いて、画像データに 2 次元連結 Lasso を適用します。本解析では、縦および横方向に隣接するピクセルの画素値の差分に L1 正則化を課しました。このようにすることで、各ピクセルにおいて周辺の画素値に近づくような推定が行われます。ここで、人工的にノイズ(ガウス乱数)を付与した画像[図 3(b)]に 2 次元連結 Lasso を適用した結果を示します。2 次元連結 Lasso により、図 3(b)のノイズが低減されていることがわかります。

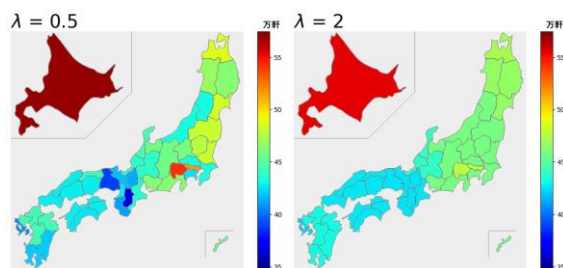


図 2 グラフデータへの連結 Lasso 適用結果

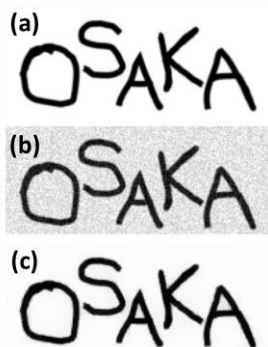


図 3 (a) 元画像、(b) ノイズを加えた画像、(c) 2 次元連結 Lasso を適用した画像

交互方向乗数法 (ADMM)

本稿では、連結 Lasso の最小化問題に交互方向乗数法 (ADMM) を用いました。ADMM では、式(4)の最小化問題を次の等式制約付き最小化問題へと置き換えます。 $D\beta = \gamma$ で条件づけられた変数 $\gamma = (\gamma_1, \dots, \gamma_p)^T$ を導入することで制約条件に対応した最適化を行います。

$$\min_{\beta, \gamma} \left\{ \frac{1}{2N} \|y - X\beta\|_2^2 + \lambda \|\gamma\|_1 \right\} \quad \text{subject to } D\beta = \gamma$$

上式より、制約付き最小化問題などに用いられる拡張ラグランジアンは以下ようになります。

$$\frac{1}{2N} \|y - X\beta\|_2^2 + \lambda \|\gamma\|_1 + u^T (D\beta - \gamma) + \frac{\rho}{2} \|D\beta - \gamma\|_2^2$$

ここで、 $u = (u_1, \dots, u_p)^T$ はラグランジュ乗数、 $\rho (> 0)$ は調整パラメータです。ADMM では、拡張ラグランジアンをもとに、 β という単一のパラメータではなく、3 つのパラメータ β, γ, u を交互に更新することで最小化を実行します。更新する 3 つのパラメータは互いに依存しておらず、式(4)の最小化は単純な下記ステップで実行できます。

$$\beta^{(t+1)} = (X^T X + N\rho D^T D)^{-1} \left\{ X^T y + N\rho D^T \left(\gamma^{(t)} - \frac{1}{\rho} u^{(t)} \right) \right\}$$

$$\gamma_j^{(t+1)} = S \left((D\beta^{(t+1)})_j + \frac{1}{\rho} u_j^{(t)}, \frac{\lambda}{\rho} \right) \quad (j = 1, \dots, p)$$

$$u^{(t+1)} = u^{(t+1)} + \rho (D\beta^{(t+1)} - \gamma^{(t+1)})$$

ここで、 $S(\cdot, \cdot)$ は軟閾値作用素と呼ばれるもので、変数 x の絶対値を縮小させます(図 4)。なお、 $(D\beta^{(t+1)})_j$ はベクトル $D\beta^{(t+1)}$ の第 j 成分です。

上記数式の形に最適化問題を定式化できれば ADMM は適用可能です。そのため、ADMM は、様々な制約条件に適用できる汎用性の高い最適化アルゴリズムであるといえます。

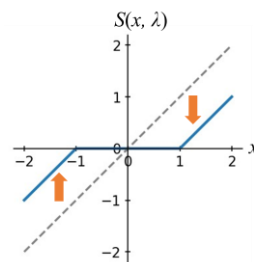


図 4 軟閾値作用素 ($\lambda = 1$)

おわりに

本稿で紹介した手法はよく知られたスパース推定法です。R 言語などでスパース推定のための多くのパッケージが公開されており、連結 Lasso などを実行することができます。

参考文献

- 1) 永廣卓哉: データ解析入門 17 <Lasso・Ridge・Elastic Net>、ORIST テクニカルシート、No. 23-06 (2023)
- 2) <https://cran.r-project.org/src/contrib/Archive/cghF/Lasso/> (accessed on November 18th, 2022)
- 3) 川野秀一・松井秀俊・廣瀬慧、スパース推定法による統計モデリング、共立出版(2018)
- 4) <https://todo-ran.com/t/kiji/10328> (accessed on November 14th, 2022)

発行日 2023 年 12 月 1 日

作成者 高分子機能材料研究部 生活環境材料研究室 永廣卓哉

Phone: 0725-51-2611 E-mail: ehirot@orist.jp