日本統計学会誌
第 39 巻, 第 2 号, 2010 年 3 月
211 頁 ～ 242 頁

回帰モデリングとL_1型正則化法の最近の展開

川野　秀一*, 廣瀬　慧*, 立石　正平*, 小西　貞則†

Recent Development in Regression Modeling and L_1 Type Regularization

Shuichi Kawano*, Kei Hirose*, Shohei Tateishi* and Sadanori Konishi†

複雑な非線形現象を捉えるための統計的モデリング手法の一つとして，基底展開法に基づく非線形回帰モデリングがある．本稿では，まず様々な基底関数に基づく非線形回帰モデルとモデルの推定を中心とした研究について述べる．次に，極めて次元の高いデータに基づく線形回帰モデルの推定と変数選択において，それらを同時に実行可能な手法として研究が進展中のL_1型正則化推定法に基づくモデリングの研究を紹介する．

This paper describes two topics about which there has been a great discussion in the past 15 years. Firstly, we present an account of the recent developments of nonlinear statistical modeling via basis expansions that has been widely used for analyzing data with complex structure. Second topic is high dimensional data analysis which has become increasingly frequent and important in various fields of research such as life science, system engineering and machine learning. We illustrate various L_1 type regularization methods that many statisticians have recently proposed as techniques for performing model selection and estimation simultaneously.

キーワード: L_1型正則化法，基底展開法，高次元データ，非線形回帰，モデル選択

1. はじめに

計算機システムと計測・測定技術の高度な発展は，生命科学，材料科学，地球環境科学，システム工学，マーケティングなど諸科学の様々な分野で大量かつ複雑なデータの獲得と蓄積を可能とした．特に，複雑な非線形構造を内在するデータや超高次元データなど，多様な様相を呈する現象を背後に持つデータが急速に増加している．蓄積されたデータから，その背後にある自然現象や社会現象を解明し，新たな知識発見と問題解決の糸口を探るには，現象の情報源であるデータの中から有益な情報やパターンを効率的に抽出するための手法の開発が不可欠である．特に，現象の結果とそれに影響を及ぼすと考えられる複数の要因を結びつけ，現象発生の確率的メカニズムを捉えるための基礎的な役割を担うのが統計的モデリングである．

* 九州大学大学院数理学府：〒 819-0395 福岡県福岡市西区元岡 744 番地.
† 九州大学大学院数理学研究院：〒 819-0395 福岡県福岡市西区元岡 744 番地.

本稿では，まずはじめに諸科学の様々な分野で応用されている非線形回帰モデリングに関する研究について述べる．特に，種々の非線形回帰モデルを統一的な枠組みで捉えることができる基底展開法に基づく非線形回帰モデリングのプロセスを通して，一連の研究成果を紹介する．

線形回帰モデルの回帰係数の2乗和を罰則項（正則化項）とした正則化推定法はリッジ推定 (Horel and Kennard (1970)) と呼ばれ，当てはめたモデルの安定化に寄与する方法として知られている．これに対して，回帰係数の絶対値（L_1 ノルム）の和を正則化項としたのが Tibshirani (1996) によって提唱された lasso と呼ばれる推定法である．その特徴は，モデルの推定と変数選択を同時に実行できる点にある．このため，データ数に比して極めて高次元のデータに基づくモデルの変数選択を可能とする一つの方法として注目を集め，近年，様々な L_1 タイプの正則化項に基づく L_1 型正則化法による回帰モデリングの研究が進展しつつある．本稿の2つめの目的は，この L_1 型正則化法の実行プロセスと最近の研究を紹介することにある．

次節で，これまでに述べた2つの目的に対する本稿の構成を，一般的な回帰モデリングの枠組みを通して具体的に述べる．

2. 正則化回帰モデリング

本節では，一般的な枠組みで回帰モデリングについて述べ，次節以降で具体的に現象を近似するモデル，モデルの推定と評価・選択について議論する．

2.1 モデルの想定

目的変数 Y と p 個の説明変数 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ に関して，n 組のデータ $\{(y_\alpha, \boldsymbol{x}_\alpha); \alpha = 1, \ldots, n\}$ が観測されたとする．一般に回帰モデルは，各データ点 $\boldsymbol{x}_\alpha$ における Y_α の確率的変動を表す成分と，その点での期待値 $E[Y_\alpha|\boldsymbol{x}_\alpha] = u(\boldsymbol{x}_\alpha)$ $(\alpha = 1, \ldots, n)$ に対して想定する現象の構造を表す成分（平均構造）から成る．

いま，データ $y_1, \ldots, y_n$ は

$$y_\alpha = u(\boldsymbol{x}_\alpha) + \varepsilon_\alpha, \quad \alpha = 1, \ldots, n \tag{2.1}$$

に従って生成されたとする．これまでに観測・測定データからノイズを除去し，現象の構造を反映する平均構造 $u(\boldsymbol{x})$ を近似するための様々なモデルが提案されている．これらのモデルの多くは，次のように基底関数と呼ばれる既知の関数 $\phi_j(\boldsymbol{x})$ $(j = 1, \ldots, m)$ の線形結合によって表すことができる．

$$u(\boldsymbol{x}; \boldsymbol{w}) = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}). \tag{2.2}$$

例えば，線形回帰モデルは，説明変数 $\boldsymbol{x}=(x_1,\ldots,x_p)^T$ に対して $\phi_j(\boldsymbol{x})=x_j$ $(j=1,\ldots,p)$ と置いて，新たに基底関数 $\phi_0(\boldsymbol{x})=1$ と切片 w_0 を付け加えたものである．また，説明変数 x に対して基底関数 $\phi_0(x)=1$ と切片 w_0 に加えて $\phi_j(x)=x^j$ としたのが，多項式回帰モデルである．さらに，スプラインをはじめとして複雑な非線形構造を内包する現象を捉えるためのモデルが種々提案されており，これらのモデルについては第3節で紹介する．

その他の非線形構造を捉える手法としては，カーネル関数 (Wand and Jones (1995), Simonoff (1996))，局所尤度法 (Fan and Gijbels (1996), Loader (1999))，サポートベクター回帰 (Smola and Schölkopf (2004)) などが挙げられる．

2.2 モデルの推定

基底関数展開に基づくモデル

$$y_\alpha=\sum_{j=1}^{m}w_j\phi_j(\boldsymbol{x})+\varepsilon_\alpha,\quad \alpha=1,\ldots,n \tag{2.3}$$

に対して，誤差項 $\varepsilon_1,\ldots,\varepsilon_n$ は互いに無相関で $E[\varepsilon_\alpha]=0$，$E[\varepsilon_\alpha^2]=\sigma^2$ とする．通常用いられる最小2乗法によって非線形回帰モデルを推定しようとすると，しばしば誤差を過剰に取り込んでしまい，その結果推定曲線あるいは推定曲面は大きく変動し，モデルのデータへの過適合を引き起こしやすいことが知られている．

そこで，モデルの変動が大きくなるにつれて増加する正則化項（罰則項）$R(\boldsymbol{w})$ を誤差の2乗和に課した関数

$$S_\gamma(\boldsymbol{w})=\sum_{\alpha=1}^{n}\left\{y_\alpha-\sum_{j=1}^{m}w_j\phi_j(\boldsymbol{x}_\alpha)\right\}^2+\gamma R(\boldsymbol{w}) \tag{2.4}$$

の最小化によってモデルを推定する方法が正則化最小2乗法である．ここで，$\gamma(>0)$ は，正則化パラメータ，あるいは平滑化パラメータと呼ばれ，モデルの適合度と当てはめたモデルの滑らかさを連続的に調整するとともに，推定量の安定化に寄与する役割を有している．

実際上，非線形回帰モデルの正則化推定に対しては，主としてパラメータの2次形式で表現される正則化項を付与した関数

$$S_\gamma(\boldsymbol{w})=\sum_{\alpha=1}^{n}\left\{y_\alpha-\sum_{j=1}^{m}w_j\phi_j(\boldsymbol{x}_\alpha)\right\}^2+\gamma\boldsymbol{w}^TK\boldsymbol{w} \tag{2.5}$$

の最小化に基づく方法が用いられてきた．ただし，$\boldsymbol{w}$ は，パラメータベクトル $\boldsymbol{w}=(w_1,\ldots,w_m)^T$ とし，K は $m\times m$ 非負値定符号行列とする．第4.1節では，(2.5) 式の正則化最小2乗法に基づく非線形回帰モデリングについて議論する．

次に，目的変数 Y に関するデータ $y_1,\ldots,y_n$ は，互いに独立に確率分布モデル $f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta})$ に従って観測されたとする．このとき，対数尤度関数 $\ell(\boldsymbol{\theta})=\sum_{\alpha=1}^{n}\log f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta})$ に対し

て，正則化項を課した関数

$$\ell_\lambda(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{n\lambda}{2} R(\boldsymbol{\theta}) \tag{2.6}$$

の最大化に基づく方法が正則化最尤法である．第4.2節では，誤差項に正規分布を仮定したガウス型非線形回帰モデルの正則化法に基づくモデルの構成法について述べる．

近年，生命科学，生物学，システム工学などの分野では，データ数に比して次元数が極めて高いデータが観測されるようになり，特に，次元数 >> データ数のもとでの新たな解析手法の開発研究が求められるようになった．このような状況下で，線形回帰モデルの推定と変数選択に一つの方法を提示していると考えられるのが，Tibshirani (1996) によって提唱されたLASSO (Least Absolute Shrinkage and Selection Operator) によるモデルの推定法である．

Lassoは，線形回帰モデルの切片を除く回帰係数の絶対値の和として与えられる L_1 タイプの正則化項を付与した

$$S_\gamma(\boldsymbol{\beta}) = \sum_{\alpha=1}^{n} \left\{ y_\alpha - \beta_0 - \sum_{j=1}^{p} \beta_j x_{\alpha j} \right\}^2 + \gamma \sum_{j=1}^{p} |\beta_j| \tag{2.7}$$

の最小化に基づく推定法である．線形回帰モデルに対しては，その制約の性質からパラメータの一部を完全に0と推定するという特徴を有しており，モデルの推定と変数選択を同時に実行できる手法として注目を集めた．このlassoの研究に端を発して，データ数に比して極めて高次元のデータに基づくモデルの推定と変数選択を目的とした，様々な L_1 タイプの正則化項を課した推定法が提案されつつある．

回帰係数の絶対値の和で表されたlassoの正則化項は，パラメータに関して微分できないことから推定量の解析的導出が難しく，推定の計算アルゴリズムを必要とする．この L_1 タイプの正則化項を課した L_1 型正則化法に基づくモデリングについては，推定アルゴリズムを含めて第5節で詳解する．

2.3 モデルの評価と選択

正則化法によって推定した非線形回帰モデルは，基底関数の個数および平滑化パラメータの値に依存する．非線形構造を適切に捉えるモデルを構築するためには，モデルの平滑化の程度を調整するパラメータの最適な値を，観測されたデータに基づいて選択することが求められる．この問題をモデル選択として捉えて，予測誤差の推定量，さらに情報量およびベイズの観点から様々なモデル評価基準が提唱され，実際問題への適用研究が行われてきた．これらの研究については，小西・北川 (2004), Konishi and Kitagawa (2008) などを参照されたい．

一方，データ数に比して次元数が極めて高いデータに基づくモデリングは，例えば，線形回帰モデルの変数選択を考えると分かるように，すべてのモデルの組み合わせの中から最適な変数の組を選択する方法には限界がある．また，逐次的に選択する方法も考えられるが，この方法では最適な変数の組を必ずしも選ぶとは限らず，選択の不安定性が指摘されている (Breiman (1996))．これに対して，L_1 タイプの正則化項を課した推定法は，適切に γ の値を選択してモデルを推定すると回帰係数の一部を完全に 0 と推定するという特徴を有しており，モデルの推定と変数選択を同時に実行できる．

