数理解析研究所講究録
第1916巻 2014年 103-122

一般化線形回帰問題と情報幾何

東京大学・情報理工学系研究科　廣瀬 善大
Yoshihiro Hirose
Department of Mathematical Informatics,
Graduate School of Information Science and Technology,
The University of Tokyo

概要

本稿では，統計学における一般化線形回帰と情報幾何との関わりについて解説する．特に，正規線形回帰におけるパラメータ推定手法のひとつである Least Angle Regression (LARS) アルゴリズムと正則化を軸に話を進める．一般化線形回帰の問題設定と情報幾何の関連を述べ，情報幾何に基づくいくつかのパラメータ推定手法を紹介する．文献 [8] において提案された推定手法 bisector regression，文献 [4] の DGLARS (Differential Geometric LARS) はいずれも情報幾何と関連の深いパラメータ推定手法である．

1　導入

本稿では，統計学における一般化線形回帰と情報幾何との関わりについて解説する．特に，正規線形回帰におけるパラメータ推定手法のひとつである Least Angle Regression (LARS, [5]) アルゴリズムと正則化を軸に話を進める．

統計学における基本的な問題のひとつに線形回帰問題がある．興味のある変数（反応変数）をその他の変数（説明変数）で表現するのが回帰問題であり，特に反応変数が説明変数の1次式で表されることを仮定したのが線形回帰問題である．反応変数と説明変数の組がいくつか観測されたもとで，パラメータである回帰係数（説明変数の線形結合の係数）を推定する．線形回帰は数学的に扱いやすく，また広く使われている方法であり，理論的にも応用上も重要な問題である．確率分布としては分散既知の正規分布を利用しており，ユークリッド幾何を用いて説明することができる．たとえば，有名な推定手法のひとつである最小二乗法は最尤法と一致し，得られる推定量は観測された反応変数ベクトルを説明変数ベクトルの張る線形部分空間に直交射影したものである．

線形回帰問題におけるパラメータ推定手法のひとつとして LARS がある．LARS アルゴリズムは，ユークリッド空間内で推定量を表す点を移動させることにより複数の推定値を出力するアルゴリズムである．推定量の点を移動させる際に角の二等分線を利用している．ユークリッド幾何における内積（あるいは角度）は，統計学における相関に対応する

ものであり，LARSは反応変数と各説明変数の間の相関（内積あるいは角度）に基づいてパラメータを推定する．相関に基づいてパラメータ推定を行う手法はそれまでにもあったが，LARSはその一部を含み，さらに計算が容易に行えるアルゴリズムである．

LARSは正則化の一種であるLASSO [14, 15] と関連があることが知られている．正則化とは，最尤法のように尤度だけを最大化するのではなく，得られる推定値に制約を加えて尤度を最大化する手法である．LASSOは正則化の代表的な手法であり，近年，関連した多くの研究がなされてきた．LASSOはl_1制約付きの対数尤度最大化の問題であり，特に得られる推定値が疎になるという特徴がある．ここで推定値が疎であるとは，パラメータの一部の成分が厳密に0になることを意味する．つまり，LASSOはパラメータ推定と同時に説明変数の選択を行う手法であり，LARSも同様の性質をもつ．

一般化線形回帰問題とは，線形回帰問題において仮定されていた仮定，つまり反応変数が説明変数の1次式で表現されるという仮定を緩めたものである．一般化線形回帰問題においては，反応変数をある関数で変換して得られる変数が説明変数の1次式で表現されると仮定する．反応変数の変換は考える問題ごとに異なり，変換に用いられる関数はリンク関数と呼ばれる．線形回帰問題では，リンク関数として恒等関数を用いていると考えることができ，一般化線形回帰問題が線形回帰問題を含むより広い問題であることが分かる．また，線形回帰問題はユークリッド幾何により説明することができたが，一般化線形回帰問題を説明するためにはより一般的な微分幾何学的アプローチとして情報幾何 [1, 2, 11] が必要になる．線形回帰では分散既知の正規分布を考えていたのに対し，一般化線形回帰では指数型分布族を考える必要があるからである．

情報幾何はユークリッド幾何の一般化のひとつである．確率分布のなす多様体上でフィッシャー情報量を計量として用いる．通常のリーマン幾何ではレビ–チビタ接続が利用されることが多いが，情報幾何においてはパラメータ付けられたアファイン接続の族が重要な役割を担う．接続が与えられると測地線や平坦さを扱うことができるようになる．特に，±1-接続と呼ばれる接続はたがいに双対で，指数型分布族の多様体を平坦にする自然な接続であり，指数型分布族に関係する様々な場面で利用される．0-接続がレビ–チビタ接続に対応する．1-接続に関して平坦な空間は-1-接続の意味でも平坦であることが知られており，その逆も成り立つ．そのため±1-接続に関して平坦な空間は双対平坦空間と呼ばれる．指数型分布族のなす多様体は±1-平坦であることが知られており，±1-接続に関する双対構造をもとに統計的推定を行うことができる．なお，統計学において自然パラメータと期待値パラメータとして知られているパラメータが幾何学的にはそれぞれの接続に関するアファイン座標系に対応している．情報幾何は，たとえば統計学において高次の漸近論などで有用であり，さらに制御論や最適化などの分野でも有用である．

一般化線形回帰に対して情報幾何を利用したパラメータ推定手法がいくつか提案されている．Bisector regression [8] は，LARS が推定量の移動で利用していた角の二等分線を一般化し，その曲線上で推定量を移動させることによりいくつかの推定値を出力するアルゴリズムである．DGLARS (Differential Geometric LARS, [4]) も微分幾何的な手法で LARS を拡張したものである．これらの手法の簡単な解説を行うことが本稿の目的のひとつである．

本稿の構成は以下の通りである．第 2 章において線形回帰問題と一般化線形回帰問題を紹介する．特に，一般化線形回帰問題を情報幾何的視点から解説する．第 2 章までが主に回帰問題と情報幾何との関わりを述べた部分である．第 3 章以降は主に情報幾何を利用したパラメータ推定手法の紹介である．第 3 章では線形回帰におけるパラメータ推定手法を紹介する．主に LARS アルゴリズムについて簡単に説明する．第 4 章は情報幾何を利用したパラメータ推定手法の紹介である．線形回帰のパラメータ推定手法である LARS アルゴリズムを一般化線形回帰の場合にまで拡張させた bisector regression, DGLARS を紹介する．本稿のまとめを第 5 章で行う．

2 回帰問題

線形回帰問題と一般化線形回帰問題について説明する．2.1 節では線形回帰の，2.2 節では一般化線形回帰の解説をする．2.3 節において，本稿の目的のひとつである，一般化線形回帰と情報幾何との関わりを説明する．

以下では，$n > d$ とし，$\{y_a, x^a = (x_1^a, x_2^a, \ldots, x_d^a)\}_{a=1,2,\ldots,n}$ が観測されているものとする．説明変数を行列として表した計画行列 X は，$X = (x_i^a)_{1 \le a \le n, 1 \le i \le d} = (x_1, x_2, \ldots, x_d)$ で与えられ，$n \times d$ 行列である．ただし，説明変数ベクトルは $x_i = (x_i^1, x_i^2, \ldots, x_i^n)^\top (i = 1, 2, \ldots, d)$ である．$\mathbf{1}$ を n 個の 1 をもつベクトル，すなわち $\mathbf{1} = (1, 1, \ldots, 1)^\top$ とし，$n \times (d+1)$ 行列 $\tilde{X}$ を $\tilde{X} = (\mathbf{1}|X)$ と定義する．$y = (y_1, y_2, \ldots, y_n)^\top$ を $x_i (i = 1, 2, \ldots, d)$ で表現するのが回帰問題である．

