双対平坦空間の情報幾何に基づく最小角回帰法の拡張

数理情報学専攻 66221 廣瀬善大
指導教員 駒木文保 准教授

1 はじめに

線形回帰問題は, 興味のある反応を説明変数の線形結合で表すものであり, 手元に得られたデータから線形結合のパラメータである係数を求める問題である. このとき, 手元のデータだけでなく将来観測されるデータを予測するという観点からパラメータを求めることを考える [4]. この問題を解くための手法として, アルゴリズムによって定義された手法である最小角回帰法（Least Angle Regression, LAR）が提案されている [3]. この手法はユークリッド空間における角の二等分線に基づいたアルゴリズムと考えることができ, 自動的に説明変数を逐次選択する働きをもつ. 本発表では, 双対平坦空間における情報幾何の方法を用いることにより, ユークリッド空間における角の二等分線を拡張した曲線を考え, その曲線を用いてより一般的な立場で LAR を拡張する. その際, 推定量の更新を逆向きに行うように改良する. 拡張された LAR は一般化線形回帰問題に対するパラメータ推定手法であり, 説明変数を選択する働きをもつアルゴリズムになる.

2 最小角回帰法 (LAR) の拡張

2.1 情報幾何

双対平坦空間は, ユークリッド空間を一般化した空間である [1, 2]. 特に, e-アファイン座標, m-アファイン座標という 2 つの便利な座標と, それぞれに対応するポテンシャルと呼ばれる凸関数をとることができる. そして, 測地線, ダイバージェンスは, それぞれユークリッド空間における直線, 距離を双対平坦空間において一般化したものであり, この 2 つを用いて LAR を双対平坦空間において拡張する. 特に, ダイバージェンス $D(\cdot,\cdot)$ に関する拡張ピタゴラスの定理を利用する.

2.2 角の二等分線の拡張

p 次元双対平坦空間において, p 個の $p-1$ 次元 e-平坦空間 $S_i = \{\theta_i = 0\}\,(i = 1, 2, \ldots, p)$ と点 $\widehat{\boldsymbol{\theta}}$ を考える. ただし, ここで p は任意の自然数として固定する. また $\widehat{\boldsymbol{\theta}}$ は $S_i\,(i = 1, 2, \ldots, p)$ 上にはないものとする.

まず, $\widehat{\boldsymbol{\theta}}$ から $S_i\,(i = 1, 2, \ldots, p)$ への m-射影を $\bar{\boldsymbol{\theta}}_i$ とおき, $\widehat{\boldsymbol{\theta}}$ から $\bar{\boldsymbol{\theta}}_i$ への m-測地線を l_i とおく. そして, $t > 0$ を任意に固定して, $D(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_i(t)) = t$ となる点 $\widetilde{\boldsymbol{\theta}}_i(t) \in l_i$ をとる. ここで, $\widetilde{\boldsymbol{\theta}}_i(t)$ において l_i と直交する e-平坦空間を $S_i'(t)$ とおくと, $S_i'(t) = \{\theta_i = \alpha_i(t)\,(= \text{const.})\}$ と表すことができるので, $S_i'(t)\,(i = 1, 2, \ldots, p)$ の交点 $\widetilde{\boldsymbol{\theta}}(t)$ は $\widetilde{\theta}_i(t) = \alpha_i(t)\,(i = 1, 2, \ldots, p)$ として求まる.

今, 固定していた $t \geq 0$ を変化させると, それに伴なって点 $\widetilde{\boldsymbol{\theta}}(t)$ がある曲線 l を描く. このとき曲線 l は, 以下の意味でユークリッド空間における角の二等分線を拡張したものになっている. すなわち, l 上の任意の点 $\boldsymbol{\theta} \in l$ はある $t_0 \geq 0$ に対して $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}(t_0)$ と表されていて, m-測地線 l_i 上に点 $\widetilde{\boldsymbol{\theta}}_i(t_0)$ を考えることができる. このとき, $\widetilde{\boldsymbol{\theta}}_i(t)$ の定義から,

$$D(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_1(t_0)) = D(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_2(t_0)) = \cdots = D(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_p(t_0)) = t_0$$