しかし，データに依存してどのように γ の値を選択すればよいかという問題は，第 6 節で述べるようにいくつかの方法が提唱されてはいるが，研究の進展が待たれる分野である．これは，L_1 型正則化法は，パラメータベクトルの 2 次形式で表される正則化項を課した正則化法と異なり，パラメータに関して微分不可能となること，また，極めて高次元のデータ，さらにはデータ数より次元数が高いデータを分析の対象とすることから，データ数に関する漸近理論の適用が難しいことなどが挙げられる．このような問題に対しては，数理的アプローチに計算アルゴリズムを融合した新たなモデル評価基準の開発研究が必要と思われる．

3. 非線形回帰モデル

本節では，説明変数に関するデータの次元が 1 次元の場合と高次元の場合に分けて，それぞれ 2 次元データへの曲線推定と多次元データへの曲面推定を目的として基底関数に基づく非線形回帰モデルを紹介する．

3.1 曲線推定

3.1.1 フーリエ級数

周期性を有するデータに対してよく用いられる基底関数として，フーリエ級数が挙げられる．説明変数 x (1 次元) に関して，フーリエ級数は次式で表される (日野 (1977))．

$$\phi_0(x) = \frac{1}{\sqrt{T}},$$

$$\phi_j(x) = \begin{cases} \sqrt{\dfrac{2}{T}} \sin \omega_j x, \quad \omega_j = \dfrac{(j+1)\pi}{T} & (j: \text{奇数}) \\ \sqrt{\dfrac{2}{T}} \cos \omega_j x, \quad \omega_j = \dfrac{j\pi}{T} & (j: \text{偶数}) \end{cases} \quad (j = 1, 2, \ldots, m). \tag{3.1}$$

ここで，説明変数の定義域を $[0, T]$ とする．図 1 は，$T = 1$ としたときのフーリエ級数を表したものであり $m = 3$ と設定した．フーリエ級数に基づく非線形回帰モデルは，取り扱いが比較的容易であるため，古くから様々な分野で用いられている．実際，この基底関数を用いた回帰モデルを構成し，周期性を有するデータへの適用を行ったところ，図 2 左の

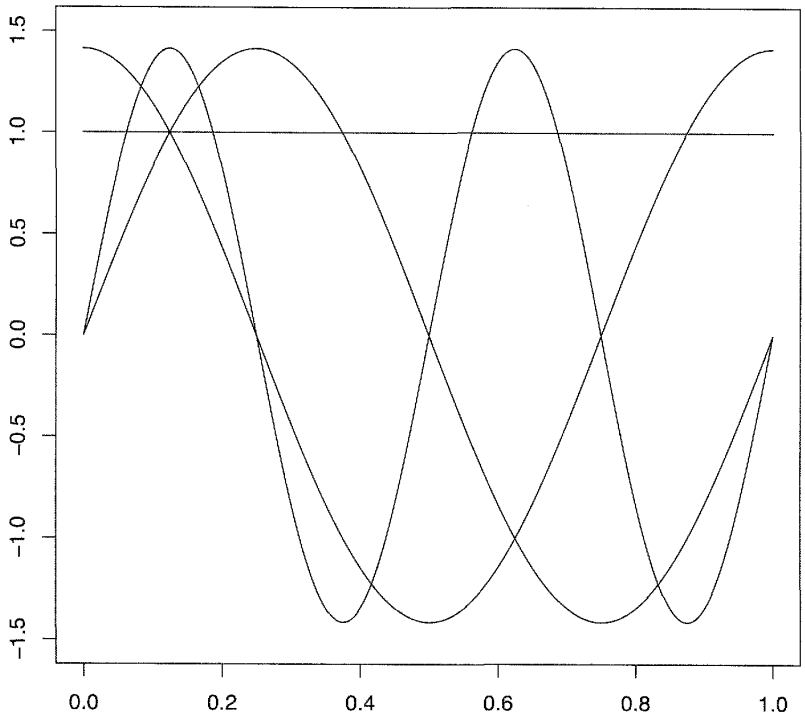

図 1　区間 $[0, 1]$ 上でのフーリエ級数

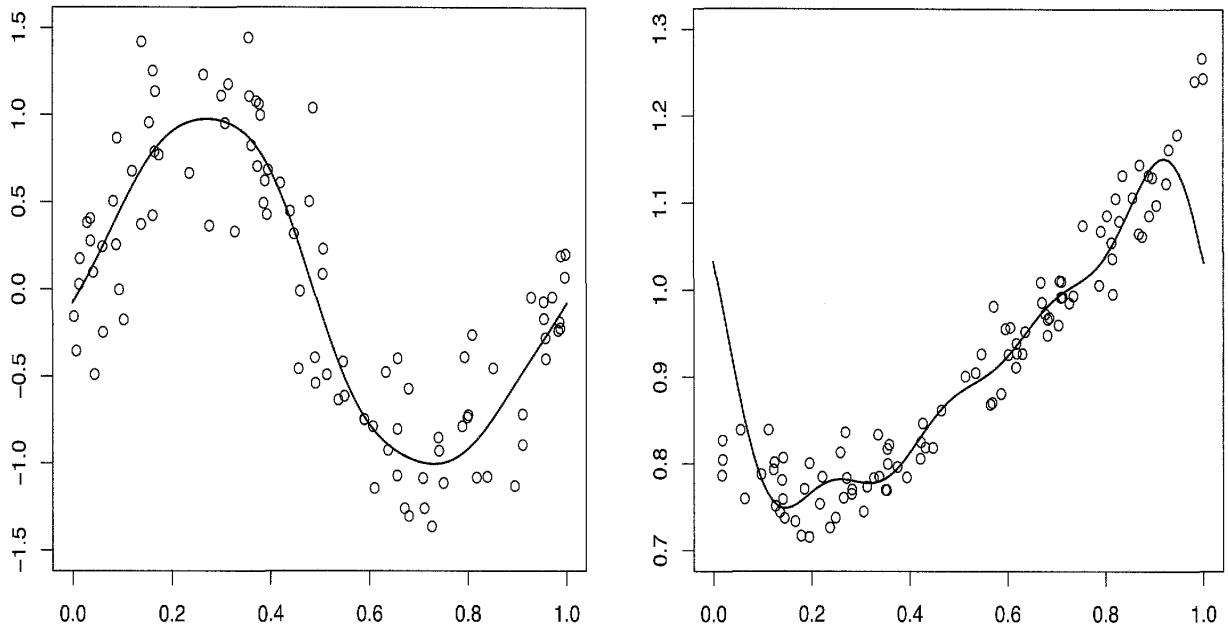

図 2 フーリエ級数に基づく非線形回帰モデル (左図：周期性を有するデータと推定曲線．右図：周期性を有さないデータと推定曲線)

ような回帰曲線を得た．この図より，回帰モデルの推定曲線はデータの構造をよく捉えていることがわかる．

しかしながら，フーリエ級数に基づく回帰モデルを，周期性を有さないデータに対して適用した際には，回帰モデルの推定曲線はデータの構造を適切に捉えきれないことが知られている．実際に，周期性を有さないデータに対してフーリエ級数を用いた回帰モデルの適用を行ったところ，図 2 右のような回帰曲線を得た．この図より，回帰モデルの推定曲

線は境界付近でデータの構造をうまく捉えきれていないのがわかる．

このようにフーリエ級数を用いた回帰モデルは，周期性を有するデータには有効に機能するが，周期性を有さないデータには必ずしも有効に機能しないことがわかる．そこで，次項では多様な非線形構造を捉えることができる基底関数としてスプライン関数について述べる．

3.1.2 スプライン

説明変数 x (1 次元) に関するデータ $x_1, \ldots, x_n$ に対して，n 個のデータは区間 $[a, b]$ 上で次のようにして大きさの順に並んでいるものとする．

$$a < x_1 < x_2 < \cdots < x_{n-1} < x_n < b. \tag{3.2}$$

いま，区間 (x_1, x_n) を分割する m 個の点を $\xi_1 < \xi_2 < \cdots < \xi_m$ と置く．これらの点は節点と呼ばれる．スプラインの基本的な考えは，配置された節点に基づいた各小区間 $[a, \xi_1], [\xi_1, \xi_2], \ldots, [\xi_m, b]$ 上で区分的に多項式をあてはめることにある．

実際上，最もよく用いられているのは 3 次多項式を節点で滑らかに接続した 3 次スプラインである．すなわち，各節点で 2 つの 3 次多項式の 1 次，2 次導関数が連続となるように制約をつけてモデルのあてはめを行ったものであり，この結果，節点 $\xi_1 < \xi_2 < \cdots < \xi_m$ を持つ 3 次スプラインは次式で与えられる (Hastie and Tibshirani (1990)).

$$u(x; \boldsymbol{\theta}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{m} \theta_j (x - \xi_j)_+^3. \tag{3.3}$$

ここで，$\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \theta_1, \theta_2, \ldots, \theta_m)^T$ とし，$(x - \xi_i)_+ := \max\{0, x - \xi_i\}$ は打ち切りベキ関数である．また，基底関数は $\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = x^2, \phi_3(x) = x^3, \phi_4(x) = (x - \xi_1)_+^3, \ldots, \phi_{m+3}(x) = (x - \xi_m)_+^3$ となる．3 次スプライン関数に基づく基底関数を図 3 左に挙げる．ここで，節点は $\xi_j = j$ $(j = 1, \ldots, 10)$ と設定した．

一般に境界付近での 3 次多項式のあてはめは，推定した曲線の変動が大きく適当でないことが知られている (Hastie *et al.* (2009, p.144)). そこで，3 次スプラインに対して両端区間 $[-\infty, \xi_1], [\xi_m, +\infty]$ では 1 次式であるという条件を付加したのが，次の自然 3 次スプラインである (Green and Silverman (1994)).

$$u(x; \boldsymbol{\theta}) = \beta_0 + \beta_1 x + \sum_{j=1}^{m-2} \theta_j \{d_j(x) - d_{m-1}(x)\}. \tag{3.4}$$

ただし，推定すべきパラメータを $\boldsymbol{\theta} = (\beta_0, \beta_1, \theta_1, \theta_2, \ldots, \theta_{m-2})^T$ とし，

$$d_j(x) = \frac{(x - \xi_j)_+^3 - (x - \xi_m)_+^3}{\xi_m - \xi_j} \tag{3.5}$$

とおく．また，基底関数は $\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = d_1(x) - d_{m-1}(x), \ldots, \phi_{m-1}(x) = d_{m-2}(x) - d_{m-1}(x)$ となる．図 3 右は自然 3 次スプライン基底関数を表す．

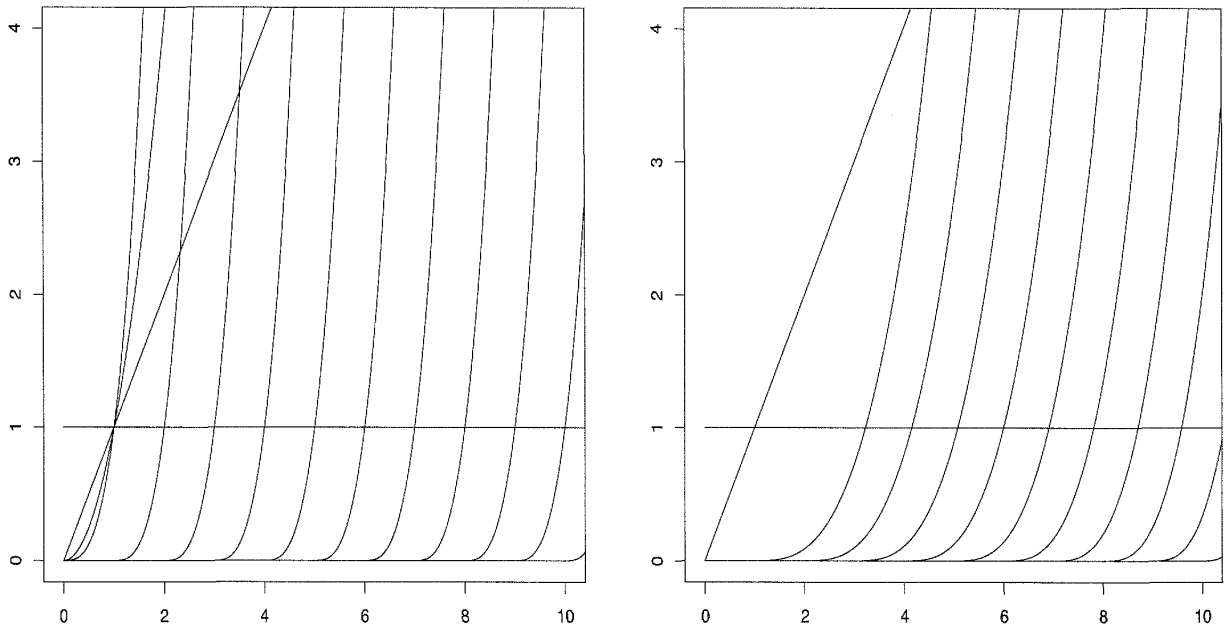

図 3 3 次スプライン関数 (左図) および自然 3 次スプライン関数 (右図)