2.1 線形回帰

線形回帰では反応変数を各説明変数の線形結合で説明することを仮定する．特に，誤差の確率分布として分散既知の正規分布を仮定した正規線形回帰は扱いやすさなどにより広く利用されている．線形回帰では以下が仮定されている:

$$\mathrm{E}[y] = \tilde{X}\tilde{\theta}.$$

ただし，$\theta = (\theta^1, \theta^2, \ldots, \theta^p)^\top \in \mathbf{R}^d$，$\tilde{\theta} = (\theta^0, \theta^\top)^\top \in \mathbf{R}^{d+1}$ である．パラメータである回帰係数 $\tilde{\theta}$（あるいは θ）を推定することが目的である．

一般化線形回帰を情報幾何で扱うための準備として，正規線形回帰を確率分布の幾何（ここではユークリッド幾何）の視点から説明する．分散既知（簡単のため分散 1）で平均 $\mu \in \mathbf{R}^n$ の n 変量正規分布の確率密度関数は

$$f(y|\,\mu) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{(y-\mu)^\top (y-\mu)}{2} \right\}$$

である．反応変数の各成分 y_a は，各説明変数が与えられた下で独立に正規分布にしたがう．$\mu = \mathrm{E}[y] = \tilde{X}\tilde{\theta}$ であるから，反応変数 y の確率密度関数は

$$f(y|\,\tilde{\theta}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{(y-\tilde{X}\tilde{\theta})^\top (y-\tilde{X}\tilde{\theta})}{2} \right\}$$

である．

n 変量正規分布の間の離れ具合をユークリッド距離（の二乗）で測ることにする．これは推定において対数尤度 $\log f(y|\,\mu)$ で距離を測ることに対応する．パラメータ推定法として有名な最小二乗法は，正規分布の場合には最尤法と一致することが分かる．$\mathbf{R}^n$ において，計画行列 $\tilde{X}$ の各列ベクトルの張る線形部分空間に y を直交射影して得られるのが最尤推定量（最小二乗推定量）である．対数尤度に基づくパラメータ推定はユークリッド幾何で説明することができ，さらに次節以降で説明される情報幾何の特別な場合に相当する．

最後に，切片の直交化について簡単に説明する．線形回帰においては計画行列 X の正規化と呼ばれる前処理を仮定することが多い．具体的には，

$$\frac{1}{n}\sum_{a=1}^{n}(x_i^a)^2 = 1, \quad \sum_{a=1}^{n} x_i^a = 0 \qquad (i = 1, 2, \ldots, d) \tag{1}$$

を仮定して切片 θ^0 を推定しない問題設定を扱う．与えられた計画行列 X を式 (1) を満たすように線形変換することを考える．式 (1) の第 2 式より，ベクトル $\mathbf{1}$ と各列ベクトルは直交することになる．以上のように切片を直交化することで，切片 θ^0 の分だけ次元を下げて推定を行うことができる．切片 θ^0 の推定はパラメータ θ とは独立に行うことができる．

2.2 一般化線形回帰

一般化線形回帰におけるパラメータ推定は，対応する指数型分布族の中から確率分布をひとつ選択する問題である ([12])．指数型分布族とは，確率密度関数（あるいは確率関

数）が

$$f(y|\xi) = \exp\left(y^\top \xi - \psi(\xi)\right)$$

と表される確率分布の族のことである．ただし，$\xi = (\xi^1, \xi^2, \ldots, \xi^n)^\top$ である．$\psi(\cdot)$ は y について積分して 1 になるよう正規化するための関数で凸関数である．パラメータ ξ は自然パラメータと呼ばれ，$\psi(\cdot)$ は ξ のポテンシャル関数と呼ばれる．自然パラメータ ξ を指定することで確率分布がひとつに決まる．

一般化線形回帰問題では，$\xi = \tilde{X}\tilde{\theta}$ を仮定する．ただし，$\theta = (\theta^1, \theta^2, \ldots, \theta^p)^\top \in \mathbf{R}^d$, $\tilde{\theta} = (\theta^0, \theta^\top)^\top \in \mathbf{R}^{d+1}$ である．パラメータである回帰係数 $\tilde{\theta}$（あるいは θ）を推定することが目的である．y の期待値 $\mu = (\mu_1, \ldots, \mu_n)^\top = (\mathrm{E}[y_1], \ldots, \mathrm{E}[y_n])^\top$ は期待値パラメータと呼ばれ，期待値パラメータを指定することで確率分布がひとつに決まる．自然パラメータ ξ と期待値パラメータ μ の間にはリンク関数と呼ばれる関数 g により $\xi = g(\mu)$ という関係がある．

一般化線形回帰の例としてロジスティック回帰を考える．考える指数型分布族は

$$f(y|\xi) = \prod_{a=1}^{n} \frac{\exp(y_a \xi^a)}{1 + \exp \xi^a} = \exp\left(\sum_{a=1}^{n} y_a \xi^a - \psi(\xi)\right)$$

である．ただし，$\xi = (\xi^1, \xi^2, \ldots, \xi^n)^\top$ は自然パラメータであり，$y \in \{0,1\}^n$ である．ξ に関するポテンシャル関数は $\psi(\xi) = \sum_{a=1}^{n} \log(1 + \exp \xi^a)$ である．期待値パラメータ μ は $\mu_a = \mathrm{E}[y_a] = \exp \xi^a / (1 + \exp \xi^a)\ (a = 1, 2, \ldots, n)$ で与えられる．リンク関数は $g(\mu) = \log \frac{\mu}{1-\mu}$ である．

2.3 一般化線形回帰と情報幾何

本節では，一般化線形回帰について情報幾何の視点から解説する．そのために，まず必要な情報幾何の説明を行う．双対平坦空間の情報幾何はユークリッド幾何の一般化のひとつである．双対平坦空間においては，ユークリッド空間と同様に直線や距離に対応する量を扱うことができる．そのためユークリッド幾何により記述されていた LARS アルゴリズムをもとにして，情報幾何を利用した推定手法がいくつか提案されている．情報幾何の詳細については，文献 [1], [2], [11] などを参照のこと．

$\Xi \subseteq \mathbf{R}^n$ をパラメータ空間とし，指数型分布族のなす空間 $S = \{f(\cdot|\xi) | \xi \in \Xi\}$ を考える:

$$f(y|\xi) = \exp\left(\sum_{a=1}^{n} y_a \xi^a - \psi(\xi)\right).$$

ここで，$\partial_a = \partial/\partial\xi^a$, $\partial^a = \partial/\partial\mu_a$ と略記することにする．対数尤度を $l(\xi, y) =$

$\log f(y|\xi)$，確率密度関数の対数をとった確率変数を $l(\xi, Y) = \log f(Y|\xi)$ と書くことにする．接ベクトル ∂_a は確率変数 $\partial_a l(\xi, Y)$ と同一視される．特に，$\mathrm{E}[\partial_a l(\xi, Y)] = 0$ が成り立つ．計量としてフィッシャー情報

$$g_{ab} = \mathrm{E}\left[(\partial_a l(\xi, Y))\,(\partial_b l(\xi, Y))\right]$$

を用いる．α-接続と呼ばれる接続の族を考え，特に $\alpha = \pm 1$ に注目する．1-接続と -1-接続はそれぞれ e-接続，m-接続とも呼ばれ，互いに双対な構造をもつ．自然パラメータ ξ に関する α-接続の係数 $\Gamma^{(\alpha)}_{abc}$ は

$$\Gamma^{(\alpha)}_{abc} = \mathrm{E}\left[\left(\partial_a\partial_b l(\xi, Y) + \frac{1-\alpha}{2}\partial_a l(\xi, Y)\partial_b l(\xi, Y)\right)\partial_c l(\xi, Y)\right]$$

で与えられる．

一般に，あるパラメータに関して接続の係数がすべて 0 であるとき空間 S は α-平坦であると言われ，そのパラメータは α-アファイン座標系と呼ばれる．S が α-平坦であれば $-\alpha$-平坦であり，その逆も成り立つ．このことから，$\pm\alpha$-平坦な空間を双対平坦空間と呼ぶ．いま S は指数型分布族であり，± 1-接続に関して双対平坦空間になることが知られている．自然パラメータ ξ が 1-アファイン座標系であり，期待値パラメータ μ が -1-アファイン座標系である．