である. また, 拡張ピタゴラスの定理により,

$$\begin{aligned} D(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= D(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_1(t_0)) + D(\widetilde{\boldsymbol{\theta}}_1(t_0), \boldsymbol{\theta}) \\ &= D(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_2(t_0)) + D(\widetilde{\boldsymbol{\theta}}_2(t_0), \boldsymbol{\theta}) \\ &= \ldots \\ &= D(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}_p(t_0)) + D(\widetilde{\boldsymbol{\theta}}_p(t_0), \boldsymbol{\theta}) \end{aligned}$$

が成り立つ. よって, 双対平坦空間におけるある種の等距離性である

$$D(\widetilde{\boldsymbol{\theta}}_1(t_0), \boldsymbol{\theta}) = D(\widetilde{\boldsymbol{\theta}}_2(t_0), \boldsymbol{\theta}) = \cdots = D(\widetilde{\boldsymbol{\theta}}_p(t_0), \boldsymbol{\theta})$$

が成り立つことが分かる. l 上の任意の点 $\boldsymbol{\theta} \in l$ においてこの等距離性が成り立つので, 曲線 l はユークリッド空間における角の二等分線を双対平坦空間において拡張したものだと考えることができる.

2.3 問題設定

一般化線形回帰モデル

$$g(\boldsymbol{\mu}) = X\boldsymbol{\theta}$$

を考える. ただし, μ はモデルの下での反応 $\boldsymbol{y}$ の期待値 $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y})$ である. また, $\boldsymbol{y}$ は反応で n ベクトル, X は計画行列で $n \times p$ 行列, $\boldsymbol{\theta}$ は係数ベクトルで p ベクトルであり, 計画行列 X の各列は標準化されているものとする.

$$\sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad (j = 1, 2, \ldots, p).$$

このとき, 考える問題は指数型分布族における分布の選択となる. 指数型分布族の自然パラメータを $\boldsymbol{\xi}$ で表すと $\boldsymbol{\xi} = X\boldsymbol{\theta}$ であり, これは指数型分布族のなす双対平坦空間における e-アファイン座標になる. また, $\boldsymbol{\xi}$ に対応する双対な座標である m-アファイン座標は上の $\boldsymbol{\mu}$ であり期待値パラメータと呼ばれる. よって, 係数ベクトル $\boldsymbol{\theta}$ を求め, 自然パラメータを指定することにより指数型分布を 1 つ選択することになる.

ところで, 指数型分布族のなす空間が双対平坦空間であるから, アファイン変換により変換された係数ベクトル $\boldsymbol{\theta}$ の空間も双対平坦空間になっている. そして, $\boldsymbol{\theta}$ と双対な座標 $\boldsymbol{\eta}$ をとることができる. 以下では, この係数ベクトルのなす双対平坦空間において推定量の更新を行う.

2.4 アルゴリズム

1. $I = \{1, 2, \ldots, p\}$, $\widehat{\boldsymbol{\theta}}_0 := \widehat{\boldsymbol{\theta}}_{MLE}$, $k = 0$ とする.
2. $j \notin I$ に対して $\theta_j = 0$ とおく. さらに, $i \in I$ なる i それぞれについて $\theta_i = 0$ とおいて得られる $p-k$ 個の $p-k-1$ 次元 e-平坦空間を $S_{i_1}, S_{i_2}, \ldots, S_{i_{p-k}}$ と表し ($i_1, i_2, \ldots, i_{p-k} \in I$ である), 点 $\widehat{\boldsymbol{\theta}}_k$ から $S_i\,(i \in I)$ への m-射影を $\bar{\boldsymbol{\theta}}_i$ とおく. $i^* = \arg\min_{i \in I} D(\widehat{\boldsymbol{\theta}}_k, \bar{\boldsymbol{\theta}}_i)$ として i^* を定め, $t = D(\widehat{\boldsymbol{\theta}}_k, \bar{\boldsymbol{\theta}}_{i^*})$ とおく.
3. $\widehat{\boldsymbol{\theta}}_k$ から $\bar{\boldsymbol{\theta}}_i\,(i \in I)$ への m-測地線を l_i と表し, l_i 上で $D(\widehat{\boldsymbol{\theta}}_k, \widetilde{\boldsymbol{\theta}}_i) = t$ となる点 $\widetilde{\boldsymbol{\theta}}_i$ をとる.
4. $\widetilde{\boldsymbol{\theta}}_i$ において l_i と直交する e-平坦な空間 S_i' は $\{\theta_j = 0\,(j \notin I), \theta_i = \alpha_i (= \text{const.})\}$ と表されるので, $S_i'\,(i \in I)$ の交点は $\theta_j = 0\,(j \notin I, j = i^*), \theta_i = \alpha_i\,(i \in I, i \neq i^*)$ となる. この交点を $\widehat{\boldsymbol{\theta}}_{k+1}$ とおく.
5. $k = p-1$ ならばステップ 6 へ進む. $k < p-1$ ならば, $k := k+1$, $I := I \cup \{i^*\}$ としてステップ 2 へ進む.
6. $\widehat{\boldsymbol{\theta}}_p := \mathbf{0}$ として, アルゴリズム終了.