3 次スプラインおよび自然 3 次スプラインを基底関数として用いる場合, 非線形度に関わる節点の個数と位置の決定は重要な問題となる. 一つの方法として, 節点をすべてのデータ点上に配置したモデルを考えることもできるが, 井元・小西 (1999) によると, このようにして構成されたモデルのパラメータ数は標本数を超え, さらに, データ点の値が重複するとパラメータ推定の際に困難が生じるなどの欠点を持つと報告している. 節点の個数と位置をパラメータとして同時最適化を行う方法も考えられるが, 計算上の困難さを伴うことが多い. これらの研究については, 例えば, Denison *et al.* (1998), Osborne *et al.* (1998), Wand (1999), Miyata and Shen (2005) 等を参照されたい.

スプライン関数については市田・吉本 (1979), Wahba (1990), Schumaker (1993), Eubank (1999) などを, またスプライン関数に基づく非線形回帰モデリングについては坂本ほか (2008) を参照されたい.

3.1.3 B-スプライン

説明変数 x に関するデータは大きさの順に並んでいるものとする. B-スプラインは, 各節点において滑らかに連結する区分的多項式により構成される.

いま, r 次の B-スプライン基底関数 $B_j(x;r)$ を構成するために必要な節点を

$$t_1 < t_2 < \cdots < t_{r+1} = x_1 < \cdots < t_{m+1} = x_n < \cdots < t_{m+r+1} \tag{3.6}$$

とおく. この B-スプライン基底関数を構成するには, de Boor (2001) のアルゴリズムがよく用いられる. まず 0 次の B-スプライン基底関数を次で定義する.

$$B_j(x;0) = \begin{cases} 1, & t_j \le x < t_{j+1}, \\ 0, & \text{その他}. \end{cases} \tag{3.7}$$

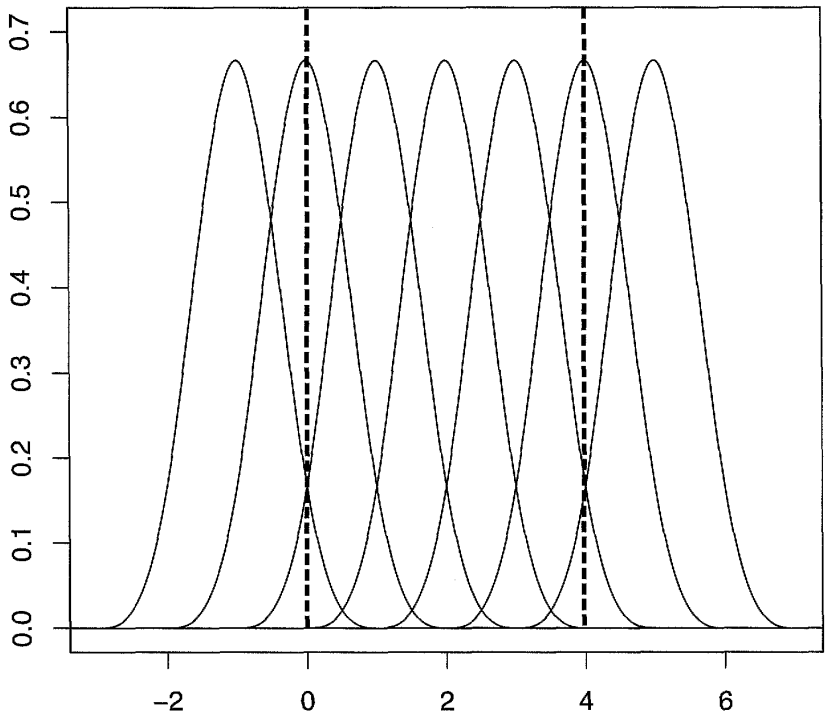

図 4　3 次の B-スプライン関数

この 0 次の B-スプライン基底関数に対して，次の逐次計算により r 次の B-スプライン基底関数を構成する．

$$B_j(x;r) = \frac{x - t_j}{t_{j+r} - t_j} B_j(x;r-1) + \frac{t_{j+r+1} - x}{t_{j+r+1} - t_{j+1}} B_{j+1}(x;r-1). \tag{3.8}$$

r 次の B-スプライン基底関数の主な性質としては，(i) 各区間 $[t_j, t_{j+1}]$ $(j = r+1, \ldots, m)$ は $(r+1)$ 個の r 次多項式からなり，(ii) 節点で C^{r-1} 級関数となることである．B-スプライン関数の詳細については de Boor (2001) を参照されたい．

実際には，節点を等間隔に配置した 3 次の B-スプライン関数を基底関数として用いることが多い (例えば，Eilers and Marx (1996), Marx and Eilers (1998), Imoto and Konishi (2003), Lang and Brezger (2004), Wager *et al.* (2007) 参照)．つまり，節点を等間隔に

$$t_1 < t_2 < t_3 < t_4 = x_1 < t_5 < \cdots < t_m < t_{m+1} = x_n < t_{m+2} < t_{m+3} < t_{m+4}, \tag{3.9}$$

と置き，n 個のデータを $(m-3)$ 個の区間に分割して構成する．すると，各区間 $[t_j, t_{j+1}]$ $(j = 4, \ldots, m)$ はそれぞれ 4 つの基底関数で覆われることになる．図 4 は 3 次の B-スプライン基底関数を示したものである．ここで，節点を $t_1 = -3, t_2 = -2, \ldots, t_{11} = 7$ と等間隔に配置した．図 4 より，$t_4 = 0$ から $t_8 = 4$ までの区間は 4 つの基底関数に覆われているのがわかる．

本節では，データを発生する真の構造が滑らかな関数の場合を考えてきた．しかし，得られるデータの構造が局所的に大きく変化して，ピークなどを有していることも少なくな

い．このような変化を捉えるための基底関数として，ウェーブレットを用いることが考えられる．ウェーブレットを基底関数に用いた回帰モデルとしては，Donoho and Johnston (1995)，Hall and Patil (1996)，Antoniadis and Fan (2001)，Fujii and Konishi (2006) などが挙げられる．

3.2 高次元曲面推定

3.2.1 テンソル積による多次元スプライン

いま，p 次元説明変数 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ に関して，n 個のデータ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ が観測されたとする．高次元データに対する基底関数としては，まずテンソル積を用いた多次元スプラインが考えられる (Green and Silverman (1994), Hastie *et al.* (2009))．

テンソル積に基づく基底関数は，次の手順に従って構成される．説明変数 x_1 に関して，基底関数 $h_{k_1}(x_1)$ $(k_1 = 1, \ldots, m_1)$ を用意する．次に，説明変数 x_2 に関して，基底関数 $h_{k_2}(x_2)$ $(k_2 = 1, \ldots, m_2)$ を用意する．同様に，p 変量すべての説明変数に対して基底関数を用意する．このとき，テンソル積を用いることによって，$m_1 \times m_2 \times \cdots \times m_p$ 次元多次元スプライン関数を次で定義する．

$$\phi_{k_1 k_2 \ldots k_p}(\boldsymbol{x}) = h_{k_1}(x_1) h_{k_2}(x_2) \cdots h_{k_p}(x_p), \quad (k_j = 1, \ldots, m_j;\ j = 1, \ldots, p). \tag{3.10}$$

この基底関数を用いることにより，説明変数が高次元の場合における回帰モデルは，

$$y_\alpha = w_0 + \sum_{k_1=1}^{m_1} \sum_{k_2=1}^{m_2} \cdots \sum_{k_p=1}^{m_p} w_{k_1 k_2 \ldots k_p} \phi_{k_1 k_2 \ldots k_p}(\boldsymbol{x}_\alpha) + \varepsilon_\alpha, \quad \alpha = 1, \ldots, n \tag{3.11}$$

で与えられる．ここで，w_0, $w_{k_1 k_2 \ldots k_p}$ $(k_j = 1, \ldots, m_j;\ j = 1, \ldots, p)$ がモデルのパラメータである．

しかし，テンソル積による多次元スプラインでは，説明変数の個数が多くなるに従って，用意しなければならない基底関数の個数 (節点の個数) が指数上に増加していき，いわゆる「次元の呪い」が生じる．また，データが存在していない部分にも基底関数が配置される可能性もあり，一般に説明変数の次元が 3 次元以上の場合には取り扱いは困難になってくる．これらの問題点を克服するため，Friedman (1991) は多変量適応的回帰スプライン (MARS: Multivariate Adaptive Regression Splines) を提案している．MARS では，必要と考えられる基底関数を何段階にもわたる複雑なアルゴリズムを用いて選択していく．詳しくは Friedman (1991) を参照されたい．

3.2.2 動径基底関数

いま，p 次元説明変数ベクトル $\boldsymbol{x}$ に関する n 個のデータを $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ とする．動径基底関数は，一般に p 次元ベクトル $\boldsymbol{x}$ と基底関数の位置を定める p 次元中心ベクトル $\boldsymbol{\mu}$ との間のユークリッド距離 $z = ||\boldsymbol{x} - \boldsymbol{\mu}||$ に依存するある非線形関数 $\phi(z)$ を用いて定式化される．

この基底関数に基づく回帰モデルは

$$y_\alpha = w_0 + \sum_{j=1}^{m} w_j \phi\left(||\boldsymbol{x}_\alpha - \boldsymbol{\mu}_j||\right) + \varepsilon_\alpha, \quad \alpha = 1, \ldots, n \tag{3.12}$$

で与えられる (Bishop (1995), Ripley (1996)). 基底関数として実際上よく用いられるのは，次のガウス型基底関数である.

$$\phi_j(\boldsymbol{x}) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_j||^2}{2h_j^2}\right), \quad j = 1, \ldots, m. \tag{3.13}$$

ここで，h_j^2 は関数の広がりの程度を表す量である．この他の動径基底関数としては，thin plate spline 関数 $\phi(z) = z^2 \log z$ や multiquadrics 関数 $\phi(z) = \sqrt{z^2 + c^2}$ $(c > 0)$ などが知られている．動径基底関数に関しては，佐藤 (1996) 及び Buhmann (2009) などを参照されたい.

ガウス型基底関数に基づく非線形回帰モデルの未知パラメータは，係数パラメータ $\{w_0, w_1, \ldots, w_m\}$ に加えて基底関数に含まれる $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m, h_1^2, \ldots, h_m^2\}$ である．これらのパラメータを同時に推定する方法も考えられるが (例えば，Leitenstorfer and Tutz (2007) 参照)，推定の一意性や数値的最適化における局所解の問題等が生じ，また基底関数の個数の選択も考慮に入れたとき，計算時間が膨大になることが予想される．このような問題を克服し応用上有用な手法として，まず説明変数に関するデータから基底関数を事前に決定して，既知の基底関数をもつモデルをデータに当てはめる 2 段階推定法が用いられる．その一つの方法として，クラスタリング手法を適用して基底関数を事前に決定する方法がある (Moody and Darken (1989), Ando *et al.* (2008)).