期待値パラメータ μ に関するポテンシャル関数 ϕ を $\phi(\mu) = \mathrm{E}[\log f(Y|\mu)] = \mathrm{E}[l(\mu, Y)]$ で定義する．このとき，関係式

$$\partial_a\psi = \mu_a,\ \partial^a\phi = \xi^a,\ \partial_a\partial_b\psi = g_{ab},\ \partial^a\partial^b\phi = g^{ab},$$
$$\phi(\mu) + \psi(\xi) - \mu^{\top}\xi = 0$$

が成り立つ．ただし，(g^{ab}) はフィッシャー情報行列 (g_{ab}) の逆行列である．

α-接続に関する測地線，α-測地線を定義する．S の曲線 $\gamma : t \mapsto \gamma(t)$ が α-測地線であるとは，任意の $t \in [0,1]$ と $c = 1, 2, \ldots, n$ に対して

$$\ddot{\gamma}^c(t) + \sum_{a,b}\dot{\gamma}^a(t)\dot{\gamma}^b(t)\left(\Gamma^{(\alpha)c}_{ab}\right)_{\gamma(t)} = 0$$

が成り立つことである．ただし，$\Gamma^{(\alpha)c}_{ab}$ は $\Gamma^{(\alpha)c}_{ab} = \sum_l \Gamma^{(\alpha)}_{abe}g^{ec}$ で与えられ，$(\gamma^1(t), \gamma^2(t), \ldots, \gamma^n(t))^{\top}$ は曲線 $\gamma(t)$ の α-アファイン座標である．測地線は直線に対応するものである．$\alpha = 1$ とすると，$\xi_{(1)}$ と $\xi_{(2)}$ を結ぶ 1-測地線は，1-アファイン座標系である自然パラメータが

$$\xi_\gamma(t) = t\xi_{(1)} + (1-t)\xi_{(2)},\ t \in [0,1]$$

と表される曲線である．同様に，$\mu_{(1)}$ と $\mu_{(2)}$ を結ぶ -1-測地線は，期待値パラメータが

$$\mu_\gamma(t) = t\mu_{(1)} + (1-t)\mu_{(2)},\ t \in [0,1]$$

と表される曲線である．

点 $\mu_{(1)}$ から点 $\xi_{(2)}$ へのカルバック–ライブラーダイバージェンス（KL ダイバージェンス）は

$$D(\mu_{(1)}|\,\xi_{(2)}) = \mathrm{E}_{\mu_{(1)}}\left[\log \frac{f(Y|\mu_{(1)})}{f(Y|\xi_{(2)})}\right]$$

で定義される．KL ダイバージェンスはユークリッド距離の二乗に対応する量である．KL ダイバージェンスはポテンシャル関数 ψ, ϕ を用いて

$$D(\mu_{(1)}|\,\xi_{(2)}) = \phi(\mu_{(1)}) + \psi(\xi_{(2)}) - \mu_{(1)}^\top \xi_{(2)}$$

と表すことができる．

-1-射影を定義する．S' を S の部分空間とし，$p \in S$ とする．点 p の S' への -1-射影とは，p からの KL ダイバージェンスが最小であるような S' の点である．

$\tilde{M} = \{f(\cdot|\xi)|\,\xi = \tilde{X}\tilde{\theta}\}$ とおく．一般化線形回帰のパラメータ $\tilde{\theta}$ は $\tilde{M}$ における 1-アファイン座標系であり，$\tilde{\eta} = \tilde{X}^\top \mu$ は $\tilde{M}$ の -1-アファイン座標系である．S における ξ と μ についての上述の定義が，$\tilde{M}$ において $\tilde{\theta}$ と $\tilde{\eta}$ についても同様に定義できる．以後，$\partial_i = \partial/\partial\theta^i$, $\partial^i = \partial/\partial\eta_i$ と略記し，添え字 a, b, c は ξ と μ に対して，添え字 i, j, k は θ と η に対して使用することにする．最尤分布は，S において，観測 y に対応する確率分布 $f(\cdot|\mu = y)$ を $\tilde{M}$ に -1-射影して得られる．

線形回帰における切片の直交化との対応を考える．$\tilde{\theta}$ の最尤推定量を $\hat{\theta}_{\mathrm{MLE}}$，$\tilde{\eta}$ の最尤推定量を $\hat{\eta}_{\mathrm{MLE}}$ とおき，$M = \{f(\cdot|\xi) \in \tilde{M}|\,\eta_0 = (\hat{\eta}_{\mathrm{MLE}})_0\}$ とおく．M は -1-平坦なサブモデルであり，双対平坦空間である．$N = \{f(\cdot|\xi)|\,\xi = \theta^0\mathbf{1}\}$ とおくと，N は切片だけからなるサブモデルである．M と N は直交することが分かり，θ^0 の推定とは別に M において θ を推定することができる．ただし，M は $(\hat{\eta}_{\mathrm{MLE}})_0$ の値に依存するサブモデルであり，線形回帰のように切片 θ^0 と θ を独立に推定することはできない点に注意が必要である．なお，説明変数選択の観点から考えると，どのような変数の組合せでも最尤推定量が M 上にあることが知られている．

今後，$\tilde{M}$ の元 $f(\cdot|\tilde{X}\tilde{\theta})$ のことを $f(\cdot|\tilde{\theta})$，M の元を $f(\cdot|\theta)$ とも表記することとする．

3 LARS と正則化

線形回帰におけるパラメータ推定手法として LARS と LASSO を紹介する．LARS は線形回帰問題におけるパラメータ推定（回帰係数の推定）のために提案されたアルゴリズ

ムであり，同時にモデル推定も行う．アルゴリズムの出力としてパラメータの推定値とモデルの系列が得られる．説明変数間の相関を内積（角度）とみなし，推定量の構成において反応変数との相関が大きな（角度の小さな）説明変数を順次追加することで推定値の系列を構成する．LARS アルゴリズムはユークリッド幾何により記述され，幾何学的に単純であり統計学的にも自然に解釈される．また，LARS は正則化の代表的な手法である LASSO との関連が指摘されている．LARS 自体は正則化ではないが，LARS アルゴリズムを修正することで LASSO による推定値を出力できることが知られている．正則化とは，最尤推定量の過学習を避けるために，対数尤度に正則化項（罰則化項）を加えたものを最大化する手法である ([7])．

計画行列は正規化されているものとする．つまり，式 (1) が成り立つことを仮定する．推定するパラメータは線形結合の係数 $\theta \in \mathbf{R}^d$ である．

3.1 LASSO

LASSO はもともと Least Absolute Shrinkage and Selection Operator の略称であったが，略称の LASSO という呼び方が広く普及している．正則化の代表的な手法であり，統計学や機械学習の分野で関連した話題が盛んに研究されている．LASSO は最適化問題として定式化され，何らかのアルゴリズムによりその問題を解くことで推定値が得られる．

LASSO は以下の制約付き最小二乗法（制約付き対数尤度最大化問題）として定義される:

$$\min_{\theta \in \mathbf{R}^d} \|y - X\theta\|_2^2 \qquad \text{s.t.} \quad \|\theta\|_1 \leq c.$$

あるいはラグランジュ形式

$$\min_{\theta \in \mathbf{R}^d} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1$$