3 データ解析

この節では, 心臓病に関するデータ [4] に対するロジスティック回帰に拡張版 LAR を適用した結果を示す (図 1). データは反応 y と 9 つの説明変数 $x_1, x_2, \ldots, x_9$ からなり, $n = 462$ 人分のデータである. 反応 y は chd (心筋梗塞の有無) であり, 全体の 5.1% が心筋梗塞をもっていた. 説明変数は, x_1: sbp (最大血圧), x_2: tobacco (タバコ), x_3: ldl (悪玉コレステロール), x_4: adiposity (脂肪過多症), x_5: famhist (家族に心筋梗塞患者がいるかどうか), x_6: typea (A 型行動), x_7: obesity (肥満), x_8: alcohol (アルコール), x_9: age (年齢) である. この心臓病データに対するロジスティック回帰について拡張版 LAR を適用すると, 説明変数 $x_1, x_2, \ldots, x_9$ のうちで推定量を構成しなくなる順番は $x_8, x_4, x_3, x_1, x_7, x_6, x_2, x_9, x_5$ であった.

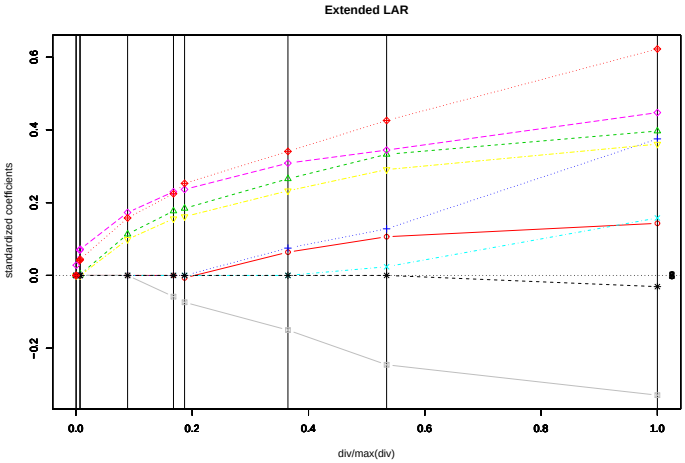

図 1. 心臓病データのロジスティック回帰に対する拡張版 LAR

4 おわりに

ユークリッド空間における角の二等分線を双対平坦空間において拡張し, その曲線を用いて LAR の拡張を行った. その際, 推定量の更新の向きを逆にした. また, 実際のデータを解析した.

今後の課題は, まず, 出力される複数の推定量の中からどの推定量を選ぶべきかという問題が挙げられる. また, 説明変数に線型独立性を仮定できないような場合, 特に $n \times p$ 計画行列について $n < p$ であるような場合のパラメータ推定・説明変数選択の方法を考えることも課題である.

参考文献

[1] S. Amari: *Differential-Geometical Methods in Statistics*. Springer Lecture Notes in Statistics 28, 1985.
[2] S. Amari, H. Nagaoka: Methods of Information Geometry. Translations of Mathematical Monographs, Vol. 191, Oxford University Press, 2000.
[3] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani: Least Angle Regression (with discussion). *The Annals of Statistics*, vol. 32 (2004), pp. 407–499.
[4] T. Hastie, R. Tibshirani, J. Friedman: *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2001.