この方法は，n 個の説明変数に関するデータ $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ を，例えば，k-平均法によって基底関数の個数に相当する m 個のクラスタ $C_1, \ldots, C_m$ に分割し，各クラスタ C_j に含まれる n_j 個のデータに基づいて中心ベクトル $\boldsymbol{\mu}_j$ と h_j^2 を次のように決定する.

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n_j} \sum_{\boldsymbol{x}_\alpha \in C_j} \boldsymbol{x}_\alpha, \quad \hat{h}_j^2 = \frac{1}{n_j} \sum_{\boldsymbol{x}_\alpha \in C_j} ||\boldsymbol{x}_\alpha - \hat{\boldsymbol{\mu}}_j||^2. \tag{3.14}$$

これらの推定値を (3.13) 式のガウス型基底関数に代入して

$$\phi_j(\boldsymbol{x}) = \exp\left(-\frac{||\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j||^2}{2\hat{h}_j^2}\right), \quad j = 1, \ldots, m \tag{3.15}$$

を j 番目の基底関数として用いる．このとき，ガウス型基底関数に基づく非線形回帰モデルは，

$$y_\alpha = w_0 + \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}_\alpha) + \varepsilon_\alpha, \quad \alpha = 1, \ldots n \tag{3.16}$$

として与えられる．

ガウス型基底関数に基づく回帰モデルについては，Bishop (1991), Konishi *et al.* (2004), Kawano and Konishi (2007), Ando *et al.* (2008) などを，また，thin plate spline 関数に基づく回帰モデルについては，Wahba (1983), McCaffery *et al.* (1992), Gu (2002), Wood (2003, 2004) などを参照されたい．

4. モデルの推定

本節では，基底関数 $\phi_j(\boldsymbol{x})$ $(j=1,\ldots,m)$ に基づく回帰モデル

$$y_\alpha = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}_\alpha) + \varepsilon_\alpha, \quad \alpha = 1,2,\ldots,n \tag{4.1}$$

に含まれるパラメータ w_j $(j=1,\ldots,m)$ の推定方法について考える．ここで，ε_α は互いに無相関で，$E[\varepsilon_\alpha]=0$, $E[\varepsilon_\alpha^2]=\sigma^2$ とする．

以降，(4.1) 式の回帰モデルを与える n 個の式をベクトルと行列を用いて次のように表記しておく．

$$\boldsymbol{y} = \Phi \boldsymbol{w} + \boldsymbol{\varepsilon}. \tag{4.2}$$

ここで，$\boldsymbol{y}=(y_1,\ldots,y_n)^T$ は n 次元観測値ベクトル，Φ はその第 (a,b) 要素を $\phi_b(\boldsymbol{x}_a)$ $(a=1,\ldots,n,\ b=1,\ldots,m)$ とする $n\times m$ 基底関数行列，$\boldsymbol{w}=(w_1,\ldots,w_m)^T$ は m 次元パラメータベクトル，$\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)^T$ は n 次元誤差ベクトルである．

4.1 正則化最小 2 乗法

基底展開法に基づく非線形回帰モデルを想定する場合，基底関数の増加は，パラメータ数の増加に伴うモデルのデータへの過適合と推定の不安定性を引き起こすおそれがある．このような場合には，誤差の 2 乗和に，正則化項 $R(\boldsymbol{w})$ を課した関数

$$S_\gamma(\boldsymbol{w}) = (\boldsymbol{y}-\Phi\boldsymbol{w})^T(\boldsymbol{y}-\Phi\boldsymbol{w}) + \gamma R(\boldsymbol{w}) \tag{4.3}$$

の最小化によってパラメータを推定する．正則化項 $R(\boldsymbol{w})$ としては，関数の曲率を考慮した 2 階微分の積分の離散近似

$$R_1(\boldsymbol{w}) = \sum_{\alpha=1}^{n}\sum_{i=1}^{p}\left\{\frac{\partial^2\{\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}{\partial x_i^2}\right\}^2 \tag{4.4}$$

や，係数パラメータ $\boldsymbol{w}$ の差分 $R_2(\boldsymbol{w})=\sum_{j=k+1}^{m}(\delta^k w_j)^2$，2 乗和 $R_3(\boldsymbol{w})=\sum_{j=1}^{m} w_j^2$ 等が用いられる．ただし，δ は差分作用素 $\delta w_j = w_j - w_{j-1}$ を表す．

ここでは，正則化項 $R(\boldsymbol{w})$ は $m\times m$ 非負値定符号行列 K を用いて，パラメータベクトル $\boldsymbol{w}$ の 2 次形式 $\boldsymbol{w}^TK\boldsymbol{w}$ と表すことができるものとする．$R_3(\boldsymbol{w})$ は，単位行列 I に対

して，$R_3(\boldsymbol{w}) = \boldsymbol{w}^T I \boldsymbol{w}$ と表すことができ，差分に基づく正則化項 $R_2(\boldsymbol{w})$ は，差分行列 $K = D_k^T D_k$ に対して，$R_2(\boldsymbol{w}) = \boldsymbol{w}^T K \boldsymbol{w}$ と表すことができる．ただし，D_k は $(m-k) \times m$ 行列で，次式で与えられる．

$$D_k = \begin{bmatrix} {}_kC_0 & -{}_kC_1 & \dots & (-1)^k\,{}_kC_k & 0 & \dots & 0 \\ 0 & {}_kC_0 & -{}_kC_1 & \dots & (-1)^k\,{}_kC_k & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & {}_kC_0 & -{}_kC_1 & \dots & (-1)^k\,{}_kC_k \end{bmatrix}.$$

実際上よく用いられるのは，次で与えられる 2 次差分である．

$$D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -2 & 1 \end{bmatrix}. \tag{4.5}$$

特に，2 次差分の正則化項 $\boldsymbol{w}^T D_2^T D_2 \boldsymbol{w}$ に対して 3 次の B-スプライン関数を基底関数として用いたとき，Eilers and Marx (1996) は，次の近似が成り立つことを示している．

$$\boldsymbol{w}^T D_2^T D_2 \boldsymbol{w} \approx \int \left\{ \frac{\partial^2 \{\boldsymbol{w}^T \boldsymbol{\phi}(x)\}}{\partial x^2} \right\}^2 dx. \tag{4.6}$$

2 次形式で表わされる正則化項を用いると，パラメータベクトル $\boldsymbol{w}$ の正則化最小 2 乗推定量は，

$$S_\gamma(\boldsymbol{w}) = (\boldsymbol{y} - \Phi \boldsymbol{w})^T (\boldsymbol{y} - \Phi \boldsymbol{w}) + \gamma \boldsymbol{w}^T K \boldsymbol{w} \tag{4.7}$$

を最小とする解として，次式で与えられる．

$$\hat{\boldsymbol{w}} = (\Phi^T \Phi + \gamma K)^{-1} \Phi^T \boldsymbol{y}. \tag{4.8}$$

4.2 正則化最尤法

誤差項 ε_α に対して，互いに独立に平均 0，分散 σ^2 の正規分布を仮定した基底展開法に基づく回帰モデルの当てはめを考える．すなわち，次の非線形回帰モデルを想定する．

$$\boldsymbol{y} = \Phi \boldsymbol{w} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n). \tag{4.9}$$

このとき，観測値ベクトル $\boldsymbol{y}$ は，平均ベクトル $\Phi \boldsymbol{w}$，分散共分散行列 $\sigma^2 I_n$ の n 次元正規分布に従うことから，その確率密度関数は

$$f(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y} - \Phi \boldsymbol{w})^T (\boldsymbol{y} - \Phi \boldsymbol{w}) \right\}$$

で与えられ，対数尤度関数は

$$\begin{aligned}\ell(\boldsymbol{w},\sigma^2) &= \log f(\boldsymbol{y}|\boldsymbol{w},\sigma^2)\\ &= -\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}(\boldsymbol{y}-\Phi\boldsymbol{w})^T(\boldsymbol{y}-\Phi\boldsymbol{w}) \qquad (4.10)\end{aligned}$$

となる．

複雑な現象を近似するために多数の基底関数を用いた非線形モデルを当てはめるとき，モデルのパラメータを最尤法によって推定すると，しばしばデータに過度に依存したモデルが推定される．これは，構造を近似するモデルがデータの近くを通るにつれて，対数尤度関数 $\ell(\boldsymbol{w},\sigma^2)$ の値が大きくなるためである．

そこで，モデルの滑らかさが失われるにつれて増加する正則化項を対数尤度関数に課した

$$\ell_\lambda(\boldsymbol{w},\sigma^2)=\ell(\boldsymbol{w},\sigma^2)-\frac{n\lambda}{2}R(\boldsymbol{w}) \qquad (4.11)$$

の最大化によってパラメータの推定を行う．ここで，$\lambda\,(>0)$ は平滑化パラメータであり，モデルの当てはまりのよさと曲線の滑らかさを制御するハイパーパラメータである．この関数は正則化対数尤度関数と呼ばれ，その最大化による推定法は，正則化最尤法あるいは罰則付き最尤法と呼ばれる．

ガウスノイズを仮定した非線形回帰モデルに対して，2 次形式の正則化項を課した正則化対数尤度関数は

$$\ell_\lambda(\boldsymbol{w},\sigma^2)=\ell(\boldsymbol{w},\sigma^2)-\frac{n\lambda}{2}\boldsymbol{w}^TK\boldsymbol{w} \qquad (4.12)$$

と表され，モデルのパラメータ $\boldsymbol{w}$ と σ^2 の正則化最尤推定量は

$$\hat{\boldsymbol{w}}=(\Phi^T\Phi+n\lambda\hat{\sigma}^2K)^{-1}\Phi^T\boldsymbol{y},\quad \hat{\sigma}^2=\frac{1}{n}(\boldsymbol{y}-\Phi\hat{\boldsymbol{w}})^T(\boldsymbol{y}-\Phi\hat{\boldsymbol{w}}) \qquad (4.13)$$

で与えられる．

ここで，$\hat{\boldsymbol{w}}$ と $\hat{\sigma}^2$ は互いに依存しているので，実際上，次のようなアルゴリズムで計算する．

Step1: 初期値を $\sigma^{2(0)}$ とし，λ をあらかじめ与えておく．$k=1$ とする．

Step2: $\boldsymbol{w}^{(k)}=(\Phi^T\Phi+n\lambda\sigma^{2(k-1)}K)^{-1}\Phi^T\boldsymbol{y}$ で更新する．

Step3: $\sigma^{2(k)}=(\boldsymbol{y}-\Phi\boldsymbol{w}^{(k)})^T(\boldsymbol{y}-\Phi\boldsymbol{w}^{(k)})/n$ を求め，十分小さな δ に対して，$(\sigma^{2(k)}-\sigma^{2(k-1)})^2<\delta$ を評価する．この条件を満たすまで **Step2** を繰り返し，条件を満たしたパラメータを推定値とする．

5. L_1 型正則化推定法

近年，L_1 型正則化による推定方法が様々な分野で注目を集めている．L_1 型正則化による推定法では，パラメータの推定と変数の選択を同時に実行でき，生命科学を含む様々な分野でその有効性が実証されつつある．本節では，はじめに線形回帰モデルの枠組みにおいて縮小推定法と制約付き最小化問題の関係性について述べる．次に，L_1 正則化の中でも最も基礎となる lasso について述べたあと，モデルを推定するための種々のアルゴリズムを紹介する．さらに，lasso の基本的な考え方に基づくいくつかの縮小推定法を紹介する．

5.1 縮小推定法と制約

目的変数 Y と p 個の説明変数 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ に関して n 組のデータ $\{(y_\alpha, \boldsymbol{x}_\alpha);\ \alpha = 1, 2, \ldots, n\}$ が観測されたとする．このとき，次の線形回帰モデルを想定する．

$$y_\alpha = \beta_0 + \sum_{j=1}^{p} \beta_j x_{\alpha j} + \varepsilon_\alpha, \quad \alpha = 1, 2, \ldots, n. \tag{5.1}$$