を LASSO と呼ぶこともある．ただし，$\|\cdot\|_2$ はユークリッドノルム (l_2-ノルム)，$\|\theta\|_1 = \sum_{i=1}^d |\theta^i|$ は l_1-ノルム，$c \geq 0$ と $\lambda \geq 0$ は問題を特徴づけるパラメータである．パラメータを変化させることで得られる推定値も変化する．たとえば，c が十分大きな値をとる場合（あるいは $\lambda = 0$ の場合），得られる推定値 $\hat{\theta}$ は最尤推定値 $\hat{\theta}_{\mathrm{MLE}}$ となる．$c = 0$ の場合（あるいは λ が十分大きな値をとる場合），$\hat{\theta} = 0$ となる．LASSO により得られる推定値は疎になるという特徴をもつ．推定値が疎であるとは，成分のいくつかが厳密に 0 になることをいう．パラメータ値が変化するごとの推定値を知るために，推定量の描くパスを求めるという問題が出てくる（図 4 など）．次節で説明する LARS アルゴリズムは，LASSO 推定量のパスを求めるためのアルゴリズムという側面をもつ．

3.2 LARS

LARS (Least Angle Regression) は相関をもとにパラメータ推定を行うアルゴリズムである．相関は幾何学的には内積（あるいは角度）に対応しており，LARS アルゴリズムはユークリッド幾何で記述することができる．

説明変数の添え字集合 $\{1,2,\ldots,d\}$ の部分集合 $\mathcal{A}$ に対して $n \times |\mathcal{A}|$ 行列 $X_{\mathcal{A}}$ を

$$X_{\mathcal{A}} = (s_j x_j)_{j \in \mathcal{A}}$$

と定義する．ただし，$s_j = \pm 1$ として，このあと考える相関の符号を表すものとする．さらに，

$$G_{\mathcal{A}} = X_{\mathcal{A}}^{\top} X_{\mathcal{A}}, \quad A_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^{\top} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-1/2}$$

とおく．ただし，$\mathbf{1}_{\mathcal{A}}$ はすべての成分が 1 の $|\mathcal{A}|$ 次元ベクトルであり，また $A_{\mathcal{A}}$ はスカラーであることに注意する．

今，$X_{\mathcal{A}}$ の列ベクトルに対して同じ角度をなすような n 次元単位ベクトル $u_{\mathcal{A}}$ は

$$u_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}}, \quad w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} \tag{2}$$

と表すことができ，実際

$$X_{\mathcal{A}}^{\top} u_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{1}_{\mathcal{A}}, \quad \|u_{\mathcal{A}}\|^2 = 1 \tag{3}$$

が成り立つ．

LARS アルゴリズムは以下の通りである．ただし，行列 $X_{\mathcal{A}_k}$ を X_k と表すことにする．$u_{\mathcal{A}_k}$, $A_{\mathcal{A}_k}$ などについても同様とする．

LARS

入力： 計画行列 $X = (x_i^a)_{1 \le a \le n,\, 1 \le i \le d} = (x_1, x_2, \ldots, x_d)$,
反応変数 $y = (y_1, y_2, \ldots, y_n)^{\top}$

出力： 推定値の系列 $\hat{\theta}_{(0)}, \hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(d)}$, あるいは推定値のパス

1. $\widehat{\mu}_0 = 0$, $k = 0$ とする．
2. $\widehat{c}_{k+1} = X^{\top}(y - \widehat{\mu}_k)$ とおく．
3. $\widehat{C}_{k+1} = \max_j\{|\widehat{c}_{k+1,j}|\}$, $\mathcal{A}_{k+1} = \{j : |\widehat{c}_{k+1,j}| = \widehat{C}_{k+1}\}$ とおく．
4. $s_j = \mathrm{sign}\{\widehat{c}_{k+1,j}\}$ $(j \in \mathcal{A}_{k+1})$ を用いて，X_{k+1}, A_{k+1}, u_{k+1} を計算する．
5. $a_{k+1} = X^{\top} u_{k+1}$ とおく

6. $\mathcal{A}_{k+1} \neq \{1,2,\ldots,d\}$ の場合，$\widehat{\gamma}$ を

$$\widehat{\gamma} = \mathop{\min}_{j \in \mathcal{A}_{k+1}^c}{}^{+} \left\{ \frac{\widehat{C}_{k+1} - \widehat{c}_{k+1,j}}{A_{k+1} - a_{k+1,j}}, \frac{\widehat{C}_{k+1} + \widehat{c}_{k+1,j}}{A_{k+1} + a_{k+1,j}} \right\} > 0, \tag{4}$$
$$\widehat{\mu}_{k+1} := \widehat{\mu}_k + \widehat{\gamma} u_{k+1}, \quad k := k+1$$

として，ステップ 2 へ戻る．$\mathcal{A}_{k+1} = \{1,2,\ldots,d\}$ の場合，ステップ 7 へ．

7. $\widehat{\gamma}$ を

$$\widehat{\gamma} = \frac{\widehat{C}_{k+1}}{A_{k+1}}$$

と定め，

$$\widehat{\mu}_{k+1} := \widehat{\mu}_k + \widehat{\gamma} u_{k+1}$$

として，アルゴリズムを終了する．

ただし，ステップ 7 の $\min^{+}_{j \in \mathcal{A}^c}$ は正の値のうちで最小値をとることを意味する．

この節の残りでは LARS のアルゴリズムのステップ 6 に説明を加える．$\gamma > 0$ に対して

$$\mu(\gamma) = \widehat{\mu}_k + \gamma u_{k+1}$$

と定義すると，x_j と残差 $y - \mu(\gamma)$ の相関は

$$\begin{aligned} c_j(\gamma) &= x_j^\top (y - \mu(\gamma)) \\ &= x_j^\top (y - \widehat{\mu}_k) - \gamma x_j^\top u_{k+1} \\ &= \widehat{c}_{k,j} - \gamma a_{k+1,j} \end{aligned} \tag{5}$$

となる．$j \in \mathcal{A}_{k+1}$ に対して

$$\begin{aligned} |c_j(\gamma)| &= |\widehat{c}_j - \gamma a_j| \\ &= |\widehat{C}_{k+1} - \gamma A_{k+1}| \\ &= \widehat{C}_{k+1} - \gamma A_{k+1} \end{aligned} \tag{6}$$

が成り立つので，相関の絶対値の最大値 $|c_j(\gamma)|$ $(j \in \mathcal{A}_{k+1})$ は各成分で同じように減少することが分かる．$j \in \mathcal{A}_{k+1}^c$ に対しては，式 (5) と式 (6) により $c_j(\gamma)$ が相関の最大値 $\widehat{C}_{k+1} - \gamma A_{k+1}$ と一致するのは $\gamma = (\widehat{C} - \widehat{c}_j)/(A_{\mathcal{A}} - a_j)$ のときだと分かる．同様に，$-c_j(\gamma)$ が最大値と一致するのは $\gamma = (\widehat{C} + \widehat{c}_j)/(A_{\mathcal{A}} + a_j)$ のときだと分かる．したがって，式 (4) の $\widehat{\gamma}$ はこのような γ のうちで最小のものをとっていることになる．そして，その最小の γ を与えたインデックス $\widehat{j}$ が次の反復の時に集合 $\mathcal{A}_{k+2}$ に含まれることになる．

4 情報幾何を利用した推定手法

一般化線形回帰に対して情報幾何を利用したパラメータ手法をいくつか紹介する．これらの手法は一般化線形回帰以外の問題に対しても拡張されるが [9, 10]，本稿では一般化線形回帰の文脈で紹介する．パラメータである回帰係数 $\theta \in \mathbf{R}^d$ を推定することが目的である．ここでは切片などの局外パラメータの推定は扱わず，それぞれの手法の主要な部分を説明する．指数型分布族のなす双対平坦空間における推定量の移動・更新である．

情報幾何を利用しない LASSO の一般化として [6] や [13] などがある．情報幾何を使った手法としては，ここで紹介する以外にも [3], [16] がある．

4.1 Bisector Regression

Hirose and Komaki (2010) で提案された Bisector Regression (BR) アルゴリズムを説明する．$i \in \mathcal{A} \subseteq \{1, 2, \ldots, d\}, s \in \mathbf{R}$ に対して，ふたつのサブモデル $M(\mathcal{A})$ と $M(i, s, \mathcal{A})$ を

$$M(\mathcal{A}) = \left\{ f(\cdot|\theta) \mid \theta^j = 0 \ (j \notin \mathcal{A}) \right\},$$
$$M(i, s, \mathcal{A}) = \left\{ f(\cdot|\theta) \mid \theta^i = s, \theta^j = 0 \ (j \notin \mathcal{A}) \right\}$$

により定義する．

まず BR アルゴリズムで何をしようとしているのかについて幾何学的な説明をし，その後に BR アルゴリズムを紹介する．BR アルゴリズムは LARS の拡張を目指したものである．アルゴリズムの反復ごとにパラメータの推定値を出力し，推定値の入るモデルが入れ子状に順に小さくなっていくのが特徴である．これによりアルゴリズムの反復はパラメータの次元 d と同じ回数だけであることが分かり，考える説明変数の組合せの候補数が大幅に削減される．

LARS では，説明変数と反応変数の相関を内積（角度）として扱い，反応変数となす相関の大きい（角度の小さい）説明変数を推定量の構成に順次取り込んでいた．すでに推定量を構成している説明変数同士は反応変数となす相関（角度）が互いに一致しており，幾何学的には推定量が角の二等分線上を動いているものとみなせる．双対平坦空間の情報幾何を利用して LARS の拡張を行うにあたって，角の二等分線を距離を用いた形で一般化し，BR アルゴリズムはその曲線上を推定量が動くようなアルゴリズムになっている．ただし，BR は LARS の厳密な拡張ではない．LARS では原点を出発点にしてアルゴリズムが始まり，フルモデルの最尤推定量を推定量の最終的な到達点にしていた．しかし，

BR では，推定量はフルモデルの最尤推定量を出発し，最終的に原点にたどりつくようになっている．これはユークリッド空間を双対平坦空間に一般化する際に生じる困難を避けるためであった．推定量の移動（更新）によりパラメータの推定値とモデルの組の系列を出力する点は同じである．

BR アルゴリズムは 2.3 節で導入したサブモデル M ににおいて働くアルゴリズムである．M 内を動く推定量の点を考え，$M(\mathcal{A})$ の形で表される M のサブモデルに推定量が順次動いていく手続きを記述する．図 1 は推定量の点が角の二等分線に対応する曲線に沿って移動する様子を表したものであり，BR アルゴリズムではこの手続きを反復することになる．推定量の点が k 回目の反復で得られた推定値 $\hat{\theta}_{(k)}$ にいるものとする．アルゴリズム開始時には $\hat{\theta}_{(0)} = \hat{\theta}_{\mathrm{MLE}}$ である．$\hat{\theta}_{(k)}$ は $d-k$ 個の非零成分をもち，$|\mathcal{A}| = d-k$ を満たすある $\mathcal{A} \subseteq \{1, 2, \ldots, d\}$ に対して $\hat{\theta}_{(k)} \in M(\mathcal{A})$ となっている．各 $i \in \mathcal{A}$ に対して，サブモデル $M(\mathcal{A}\backslash\{i\})$ を考える．サブモデル $M(\mathcal{A}\backslash\{i\})$ は，$M(\mathcal{A})$ よりも一次元だけ次元の小さなサブモデルで $M(\mathcal{A})$ に含まれている．現在の推定値 $\hat{\theta}_{(k)}$ から各サブモデル $M(\mathcal{A}\backslash\{i\})$ への KL ダイバージェンスを測り，その最小値を t^* とおくことにする．$\hat{\theta}_{(k)}$ からの KL ダイバージェンスが t^* になるように各サブモデル $M(\mathcal{A}\backslash\{i\})$ を平行移動したものは，それぞれ $M(i, s_i^*, \mathcal{A})$ の形で表される．$M(\mathcal{A})$ における各 $M(i, s_i^*, \mathcal{A})$ の交点を $\hat{\theta}_{(k+1)}$ と決める．このとき，t^* を達成した $i^* \in \mathcal{A}$ に対して $\hat{\theta}_{(k+1)} \in M(\mathcal{A}\backslash\{i^*\})$ であるから，推定値を含むサブモデルの次元がひとつ小さくなったことが分かる．なお，上述の説明では角の二等分線に対応する曲線を陽に用いていないが，曲線の両端がそれぞれ推定値 $\hat{\theta}_{(k)}$ と $\hat{\theta}_{(k+1)}$ に対応している．

角の二等分線に対応する曲線について図 1 にしたがって簡単に説明する．現在の推定値 $\hat{\theta}_{(k)}$ から各サブモデル $M(\mathcal{A}\backslash\{i\})$ への -1-射影を $\overline{\theta}_{(k)}^{-i}$ とおく．$\hat{\theta}_{(k)}$ から各 $\overline{\theta}_{(k)}^{-i}$ への KL ダイバージェンスのうち最小値が t^* であり，最小値を達成するのは $i^* \in \mathcal{A}$ である．$\hat{\theta}_{(k)}$ から $\tilde{\theta}_{(k)}^{-i'}$ への KL ダイバージェンスが $\hat{\theta}_{(k)}$ から $\overline{\theta}_{(k)}^{-i^*}$ への KL ダイバージェンスと一致するように，サブモデル $M(\mathcal{A} \setminus \{i'\})$ を平行移動して $M(i', s_{i'}^*, \mathcal{A})$ とする．ただし，$\tilde{\theta}_{(k)}^{-i'}$ は $\hat{\theta}_{(k)}$ の $M(i', s_{i'}^*, \mathcal{A})$ への -1-射影であり，$s_{i'}^*$ は $D\left(\hat{\theta}_{(k)} \mid \tilde{\theta}_{(k)}^{-i'}\right) = t^*$ を満たす値として決まる．i に対して同様のことを考えると，$s_{i^*}^* = 0$, $\tilde{\theta}_{(k)}^{-i^*} = \overline{\theta}_{(k)}^{-i^*}$ であることが分かる．新しい推定値 $\widehat{\theta}_{(k+1)}$ は $M(i', s_{i'}^*, \mathcal{A})$ と $M(i^*, 0, \mathcal{A}) = M(\mathcal{A} \setminus \{i^*\})$ との交点であった．ここで，$0 \leq t \leq t^*$ なる t に対して，t^* の代わりに t を使って $\tilde{\theta}_{(k)}^{-i'}$ と $\tilde{\theta}_{(k)}^{-i^*}$ に対応する点を作ることができる．そして，それらをそれぞれ含むようなサブモデル $M(i', s_{i'}(t), \mathcal{A})$ と $M(i^*, s_{i^*}(t), \mathcal{A})$ の交点 $\hat{\theta}(t)$ を考えることができる（図 1 における点線上の点になる）．特に，$t = 0$ のとき $\hat{\theta}(t) = \hat{\theta}_{(k)}$ であり，$t = t^*$ のとき $\hat{\theta}(t) = \hat{\theta}_{(k+1)}$ である．拡張ピタゴ