ここで，ε_α は互いに無相関で，$E[\varepsilon_\alpha] = 0$, $E[\varepsilon_\alpha^2] = \sigma^2$ とする．この線形回帰モデルは，(4.2) 式における係数パラメータベクトル $\boldsymbol{w}$ と基底関数行列 Φ をそれぞれ $(p+1)$ 次元係数パラメータベクトル $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ と $n \times (p+1)$ 計画行列 X で置き換えることによって次の式で表わされる．

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{5.2}$$

この線形回帰モデルにおいて，2 乗和誤差関数に係数パラメータの 2 乗和 $R(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$ を正則化項として課した関数の最小化によって係数パラメータベクトル $\boldsymbol{\beta}$ の推定を行う方法は，リッジ回帰 (Hoerl and Kennard (1970)) と呼ばれる．さらに，リッジ回帰の正則化項を一般化した目的関数

$$S_\gamma(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \gamma \sum_{j=1}^{p} |\beta_j|^q \tag{5.3}$$

の最小化に基づく推定法が考えられ，ブリッジ回帰 (Frank and Friedman (1993), Fu (1998)) と呼ばれる．また，(5.3) 式は，次の条件付き最小化問題と同等であることが示される．

$$\min_{\beta} \left\{(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})\right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j|^q \leq \eta. \tag{5.4}$$

特に $q = 1$ とした L_1 ノルムの制約を課した推定法は，lasso と呼ばれる (Tibshirani (1996))．

5.2 Lasso

Lasso は，2 乗和誤差関数に L_1 ノルムの制約を課した関数

$$S_\gamma(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \gamma \sum_{j=1}^{p} |\beta_j| \tag{5.5}$$

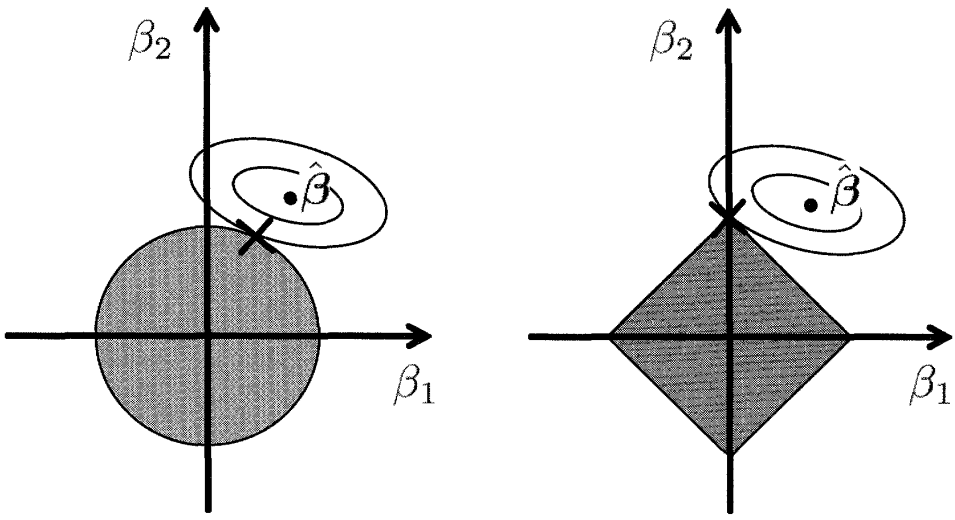

図 5 $p=2$ の場合における，誤差関数の等高線表示 (楕円) と制約領域の図．リッジ推定では円形 (左)，lasso 推定では四角形 (右) の制約条件を満たす領域となる．

の最小化によってモデルのパラメータを推定する方法で，いくつかの係数パラメータを真に 0 と縮小する推定法である．図 5 に示すように，リッジ推定が回帰係数の推定値を 0 へと縮小するのに対して，lasso 推定はいくつかの回帰係数を真に 0 と推定するという違いがある．リッジ推定では，最小 2 乗推定値 $\hat{\boldsymbol{\beta}}$ と比べ，β_1 と β_2 が共に 0 へ縮小されているのに対し，lasso 推定では，β_1 の推定値が真に 0 へ縮小されている．

リッジ推定は制約がパラメータに関して微分可能なため，推定量が閉じた形で与えられるが，lasso 推定における L_1 ノルムの制約は微分不可能なため，推定量の解析的な導出が困難となる．そこで，lasso 推定量を近似的に導出する方法が提案されている．ここでは，Fu (1998) によって提案された shooting アルゴリズムと Efron *et al.* (2004) によって提案された LARS (Least Angle Regression) について紹介する．さらに，ベイズ的に lasso 推定量を算出する手法 (Park and Cassella (2008)) について述べる．

5.2.1 Shooting アルゴリズム

Fu (1998) は，(5.4) 式の $q>1$ の場合に対するブリッジ推定量を算出するために MNR (modified Newton-Raphson) アルゴリズムを構築した．この MNR アルゴリズムの考えに基づいて，q を 1 へ限りなく近づけることによって，lasso 推定量を数値的に算出する逐次計算法を提案し，shooting アルゴリズムと呼んだ．計算アルゴリズムは次で与えられる．

Step1: 初期値 $\hat{\boldsymbol{\beta}}_{old}=(\hat{\beta}_{0,old},\ldots,\hat{\beta}_{p,old})^T=\hat{\boldsymbol{\beta}}_{ols}$ とする．ただし，$\hat{\boldsymbol{\beta}}_{ols}$ は最小 2 乗推定量である．

Step2: $j=0,\ldots,p$ に対して，$\beta_{j,old}=0$ のときは $\beta_{j,new}=0$ とし，$\beta_{j,old}\neq 0$ のときは

$$\hat{\beta}_{0,new}=\frac{-S_0}{2n},\qquad \hat{\beta}_{j,new}=\begin{cases}\dfrac{\gamma-S_0}{2\boldsymbol{x}_j^T\boldsymbol{x}_j} & (S_0>\gamma),\\ \dfrac{-\gamma-S_0}{2\boldsymbol{x}_j^T\boldsymbol{x}_j} & (S_0<-\gamma),\quad (j=1,\ldots,p),\\ 0 & (|S_0|\leq\gamma),\end{cases}$$

とする．ただし，$\boldsymbol{x}_j$ は計画行列 X の第 j 列，$\hat{\boldsymbol{\beta}}_{-j,old}=(\hat{\beta}_{0,new},\hat{\beta}_{1,new},\ldots,\hat{\beta}_{j-1,new},0,\hat{\beta}_{j+1,old},\ldots,\hat{\beta}_{p,old})^T$，$S_j(\boldsymbol{\beta},X,\boldsymbol{y})=\partial(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})/\partial\beta_j$，$S_0=S_j(\hat{\boldsymbol{\beta}}_{-j,old},X,\boldsymbol{y})$ とする．これを繰り返し，$\hat{\boldsymbol{\beta}}_{new}=(\hat{\beta}_{0,new},\ldots,\hat{\beta}_{p,new})^T$ を得る．

Step3: Step2 を収束するまで繰り返す．

5.2.2 LARS アルゴリズム

Efron *et al.* (2004) は，線形回帰モデルの変数選択のアルゴリズムとして LARS アルゴリズム (Least angle regression) を提唱した．LARS アルゴリズムによって得られた推定値は lasso 推定値とかなり似ており，さらに LARS アルゴリズムを少し修正することにより lasso 推定値を求めることができる（本稿では，これを LARS-lasso アルゴリズムと呼ぶ）．LARS-lasso アルゴリズムの特徴の 1 つとして極めて高速であるということがあり，例えば先に述べた shooting アルゴリズムよりはるかに効率的である．Shooting アルゴリズムが，γ の候補をいくつか与え，それぞれの候補に対して反復計算を行って推定値を求めるのに対して，LARS-lasso は γ の候補を自動的に選択する．さらに 1 回のステップで変数を追加・削除して推定値を求めることができるため，ほぼ p 回で全てのステップを完了することができる．

LARS アルゴリズムについて述べる前に，lasso 推定値の性質について考察する．まず，

$$\sum_{\alpha=1}^{n}y_\alpha=0,\quad \sum_{\alpha=1}^{n}x_{\alpha j}=0,\quad \sum_{\alpha=1}^{n}x_{\alpha j}^2=1,\qquad j=1,\ldots,p \tag{5.6}$$

となるように $\boldsymbol{y}$ を中心化し，X を基準化する．この中心化，基準化は LARS アルゴリズムを行う上で重要となる．また，一般性を失うことなく，$\beta_0=0$ として $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$ とおく．

Lasso 推定値は (5.5) 式の最小化によって求めることができるが，γ の値が大きければいくつかの係数の推定値は 0 となる．そこで，推定値が 0 でない添字集合 $\mathcal{A}$ を

$$\mathcal{A}=\{j\in\{1,\ldots,p\}:\ \hat{\beta}_j^{\text{lasso}}\neq 0\} \tag{5.7}$$

と定義すると，lasso 推定値 $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ は次の性質を満たすことが分かる．

$$\boldsymbol{x}_j^T(\boldsymbol{y}-X\hat{\boldsymbol{\beta}}^{\text{lasso}})=\frac{\gamma}{2}\cdot\text{sign}(\hat{\beta}_j^{\text{lasso}}),\quad \forall j\in\mathcal{A}. \tag{5.8}$$

この式は，任意の $j \in \mathcal{A}$ に対し，説明変数 $\boldsymbol{x}_j$ と残差 $(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{lasso}})$ の相関 (内積) の絶対値が等しいことを意味する．ゆえに，lasso 推定は次のような特徴を有することが分かる．

まず，$\gamma = \infty$ のときは，すべての係数の推定値が 0 となっている．γ の値を連続的に小さくして推定値を計算すると，0 でない係数 $\hat{\beta}_{j_1}^{\text{lasso}}$ が出てくる．さらに γ の値を小さくすると，0 でない係数 $\hat{\beta}_{j_2}^{\text{lasso}}$ が出現し，このとき

$$|\boldsymbol{x}_{j_1}^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{lasso}})| = |\boldsymbol{x}_{j_2}^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{lasso}})| \tag{5.9}$$

をみたす．このまま γ を小さくしていくと，$\hat{\beta}_{j_1}^{\text{lasso}}, \hat{\beta}_{j_2}^{\text{lasso}}$ が 0 にならない限り (5.9) 式が保たれ，その相関の絶対値は単調減少する．さらに γ を小さくしても $\hat{\beta}_{j_1}^{\text{lasso}}, \hat{\beta}_{j_2}^{\text{lasso}}$ が 0 にならないとすると，推定値が 0 でない係数 $\hat{\beta}_{j_3}^{\text{lasso}}$ が出現して，

$$|\boldsymbol{x}_{j_1}^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{lasso}})| = |\boldsymbol{x}_{j_2}^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{lasso}})| = |\boldsymbol{x}_{j_3}^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{lasso}})|$$

をみたす．このように，相関の絶対値の等しい方向に予測値を動かすアルゴリズムが LARS アルゴリズムである．

次に，LARS アルゴリズムを図 6 の $p = 2$ の場合の図を使って解説する．まず，最小 2 乗推定値を $\bar{\boldsymbol{y}}_2$ とおくと，任意の $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$ と書ける推定値に対し，

$$X^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = X^T(\bar{\boldsymbol{y}}_2 - \hat{\boldsymbol{\mu}}) \tag{5.10}$$