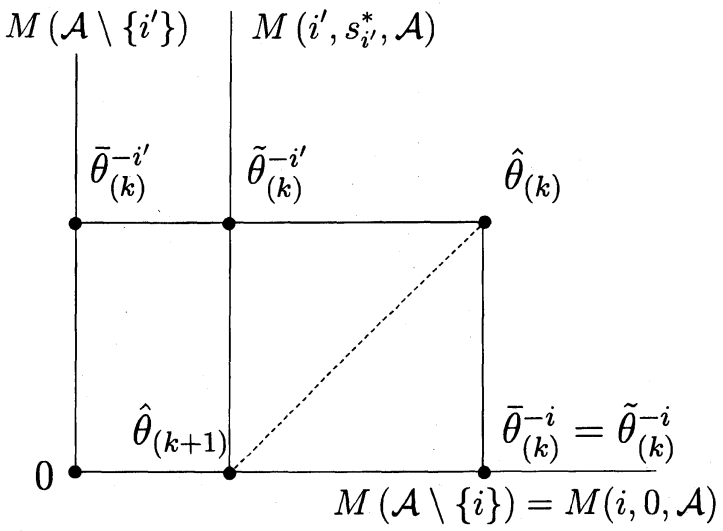

図 1　推定量の更新．$\mathcal{A} \subseteq \{1, 2, \ldots, d\}$: 説明変数を指定するインデックス．$M(i, s, \mathcal{A}) = \{\theta \,|\, \theta^i = s,\, \theta^j = 0\ (j \notin \mathcal{A})\}$．0: 原点．$\hat{\theta}_{(k)}$: k 番目の推定量．$\hat{\theta}_{(k+1)}$: $k+1$ 番目の推定量．$\bar{\theta}^{-i}_{(k)}$: サブモデル $M(\mathcal{A}\setminus\{i\}) = M(i, 0, \mathcal{A})$ における最尤推定量，あるいは $\hat{\theta}_{(k)}$ の $M(\mathcal{A}\setminus\{i\})$ への -1-射影でも同じ．$D\left(\hat{\theta}_{(k)} \,|\, \bar{\theta}^{-i'}_{(k)}\right) > D\left(\hat{\theta}_{(k)} \,|\, \bar{\theta}^{-i^*}_{(k)}\right)$ である．$\tilde{\theta}^{-i'}_{(k)}$: $\hat{\theta}_{(k)}$ の $M(i', s^*_{i'}, \mathcal{A})$ への -1-射影．s^*_i: 条件 $D\left(\hat{\theta}_{(k)} \,|\, \tilde{\theta}^{-i'}_{(k)}\right) = D\left(\hat{\theta}_{(k)} \,|\, \bar{\theta}^{-i^*}_{(k)}\right)$ より決まる値．$\hat{\theta}_{(k)}$ は $M(\mathcal{A}\setminus\{i'\})$ よりも $M(\mathcal{A}\setminus\{i^*\})$ に近いので $\tilde{\theta}^{-i^*}_{(k)} = \bar{\theta}^{-i^*}_{(k)}$ である．$\hat{\theta}_{(k+1)}$ は $M(i^*, 0, \mathcal{A})$ と $M(i', s^*_{i'}, \mathcal{A})$ の交点として定義される．点線は角の二等分線に対応する曲線である．BR アルゴリズムでは，推定量がこの曲線に沿って動いていると見ることができる．

ラスの定理より，

$$D\left(\hat{\theta}_{(k)} \,|\, \tilde{\theta}^{-i'}(t)\right) = D\left(\hat{\theta}_{(k)} \,|\, \tilde{\theta}^{-i^*}(t)\right),$$
$$D\left(\tilde{\theta}^{-i'}(t) \,|\, \hat{\theta}(t)\right) = D\left(\tilde{\theta}^{-i^*}(t) \,|\, \hat{\theta}(t)\right)$$

が成り立つ．このことから $\{\hat{\theta}(t) | 0 \le t \le t^*\}$ が角の二等分線に対応する曲線になっていることが分かる．

BR アルゴリズムは以下の通りである．ステップ 2 から 6 が反復され，反復ごとにパラメータの推定値とモデルが出力される．ステップ 4 の $l^{-i}_{(k)}$ は，$\hat{\theta}_{(k)}$ と $\bar{\theta}^{-i}_{(k)}$ を結ぶ -1-測地線である．

BR

入力：　計画行列 $X = (x^a_i)_{1 \le a \le n,\, 1 \le i \le d} = (x_1, x_2, \ldots, x_d)$,

反応変数 $y = (y_1, y_2, \ldots, y_n)^\top$

出力： 推定値の系列 $\hat{\theta}_{(0)}, \hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(d)}$,

1. $\hat{\theta}_{(0)} := \hat{\theta}_{\mathrm{MLE}}$, $\mathcal{A} := \{i | 1 \leq i \leq d\}$, $k := 0$ とおく.
2. 各 $i \in \mathcal{A}$ に対してモデル $M(\mathcal{A} \setminus \{i\})$ の最尤推定量 $\overline{\theta}_{(k)}^{-i}$ を計算する.
3. $t^* := \min_{i \in \mathcal{A}} D\left(\hat{\theta}_{(k)} \mid \overline{\theta}_{(k)}^{-i}\right)$, $i^* := \arg\min_{i \in \mathcal{A}} D\left(\hat{\theta}_{(k)} \mid \overline{\theta}_{(k)}^{-i}\right)$ とおく.
4. 各 $i \in \mathcal{A}$ に対して, 二条件 $D(\hat{\theta}_{(k)} \mid \tilde{\theta}_{(k)}^{-i}) = t^*$, $\tilde{\theta}_{(k)}^{-i} \in M(i, s_i^*, \mathcal{A})$ を満たすような s_i^* と $\tilde{\theta}_{(k)}^{-i} \in l_{(k)}^{-i}$ を求める.
5. $i \in \mathcal{A}$ に対して $\hat{\theta}_{(k+1)}^{i} := s_i^*$, $j \notin \mathcal{A}$ に対して $\hat{\theta}_{(k+1)}^{j} := 0$ とする.
6. $k+1 < d-1$ の場合, $k := k+1, \mathcal{A} := \mathcal{A} \setminus \{i^*\}$ としてステップ 2 へ戻る. $k+1 = d-1$ の場合, ステップ 7 へ進む.
7. $\hat{\theta}_{(d)} := 0$ とし, アルゴリズム終了.

ステップ 5 では, $i^* \in \mathcal{A}$ であるが $\hat{\theta}_{(k+1)}^{i^*} = 0$ であることに注意が必要である. このことにより, 1 回の反復において θ の成分がひとつ 0 になり, モデルが入れ子状に小さくなっていくことになる.

4.2 Differential Geometric LARS

[4] において提案された Differential Geometric Least Angle Regression (DGLARS) アルゴリズムを紹介する. DGLARS アルゴリズムも LARS アルゴリズムと同様に推定値の系列とその間の（近似的な）パスを出力する. パスのパラメータを γ で表わすことにする. 推定量 $\hat{\theta}$ を構成している説明変数の添え字集合を active set と呼ぶ. この active set は単調増加するので, active set が変化する点を $0 \leq \gamma^{(p)} \leq \gamma^{(p-1)} \leq \cdots \leq \gamma^{(1)}$ とおく. DGLARS アルゴリズムではパラメータ γ は $\gamma_{(1)}$ から始まり, 単調減少して 0 に到る. $\gamma = 0$ がフルモデルの最尤推定量に対応する.

DGLARS アルゴリズムの概要を幾何学的に説明する. 推定量の出発点は原点 $\hat{\theta} = 0$ である. 最終的には推定量はフルモデルの最尤推定量に到る. 現時点での推定量による「残差」と最も小さな「角度」をもつ説明変数を用いて推定量を構成する. ある曲線上を推定量が移動すると, これまで推定量を構成していなかった説明変数が新しく「残差」と小さな「角度」をなすようになる. そこで新しい説明変数を active set に追加し, active set を更新する. このような手順で逐次 active set を大きくしながら推定量を更新する.

ユークリッド空間と違い, 一般の双対平坦空間では「残差」と説明変数のなす「角度」

が自明ではない．DGLARS では，点 $f(\cdot|\theta)$ の接空間 $T_{f(\cdot|\theta)}S$ の元である残差ベクトル

$$r(\mu(\theta), y, Y) = \sum_{a=1}^{n} \{y_a - \mu_a(\theta)\} \partial^a l(\mu(\theta), Y)$$

を考える．各説明変数に対応する接ベクトルとして $\partial_i l(\theta, Y) \in T_{f(\cdot|\theta)}M$ をとる．残差ベクトル $r(\mu(\theta), y, Y)$ と基底ベクトル $\partial_i l(\theta, Y)$ との内積については

$$\partial_i l(\theta, y) = \langle \partial_i l(\theta, Y), r(\mu(\theta), y, Y) \rangle_{T_{f(\cdot|\theta)}S} \tag{7}$$

が成り立つ．推定量 $\hat{\theta}(\gamma)$ における接空間 $T_{f(\cdot|\hat{\theta}(\gamma))}S$ において，これらのベクトルのなす内積にもとづいて推定量を移動させる．DGLARS アルゴリズムでは，接ベクトル間の内積を直接計算するのではなく，符号付きラオスコア統計量

$$\begin{aligned} r_i^u(\gamma) &= (g_{ii})^{-1/2} \partial_i l(\theta(\gamma), y) \\ &= \|r(\theta(\gamma), y, Y)\|_{T_{f(\cdot|\theta)}S} \cos \rho_i(\theta(\gamma)) \end{aligned}$$

を利用する．ただし，$\rho_i(\theta)$ は，点 $f(\cdot|\theta)$ の接空間 $T_{f(\cdot|\theta)}S$ における残差ベクトル $r(\theta, y, Y)$ と基底ベクトル $\partial_i l(\theta, Y)$ とのなす角度であり，ふたつめの等号は式 (7) から導かれる．$\|r(\theta(\gamma), y, Y)\|_{T_{f(\cdot|\theta)}S}$ は i によらない量であり，角度の大きさが $r_i^u(\gamma)$ で測られることが分かる．

DGLARS アルゴリズムの概要は以下の通りである．文献 [4] において，式 (8) は generalized equiangularity condition と呼ばれている．より詳しい内容は文献 [4] を参照のこと．

DGLARS

入力： 計画行列 $X = (x_i^a)_{1 \le a \le n, 1 \le i \le d} = (x_1, x_2, \ldots, x_d)$,
反応変数 $y = (y_1, y_2, \ldots, y_n)^\top$

出力： 推定値の系列 $\hat{\theta}_{(0)}, \hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(d)}$, あるいは推定値のパス

1. $\hat{\theta}_{(0)} = 0$, $k = 1$ とおく．
2. $\mathcal{A}_1 = \{a_1\} = \arg\max_{j \in 1,2,\ldots,d} |r_j^u(0)|$, $\gamma_{(1)} = |r_{a_1}^u(0)|$ とおく．
3. $\gamma \le \gamma_{(k)}$ に対して，条件

$$\begin{aligned} &\left|r_{a_i}^u(\gamma)\right| = \left|r_{a_j}^u(\gamma)\right|, \quad \forall a_i, a_j \in \mathcal{A}_k, \\ &\left|r_{a_h^c}^u(\gamma)\right| < \left|r_{a_i}^u(\gamma)\right|, \quad \forall a_h^c \notin \mathcal{A}_k, \forall a_i \in \mathcal{A}_k \end{aligned} \tag{8}$$

を満たすような $\hat{\theta}(\gamma)$ を求める．

4. γ を小さくしていき，ある $a_h^c \notin \mathcal{A}_k$ に対して

$$\left|r^u_{a_h^c}(\gamma)\right| = \left|r^u_{a_i}(\gamma)\right|, \quad \forall a_i \in \mathcal{A}_k$$

が成り立ったら，$\hat{\theta}_{(k)} = \hat{\theta}(\gamma)$, $\gamma_{(k+1)} = \gamma$, $\mathcal{A}_{k+1} = \arg\max_{j\in 1,2,\ldots,d} |r^u_j(\gamma)|$, $k = k+1$ とおいてステップ 3 に進む．

5. $\gamma = 0$ となったらアルゴリズムを終了する．

4.3 例

一般化線形回帰のひとつとしてロジスティック回帰を考える．ロジスティック回帰については 2.2 節で簡単に説明した．データとして South African Heart Disease (SAHD) データを使用した ([7])．SAHD データは 9 個の説明変数 $x_1, x_2, \ldots, x_9$ と反応変数 y からなり，$n = 462$ 人分のデータが集められている．

SAHD データに対する BR の結果を図 2 に示す．横軸は BR が出力した各推定値の l_1-ノルムを示している．縦軸は説明変数 $x_i\,(i = 1, 2, \ldots, 9)$ の回帰係数 $\theta^i\,(i = 1, 2, \ldots, 9)$ の値を表している．BR の出発点であるフルモデルの最尤推定値が図 2 の右端に対応する．BR アルゴリズムは図の右端から左端に向かって進む．図に入っている縦線はそれぞれ BRGLM アルゴリズムが出力した推定値に対応している．左に進むにしたがって非零成分の個数がひとつずつ減っており，縦線の上の数字は非零成分の個数を示している．図 2 より，$\hat{\theta}$ の成分が 0 になった順番は $\theta^8, \theta^4, \theta^1, \theta^7, \theta^3, \theta^2, \theta^6, \theta^5, \theta^9$ である．

DGLARS による結果を図 3 に示す．計算には R の **dglars** パッケージを利用した．図 3 の横軸は推定値の l_1-ノルムを，縦軸はパラメータ $\theta^i\,(i = 1, 2, \ldots, 9)$ の値を示している．アルゴリズムは図の左端からスタートし，右端に到る．

情報幾何を利用していない [13] の方法による結果を図 4 に示す．この方法は LASSO を一般化線形回帰に拡張した正則化の方法である．計算には R の **glmpath** パッケージを利用した．図 4 の横軸は推定値の l_1-ノルムを，縦軸はパラメータ $\theta^i\,(i = 1, 2, \ldots, 9)$ の値を示している．縦線の上の数字は非零成分の個数を示している．図 4 の右端から左端に見ていくと，パラメータが 0 になる順番は $\theta^8, \theta^4, \theta^7, \theta^1, \theta^6, \theta^3, \theta^2, \theta^5, \theta^9$ であったことが分かる．

5 まとめ

統計学における一般化線形回帰と情報幾何について説明した．一般化線形回帰の問題を紹介して，情報幾何の視点から局外パラメータの説明をした．線形回帰におけるパラメー

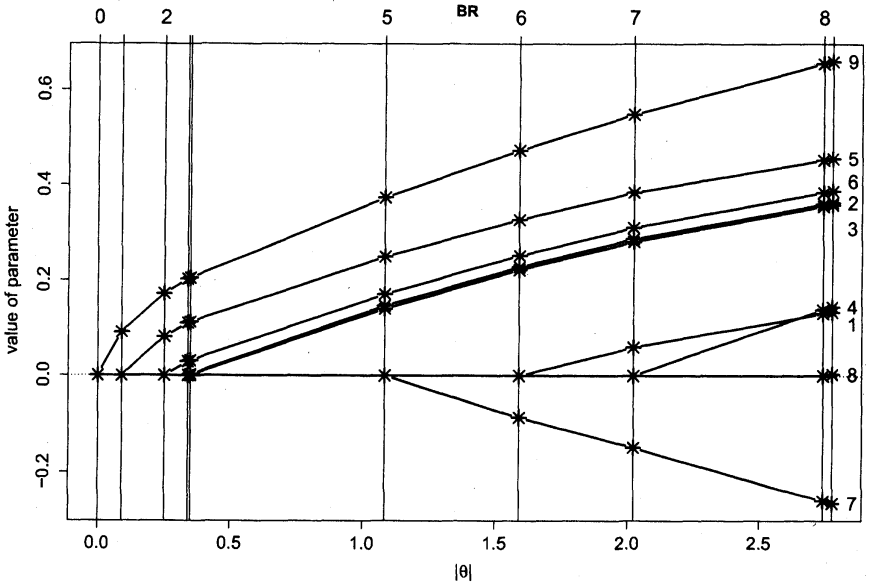

図 2　SAHD データのロジスティック回帰に対する BR の結果．横軸は各推定値の l_1-ノルムをを示している．縦軸は説明変数 $x_i\,(i = 1, 2, \ldots, 9)$ の回帰係数 $\theta^i\,(i = 1, 2, \ldots, 9)$ の値を表す．図の右端はフルモデルの最尤推定値であり，BR アルゴリズムの出発点を表している．左端は BR アルゴリズムの出力した最後の推定値であり，推定したいパラメータがすべて 0 になっている．図に入っている縦線はそれぞれ BR アルゴリズムが出力した推定値に対応している．左に進むにしたがって非零成分の個数がひとつずつ減っていく．縦線の上の数字は非零成分の個数を示している．パラメータが 0 になる順番は $\theta^8, \theta^4, \theta^1, \theta^7, \theta^3, \theta^2, \theta^6, \theta^5, \theta^9$ であった．