が成り立つことに注目しておく．

まず，予測値を初期値 $\hat{\boldsymbol{\mu}}_0 = \boldsymbol{0}$ とし，最小 2 乗推定値に近づけることを考える．ここで，図 6 より，説明変数と残差との相関の絶対値は $\boldsymbol{x}_1$ の方が大きい，すなわち，$|\boldsymbol{x}_1^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)| > |\boldsymbol{x}_2^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)|$ が成り立つ．LARS アルゴリズムでは，相関の絶対値の大きい $\boldsymbol{x}_1$ の方向に予測値を動かす．ステップワイズ法 (Weisberg (1980, 8.5 節) 参照) では予測値を $\bar{\boldsymbol{y}}_1 = |\boldsymbol{x}_1^T\boldsymbol{y}|\boldsymbol{x}_1$ まで動かすが，LARS ではその手前の $\hat{\boldsymbol{\mu}}_1$ まで動かす．ここでは説明変数 $\boldsymbol{x}_1$ と残差 $\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1$ との相関の絶対値と，$\boldsymbol{x}_2$ と残差 $\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1$ との相関の絶対値が等しい．すなわち，$|\boldsymbol{x}_1^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)| = |\boldsymbol{x}_2^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)|$ をみたす．

次に，$\hat{\boldsymbol{\mu}}_1$ から予測値を図 6 のように $\bar{\boldsymbol{y}}_2$ の方向，すなわち角度が等しくなる方向に方向転換する．これは，$|\boldsymbol{x}_1^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}})| = |\boldsymbol{x}_2^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}})|$ を保ったまま予測値を動かすことと同等である．$p = 2$ のときは次のステップで最小 2 乗推定値に達するが，$p \geq 3$ のときは角度が等しくなる方向にさらに方向転換して予測値を動かすという操作を繰り返す．

LARS アルゴリズムで得られた推定値と lasso 推定値はかなり似ているが，必ずしも完全に一致するとは限らない．なぜなら，lasso 推定においては，0 でない係数 β_j に対し，γ を連続的に小さくする過程で $\hat{\beta}_j^{\text{lasso}}$ が 0 に達したとき，添字 j に対して (5.8) 式が成り立たなくなるので予測値ベクトルの方向が変わることがあるが，LARS アルゴリズムにおいて

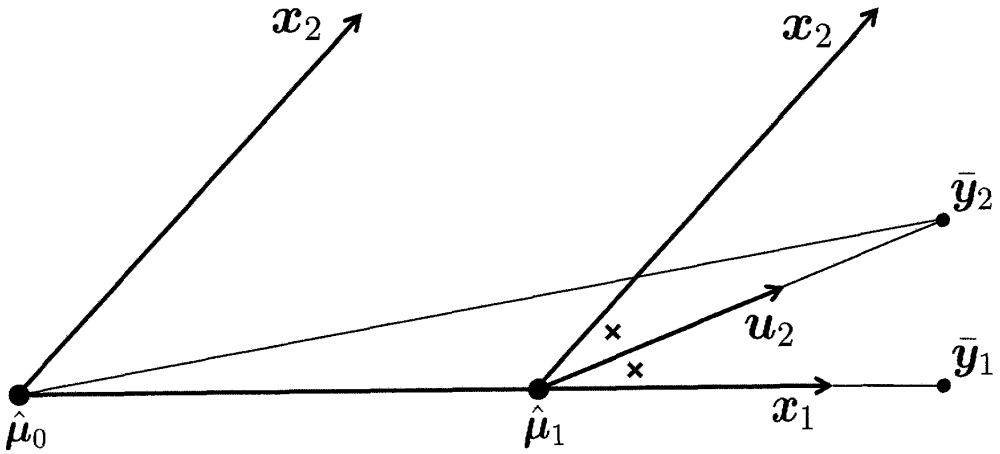

図 6　$p = 2$ としたときの LARS アルゴリズムのイメージ図.

はこのような現象が生じたとしても予測値ベクトルが同じ方向のまま進むためである．しかし，0 でなかった係数が γ を連続的に小さくする過程において再び 0 に達したとき，その対応する変数を取り除いた上で予測値の進むべき方向を再計算するように LARS アルゴリズムを修正することにより，lasso 推定値を求めることができる．

5.2.3　ベイジアン lasso

Lasso 推定は，ベイズ的観点からみると，係数パラメータの事前分布としてラプラス分布を設定したときの最大事後確率 (MAP; maximum a posteriori) 推定量と考えることができる．そこで，Park and Cassella (2008) は，次の式を用いてラプラス分布を正規分布で近似し，マルコフ連鎖モンテカルロ法の 1 つであるギブス・サンプラーアルゴリズムを用いて推定量を算出している．

$$\frac{\lambda}{2}\exp\left(-\lambda|\beta|\right) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{\beta^2}{2s}\right)\cdot\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2 s}{2}\right)ds, \quad \lambda > 0. \tag{5.11}$$

5.3　様々な L_1 型正則化法に基づくモデル

近年 lasso 推定の有効性から，L_1 制約を拡張した新たな縮小推定法が種々提案されている．本節では，これまで提案されてきた様々な L_1 型正則化法に基づくモデルを紹介する．また，lasso は主に線形回帰モデルに対して用いられるが，非線形回帰モデルに対して lasso を適用した研究としては，Osborne *et al.* (1998)，Antoniadis *et al.* (2010)，Tateishi *et al.* (2010) などが挙げられる．

5.3.1　Elastic net

いま，説明変数の次元数 p が標本数 n より大きい "$p > n$" の場合を考える (以降，このようなデータを高次元小標本データと呼ぶ)．この高次元小標本データの解析は，現在多くの研究者が取り組んでいる重要な問題の一つである．

高次元小標本データに対して変数選択問題を考える場合，総当たり法では計算時間的に困難となり，変数増減法では最適な変数を選ぶとは限らず不安定になることが報告されている (Breiman (1996))．そこで，推定と変数選択を同時に行うことができる lasso は，高次元小標本データに対する変数選択を行う上で有用な手法である．しかし，"$p > n$" の場合 lasso は高々 n 個の変数までしか選択できないことが知られている (Efron *et al.* (2004), Rosset *et al.* (2004))．また，遺伝子データのような説明変数間に高い相関性を有しているデータの変数選択を考えるとき，lasso ではこの相関を捉えきれず適切な変数選択が行われるとは限らないことが知られている．

上記の問題点を克服するため，Zou and Hastie (2005) は次の目的関数を最小化する Elastic net と呼ばれる推定法を提唱した．

$$(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \gamma \sum_{j=1}^{p} \{\alpha|\beta_j| + (1-\alpha)\beta_j^2\}. \tag{5.12}$$

ここで，γ (>0) および α $(0 \leq \alpha \leq 1)$ は正則化パラメータである．罰則項に注目すると，これは L_1 正則化と L_2 正則化をあわせた形になっているため，Elastic net はリッジ推定と lasso 推定の両方の性質を有することがわかる．Elastic net は lasso の形に書き直すことができるため (Zou and Hastie (2005, p.304))，推定アルゴリズムには lasso で用いられるアルゴリズムが使える．Elastic net の理論的考察については Yuan and Lin (2007) および Zou and Zhang (2009) を，一般化線形モデルへの拡張については Park and Hastie (2007) および Friedman *et al.* (2009) を参照されたい．

また，Elastic net によるパラメータの制約領域は，ブリッジ回帰

$$(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \gamma \sum_{j=1}^{p} |\beta_j|^q \tag{5.13}$$

における $q \in [1,2]$ のときの制約領域とほぼ同じような領域となる．例えば，$q = 1.2$, $\alpha = 0.8$ としたときのパラメータの制約領域を図 7 に示す．図 7 より，2 つの制約領域は見た目では判断がつかないほどかなり似ている．しかし，Elastic net は lasso の性質を有しているため変数を 0 に縮小させることが可能だが，ブリッジ回帰では真に 0 への縮小は難しいことに注意されたい．

5.3.2 Adaptive lasso

前述のように，lasso はパラメータの推定と変数選択を同時に行うことができる手法として注目されている．しかし，一般には lasso は変数選択の一致性を有さないことが Meinshausen and Bühlmann (2006) および Zou (2006) により指摘されている．

そこで，Zou (2006) は変数選択の一致性を有するように lasso の形を改良した次の目的

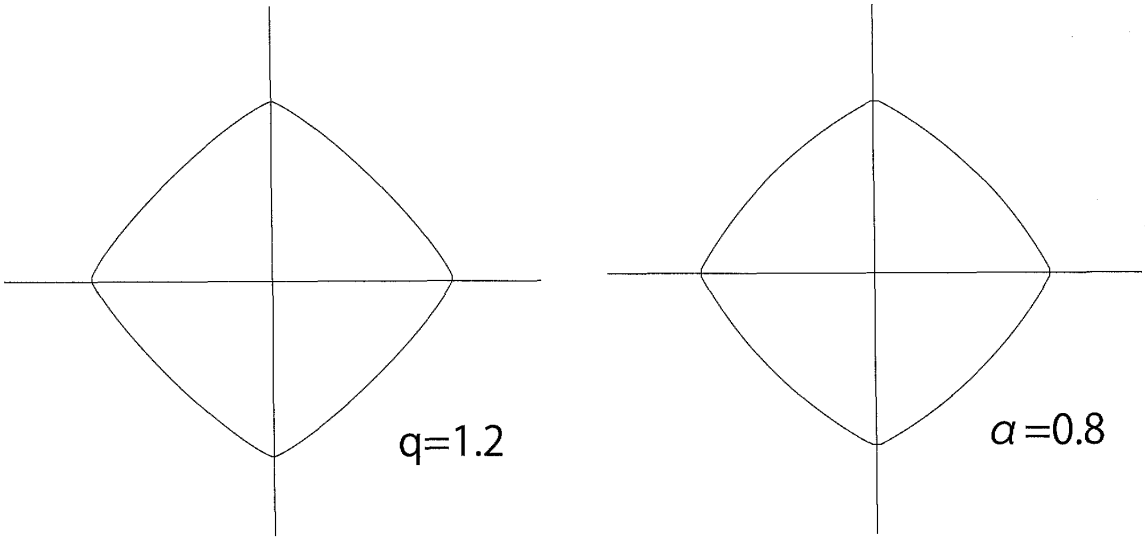

図 7　ブリッジ制約 (左図) および Elastic net 制約 (右図)

関数の最小化を考えた.

$$(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})+\gamma\sum_{j=1}^{p}\hat{w}_j|\beta_j|. \tag{5.14}$$

ここで, γ (>0) は正則化パラメータ, $\hat{w}_j=1/|\hat{\beta}_j|^\eta$ $(j=1,\ldots,p;\ \eta>0)$ は正則化項内で各係数パラメータに掛かる重みであり, $\hat{\beta}_j$ $(j=1,\ldots,p)$ は最小 2 乗推定量である. Zou (2006) はこの推定方法を Adaptive lasso と呼んだ. Adaptive lasso の推定アルゴリズムは, lasso の場合と同じものを用いることができ, Zou (2006) では計算量が比較的小さい LARS アルゴリズムを用いることを推奨している.

5.3.3 Fused lasso

説明変数に対するデータがその説明変数の番号に関して順序付けられている場合, この順序も考慮に入れた縮小推定法として, Tibshirani *et al.* (2005) は次の目的関数を最小化する Fused lasso と呼ばれる手法を提案した.

$$(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})+\gamma_1\sum_{j=1}^{p}|\beta_j|+\gamma_2\sum_{j=2}^{p}|\beta_j-\beta_{j-1}|. \tag{5.15}$$

ここで, γ_1 (>0), γ_2 (>0) はそれぞれ正則化パラメータである. Fused lasso は, 生命科学分野や画像処理など様々な分野で用いられる有用な手法として近年注目を集めている (例えば, Friedman *et al.* (2007), Tibshirani and Wang (2008) を参照). また, Fused lasso の推定方法についてはいくつかの有用なアルゴリズムが提唱されており, 詳しくは Tibshirani *et al.* (2005) および Friedman *et al.* (2007) を参照されたい.