タ推定手法である LARS と LASSO を紹介した．特に，LARS はユークリッド幾何により記述できるアルゴリズムであり，変数間の相関を内積（角度）として扱うことにより，相関にもとづく推定を行っている点を確認した．また，情報幾何を利用したパラメータ推定法を紹介した．特に，LARS の拡張という観点から関連した手法を紹介した．ダイバージェンス，あるいは接ベクトル間の内積を利用することにより，ユークリッド空間における角の二等分線を拡張し，得られた曲線を利用してパラメータ推定を行っていた．

参考文献

[1] S. Amari (1985). *Differential-Geometrical Methods in Statistics.* Springer Lecture Notes in Statistics, **28**, Springer.

[2] S. Amari and H. Nagaoka (2000). *Methods of Information Geometry.* Translations

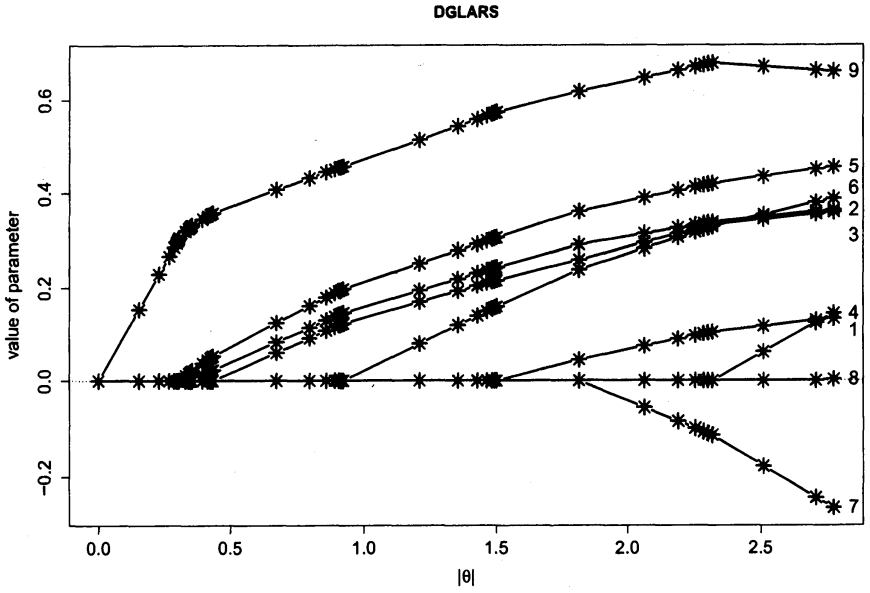

図 3　SAHD データのロジスティック回帰に対する DGLARS の結果．横軸は推定値の l_1-ノルムを，縦軸はパラメータ θ^i の値を示している．図の左端からスタートし，右端に到る．

of Mathematical Monographs, **191**, Oxford University Press.

[3] S. Amari and M. Yukawa (2013). Minkovskian Gradient for Sparse Optimization. *IEEE Journal of Selected Topics in Signal Processing*, **7**, pp. 576–585.

[4] L. Augugliaro, A.M. Mineo, and E.C. Wit (2013). Differential Geometric Least Angle Regression: A Differential Geometric Approach to Sparse Generalized Linear Models. *Journal of the Royal Statistical Society B*, **75**, pp. 471–498.

[5] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least Angle Regression (with discussion), *The Annals of Statistics*, **32**, 407–499.

[6] J. Friedman, T. Hastie, and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, **33**, 1–22.

[7] T. Hastie, R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. Springer, New York.

[8] Y. Hirose and F. Komaki (2010). An Extension of Least Angle Regression Based on the Information Geometry of Dually Flat Spaces, *Journal of Computational and*

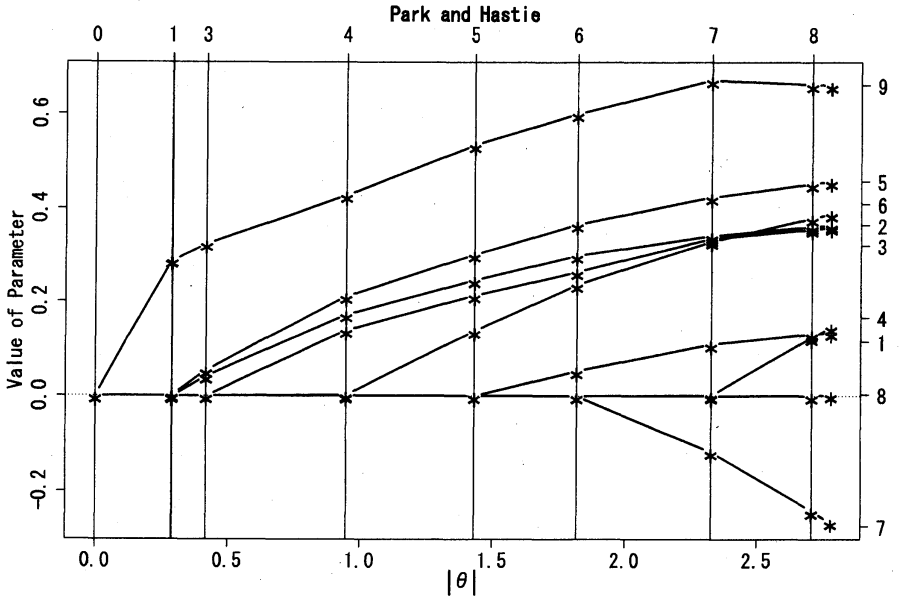

図 4 SAHD データに対する Park and Hastie の方法の結果. 横軸は推定値の l_1-ノルムを, 縦軸はパラメータ θ^i の値を示している. 縦線の上の数字は非零成分の個数を示している. 図の右端から左端に見ていくと, パラメータが 0 になる順番は $\theta^8, \theta^4, \theta^7, \theta^1, \theta^6, \theta^3, \theta^2, \theta^5, \theta^9$ であった.

Graphical Statistics, **19**, 1007–1023.

[9] Y. Hirose and F. Komaki (2013). Edge Selection Based on the Geometry of Dually Flat Spaces for Gaussian Graphical Models, *Statistics and Computing*, **23**, 793–800.

[10] 廣瀬善大, 駒木文保 (2013). 双対平坦空間の情報幾何を利用した統計的推定, 京都大学数理解析研究所講究録, **1832**, 26–44.

[11] R. Kass and P. Vos (1997). *Geometrical Foundations of Asymptotic Inference*. John Wiley, New York.

[12] P. McCullagh and J. A. Nelder (1989). *Generalized Linear Models*, 2nd Edition. Chapman and Hall/CRC, Boca Raton.

[13] M. Y. Park and T. Hastie (2007). L_1-Regularization Path Algorithm for Generalized Linear Models, *Journal of the Royal Statistical Society B*, **69**, 659–677.

[14] R. Tibshirani (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society B*, **58**, 267–288.

[15] R. Tibshirani (2011). Regression Shrinkage and Selection via the Lasso: a Retrospective, *Journal of the Royal Statistical Society B*, **73**, 273–282.

[16] M. Yukawa and S. Amari (2011). On Extensions of LARS by Information Geometry: Convex Objectives and l_p-Norm. In *APSIPA Annual Summit and Conference: Special Session on Recent Advances in Adaptive/Sparse Signal Processing.*