5.3.4 Grouped lasso

いま, いくつかの説明変数があらかじめに設定されたグループに分かれている状況を想定する. 例えば, 同じ生物学的経路を有する遺伝子やカテゴリカルデータの水準を表すダ

ミー変数などを考える場合がこの状況にあてはまる．このような状況において，変数選択はそのグループ毎に行われるのが好ましい．Yuan and Lin (2006) により提案されたgrouped lassoは，このグループ毎の選択を可能とする推定方法である．

まず，p 個の説明変数が J 個のグループに分かれているとし，各グループ内の説明変数の個数を p_j とする．このとき，j 番目のグループに属する $n \times p_j$ 計画行列を X_j としたときに (5.2) 式の回帰モデルを次のように表す．

$$\boldsymbol{y} = \sum_{j=1}^{J} X_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}. \tag{5.16}$$

ここで，$\boldsymbol{\beta}_j$ を p_j 次元係数パラメータベクトルとする．このとき，grouped lassoは次の目的関数の最小化により定式化される (Yuan and Lin (2006))．

$$\left(\boldsymbol{y} - \sum_{j=1}^{J} X_j \boldsymbol{\beta}_j\right)^T \left(\boldsymbol{y} - \sum_{j=1}^{J} X_j \boldsymbol{\beta}_j\right) + \gamma \sum_{j=1}^{J} \sqrt{p_j} ||\boldsymbol{\beta}_j||_2. \tag{5.17}$$

ここで，$\gamma\ (>0)$ は正則化パラメータであり，$||\cdot||_2$ はユークリッドノルム ($||\boldsymbol{\beta}||_2 = (\boldsymbol{\beta}^T\boldsymbol{\beta})^{1/2}$) を表す．

罰則項の形に注目すると，この罰則項は各グループの係数の2乗和に対して L_1 制約を課していることがわかる．なぜならば，各グループの係数の2乗和を $\eta_j^2 = \boldsymbol{\beta}_j^T\boldsymbol{\beta}_j$ とおくと (5.17) 式の罰則項の部分は

$$\sum_{j=1}^{J} \sqrt{p_j} ||\boldsymbol{\beta}_j||_2 = \sum_{j=1}^{J} \sqrt{p_j} |\eta_j| \tag{5.18}$$

と変形できるからである．さらに，ベクトル $\boldsymbol{\beta}_j$ のすべての成分が0になるときに限ってそのベクトルのユークリッドノルムが0になるため，(5.17) 式の罰則項を用いることによりグループ毎の選択が可能になる．Grouped lassoは，通常のlassoを推定するときに用いられるアルゴリズムをそのまま適用することができ，特にYuan and Lin (2006) ではLARSアルゴリズムを用いている．Grouped lassoの理論的性質については，Bach (2008) およびNardi and Rinaldo (2008) の研究がある．また，非線形回帰モデルへの応用については，Lin and Zhang (2006) およびRavikumar *et al.* (2009) を参照されたい．

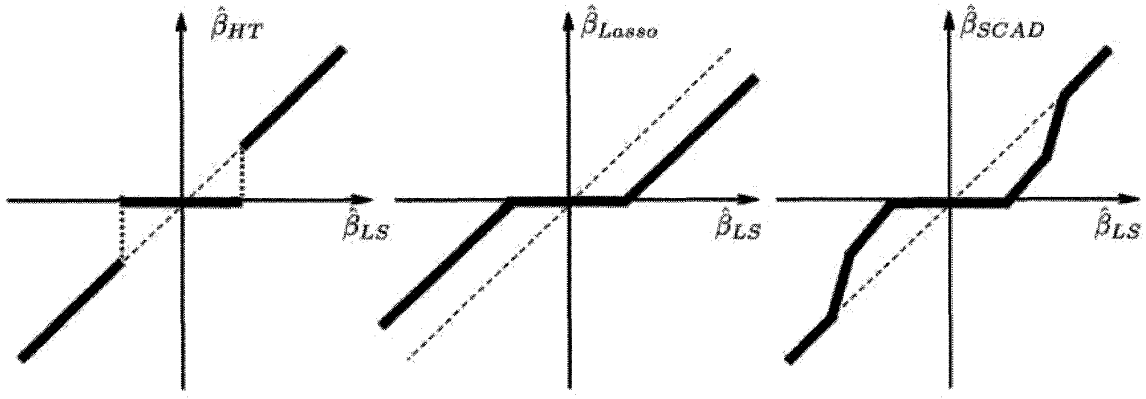

図 8 最小 2 乗推定量 (点線) と縮小推定量 (実線) との関係図. 左：Hard thresholding, 中：Lasso, 右：SCAD.

5.3.5 SCAD

Fan and Li (2001) は，連続性と不偏性を考慮に入れた次の式で与えられる SCAD (smoothly clipped absolute deviation) 制約を提案した.

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & (|\beta| \le \lambda), \\ -\dfrac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)} & (\lambda < |\beta| \le a\lambda), \\ \dfrac{(a+1)\lambda^2}{2} & (a\lambda < |\beta|). \end{cases}$$

ただし，$\lambda\,(>0)$ と $a\,(>2)$ は調整パラメータであり，推定の際には適切な値が必要となる. Fan and Li (2001) は，ベイズリスク最小化の観点から，$a = 3.7$ を用いた.

図 8 は，ガウスノイズと正規直交性を仮定した計画行列をもつ線形回帰モデルに対する 3 つの推定量の関係を示したものである. Hard thresholding (Donoho and Johnstone (1994)) では，0 と推定されない部分では最小 2 乗推定量と一致しているが，境界部分では不連続となっているため，不安定な推定となってしまう. 一方，lasso では連続的に推定量が変化するが，0 でない推定量の部分で制約が課され，バイアスをもつ. このような問題に対して，SCAD 制約には縮小レベルに対する連続性と不偏性が考慮されている. SCAD の理論的性質については，Fan and Li (2001)，Fan and Peng (2004)，Kim *et al.* (2008) 等の研究がある.

SCAD は，L_1 制約を含むため，推定量を解析的に求めるのが困難となる. そこで，ここでは，局所 2 次近似 (LQA; local quadratic approximation) の手法について述べる. まず，β の初期値 $\beta_0\ (\neq 0)$ を定める. $\beta \approx \beta_0$ のとき，$p_\lambda(|\beta|)$ の微分を次のように近似することができる.

$$[p_\lambda(|\beta|)]' = p'_\lambda(|\beta|)\mathrm{sign}(\beta) = \{p'_\lambda(|\beta|)/|\beta|\}\,\beta \approx \{p'_\lambda(|\beta_0|)/|\beta_0|\}\,\beta. \tag{5.19}$$

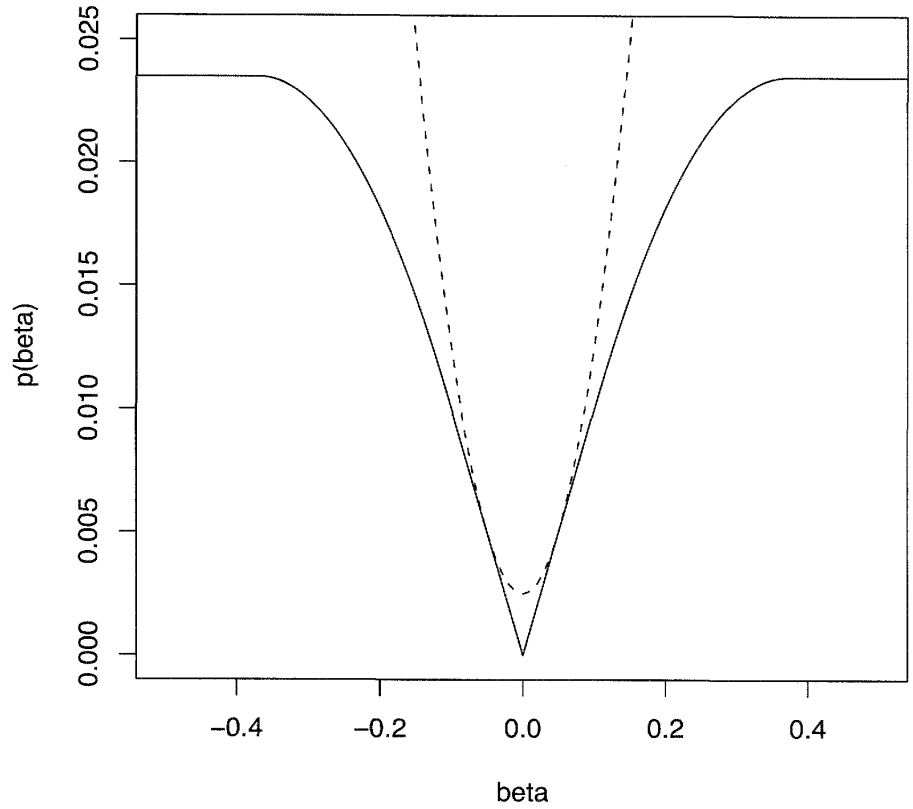

図 9 $\lambda = 0.1, \beta_0 = 0.15$ の時の SCAD ペナルティの 2 次近似の様子．横軸が β を表し，実線が $p_\lambda(|\beta|)$，点線がその 2 次近似を表す．

(5.19) 式の両辺を積分することにより，$p_\lambda(|\beta|)$ は次のように近似できる．

$$p_\lambda(|\beta|) \approx p_\lambda(|\beta_0|) + \frac{1}{2}\{p'_\lambda(|\beta_0|)/|\beta_0|\}(\beta^2 - \beta_0^2). \tag{5.20}$$

(5.19) 式，(5.20) 式の近似式は両辺に $\beta = \beta_0$ を代入すると完全に一致する．それゆえ，(5.20) 式の右辺は，関数 $p_\lambda(|\beta|)$ を点 β_0 で接する 2 次関数で近似している．図 9 は β_0 が 0.15 のときの 2 次近似を表す図である．β_0 が 0 でない限り，β が β_0 に近い時はうまく近似されている様子を表している．(5.20) 式の右辺は β に関する 2 次関数なので，微分することができ，たとえばニュートンラフソン法を組み合わせることにより推定アルゴリズムを構築できる．また，反復の途中で $|\beta|$ がある閾値 (たとえば 10^{-5}) より小さければ，$\beta = 0$ と推定する．

SCAD はセミパラメトリックモデルや比例ハザードモデルなどに対しても適用されている (例えば, Fan and Li (2002, 2004), Cai *et al.* (2005))．さらに, Wang *et al.* (2007) は grouped lasso の考えを SCAD に応用することによって grouped SCAD を提案し, varying-coefficient モデル (Hastie and Tibshirani (1993)) の推定を行った．また, Matsui and Konishi (2009) は grouped SCAD を関数回帰モデルの正則化項に応用している．

6. モデルの評価と選択

Lasso タイプの正則化項を課したモデルを推定する際，正則化パラメータ γ の値に依存して様々なモデルが構成される．したがって，モデリングの過程において，どのように γ の値を選択すればよいかということが本質となる．本節では，L_1 型正則化法によって推定されたモデルを評価するための様々な基準について述べる．L_1 正則化法は，パラメータベクトルの 2 次形式で表される正則化項を課した正則化法と異なり，パラメータに関して微分不可能となること，また，データ数より次元数が高いことからデータ数に関する漸近理論が適用できないことなどから，今後も研究の進展が待たれる分野である．

6.1 クロスバリデーション

観測データに基づいて推定されたモデルの良さを，将来得られるデータに対する予測精度によって測ることを考える．いま，lasso 型推定法によって推定された回帰式を $\hat{u}(\boldsymbol{x})$ とする．このとき，観測データとは独立に $\boldsymbol{x}_\alpha$ でランダムに得られたデータ z_α に対して（平均）予測 2 乗誤差

$$\mathrm{PSE} = \frac{1}{n}\sum_{\alpha=1}^{n} E\left[\{z_\alpha - \hat{u}(\boldsymbol{x}_\alpha)\}^2\right] \tag{6.1}$$

が最小となるモデルを最適なモデルと考える．期待値は将来のデータ z_α に関して取るものとする．この予測 2 乗誤差に含まれる将来のデータ z_α を現在のデータ y_α でおきかえた残差平方和

$$\mathrm{RSS} = \frac{1}{n}\sum_{\alpha=1}^{n}\{y_\alpha - \hat{u}(\boldsymbol{x}_\alpha)\}^2 \tag{6.2}$$

で予測 2 乗誤差を推定すると，たとえば線形回帰モデルでは全ての変数を用いて推定したモデルが最もよいモデルであるという結果を導く．

クロスバリデーション (CV: cross validation; Stone (1974)) は，モデルの推定に用いるデータとモデルの評価に用いるデータを分離して予測 2 乗誤差の推定を行う方法である．まず，データ $\{(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)\}$ の中から α 番目のデータを取り除いた $(n-1)$ 個のデータに基づいて推定された回帰式を $\hat{u}^{(-\alpha)}(\boldsymbol{x})$ とする．このようにして推定されたモデルを用いて，(6.1) 式の予測 2 乗誤差を

$$\mathrm{CV} = \frac{1}{n}\sum_{\alpha=1}^{n}\left\{y_\alpha - \hat{u}^{(-\alpha)}(\boldsymbol{x}_\alpha)\right\}^2 \tag{6.3}$$

で推定する．クロスバリデーションと AIC タイプの情報量規準の漸近的同等性が Stone (1977), Konishi and Kitagawa (2008) によって示されている．また, Fujikoshi *et al.* (2003), Yanagihara *et al.* (2006) は，漸近バイアスの補正項の精度を改善したクロスバリデーション推定量を与えた．

また，クロスバリデーションによる選択では選択されるモデルの不安定性が指摘されており，データ集合 $\{(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)\}$ を K 個に分割してクロスバリデーションを実行する K 分割交差検証法 (K-fold cross-validation) がよく用いられる．K としては 5 や 10 が主に用いられている (Hastie *et al.* (2009, 7.10 節))．

6.2 一般化自由度

Ye (1998) は，最尤法やリッジ推定法の枠を外した一般の推定量に対するモデルの複雑さを表す一般化自由度 GDF (generalized degrees of freedom) を定義した．いま，データ $\boldsymbol{y}$ は平均ベクトル $\boldsymbol{\mu}$，分散共分散行列 $\sigma^2 I_n$ の多変量正規分布 $N_n(\boldsymbol{\mu}, \sigma^2 I_n)$ に従うとする．Ye (1998) は，モデリングとはデータ $\boldsymbol{y}$ の推定値 $\hat{\boldsymbol{\mu}}$ を生成する写像 $\mathcal{M} : \boldsymbol{y} \to \hat{\boldsymbol{\mu}}$ と捉えて，モデリングに対する一般化自由度を次で定義した．

$$\mathrm{GDF} = \sum_{\alpha=1}^{n} \frac{\partial E_{\boldsymbol{Y}|\boldsymbol{\mu}}[\hat{\mu}_\alpha(\boldsymbol{Y})]}{\partial \mu_\alpha} = \sum_{\alpha=1}^{n} \frac{\mathrm{cov}(y_\alpha, \hat{\mu}_\alpha)}{\sigma^2}. \tag{6.4}$$

Efron (2004) は，一般化自由度が Mallows の C_p を拡張することによって導出されることを示した．一般化自由度は，任意のモデリング $\mathcal{M} : \boldsymbol{y} \to \hat{\boldsymbol{\mu}}$ に対して適用することができるため，lasso タイプの推定量のような正則化項が微分不可能な場合においても適用できるという特徴をもつ．また，一般化自由度は有効パラメータ数 (たとえば Hastie *et al.* (2009, 5.4.1 節)) を一般化していると考えられる．実際，$\hat{\boldsymbol{\mu}} = H(\gamma, m)\boldsymbol{y}$ と書けるとすると，

$$\sum_{\alpha=1}^{n} \frac{\mathrm{cov}(y_\alpha, \hat{\mu}_\alpha)}{\sigma^2} = \frac{\mathrm{tr}\{\mathrm{cov}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})\}}{\sigma^2} = \mathrm{tr}\{H(\gamma, m)\} \tag{6.5}$$

となり，有効パラメータ数と一致する．

(6.4) 式で与えられる自由度に含まれる $\mathrm{cov}(y_\alpha, \hat{\mu}_\alpha)$ は，積分計算が必要となるため，一般に解析的に導出することは困難である．そのため，パラメトリックブートストラップ法や SURE (Stein's unbiased risk estimate; Stein (1981)) による一般化自由度の推定が考えられる．

パラメトリックブートストラップ法によるアプローチ (Efron (2004)) は，lasso タイプの様々な推定量に対して適用できる極めて汎用性の高い手法である．まず，予測値ベクトル $\hat{\boldsymbol{\mu}}$ と，誤差分散の推定値 $\hat{\sigma}^2$（例えば，最も複雑なモデルの誤差分散の不偏推定量）を計算する．次に，ブートストラップ標本を $N_n(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2 I_n)$ から発生させ，そのブートストラップ標本 $\boldsymbol{y}^*$ から $\hat{\boldsymbol{\mu}}^*$ を推定する．このプロセスを B 回繰り返し，自由度を

$$\widehat{\mathrm{cov}}(y_\alpha, \hat{\mu}_\alpha) \approx \frac{1}{B-1} \sum_{b=1}^{B} \hat{\mu}_\alpha^{*b}(y_\alpha^{*b} - y_\alpha^{*\cdot}), \quad y_\alpha^{*\cdot} = \frac{1}{B} \sum_{b=1}^{B} y_\alpha^{*b} \tag{6.6}$$

と推定する．

ブートストラップ法によるアプローチは，極めて汎用性がある上，一般化自由度の推定量として安定しているという利点を有するものの，計算コストが大きいという難点をもつ．そこで，SURE に基づくアプローチが有用であると考えられる．まず，SURE により，$\mathrm{cov}(y_\alpha, \hat{\mu}_\alpha)$ が $\mathrm{cov}(y_\alpha, \hat{\mu}_\alpha) = E[\partial \hat{\mu}_\alpha / \partial y_\alpha]$ で与えられる．そこで，(6.4) 式の自由度に含まれる項 $\sum \mathrm{cov}(y_\alpha, \hat{\mu}_\alpha)$ を $\sum_{\alpha=1}^n \mathrm{cov}(y_\alpha, \hat{\mu}_\alpha) = \sum_{\alpha=1}^n \partial \hat{\mu}_\alpha / \partial y_\alpha$ で推定することができる．Efron *et al.* (2004) は，適当な条件のもとでは LARS 推定量に対する一般化自由度の不偏推定量が，LARS アルゴリズムで選ばれた変数の数と等しいことを示した．さらに，Zou *et al.* (2006) は，lasso 推定量に対する一般化自由度の不偏推定量が，係数の推定値が 0 でないパラメータの個数と等しいことを示した．これより，計算コストを増やさずに一般化自由度を推定することができる．

一般化自由度の回帰モデルへの適用については，Shen and Ye (2002), Shen *et al.* (2004), Zou and Li (2008) 等を参照されたい．

6.3 モデル評価基準 DIC, EIC

情報量規準 AIC(Akaike (1973, 1974)) は，真のモデルと想定したモデルの近さを捉える Kullback-Leibler 情報量 (Kullback and Leibler (1951)) に基づいて導出された極めて適用範囲の広い柔軟な評価基準であり，諸科学の様々な分野で応用されている．しかし，AIC は最尤法によって推定されたモデルを評価する基準であるため，正則化最尤法によって推定されたモデルの評価に直接適用するにはいくつかの問題点が生じる．そこで，Kullback-Leibler 情報量に基づく様々な AIC タイプの評価基準が提案された (例えば，Konishi and Kitagawa (1996, 2008))．本稿では，lasso タイプの推定法によって推定されたモデルを評価することができる基準として EIC (Ishiguro *et al.* (1997)) と DIC (Spiegelhalter *et al.* (2002)) を紹介する．

いま，目的変数と説明変数に関する n 組のデータ $\{(y_\alpha, \boldsymbol{x}_\alpha);\ \alpha = 1, \ldots, n\}$ が真の分布 $g(\boldsymbol{z}|X) = \prod_{\alpha=1}^n g(z_\alpha|\boldsymbol{x}_\alpha)$ から発生したとする．真の分布 $g(\boldsymbol{z}|X)$ を近似するため，パラメトリックモデル $f(\boldsymbol{z}|X;\boldsymbol{\theta}) = \prod_{\alpha=1}^n f(z_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta})$ を想定し，観測されたデータに基づいて構築した統計モデルを $f(\boldsymbol{z}|X;\hat{\boldsymbol{\theta}})$ とする．ただし，$\hat{\boldsymbol{\theta}}$ は n 組のデータ $\{(y_\alpha, \boldsymbol{x}_\alpha);\ \alpha = 1, \ldots, n\}$ から推定したパラメータの推定量とする．ここで，真のモデル $g(\boldsymbol{z}|X)$ と構築した統計モデル $f(\boldsymbol{z}|X;\hat{\boldsymbol{\theta}})$ の近さを Kullback-Leibler 情報量

$$I(g, f) = E_{G(\boldsymbol{z}|X)}\left[\log \frac{g(\boldsymbol{z}|X)}{f(\boldsymbol{z}|X;\hat{\boldsymbol{\theta}})}\right] \tag{6.7}$$

で測る．

Kullback-Leibler 情報量は，観測データ $\{(y_\alpha, \boldsymbol{x}_\alpha);\alpha = 1, \ldots, n\}$ とは独立に将来真の確率構造から得られるデータ $\{(z_\alpha, \boldsymbol{x}_\alpha);\alpha = 1, \ldots, n\}$ の従う分布 $g(\boldsymbol{z}|X)$ を，統計モデル

$f(\boldsymbol{z}|X;\hat{\boldsymbol{\theta}})$ で予測したときの平均的な良さを測っており，この値が小さいモデルを選択する．これは，平均対数尤度 $E_{G(\boldsymbol{z}|X)}[\log f(\boldsymbol{z}|X;\hat{\boldsymbol{\theta}})]$ を最大にするモデルを選択することと同等である．そこで，平均対数尤度を対数尤度 $\log f(\boldsymbol{y}|X;\hat{\boldsymbol{\theta}})$ で推定することを考える．しかしながら，対数尤度は平均対数尤度の推定量としてバイアスが生じ，さらにそのバイアスの大きさが想定するモデルによって異なる．そこで，そのバイアスを補正するため，対数尤度にバイアス補正項

$$b(G) = E_{G(\boldsymbol{y}|X)}[\log f(\boldsymbol{y}|X;\hat{\boldsymbol{\theta}}) - E_{G(\boldsymbol{z}|X)}[\log f(\boldsymbol{z}|X;\hat{\boldsymbol{\theta}})]] \tag{6.8}$$

を加えることにより，情報量規準は

$$-2\log f(\boldsymbol{y}|X;\hat{\boldsymbol{\theta}}) + 2b(G) \tag{6.9}$$

で定義される (小西・北川 (2004), Konishi and Kitagawa (2008)).

情報量規準を導出する際に重要となってくるのが，(6.8) 式のバイアス補正項をどのように推定するかということにある．例えば，バイアス補正項の推定量を解析的に求めるのではなくブートストラップ法 (Efron (1979)) を適用して数値的に求める方法が提案されている (Efron (1983, 1986), Ishiguro *et al.* (1997)). 特に，Ishiguro らはこれを EIC (extended information criterion) と呼んだ.

また，Spiegelhalter *et al.* (2002) は，(6.8) 式のバイアス補正項の第 2 項目の真の分布に関する期待値 $E_{G(\boldsymbol{z}|X)}$ をとった後に，さらに事後分布に関する期待値 $E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}$ をとった量が，次式で定義される有効パラメータ数

$$P_D = E_{\pi(\boldsymbol{\theta}|\boldsymbol{y})}[-2\log f(\boldsymbol{y}|\boldsymbol{\theta})] + 2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) \tag{6.10}$$

で近似できるとし，ベイズ型情報量規準 DIC を導出した．有効パラメータ数 P_D はバイアス補正項に含まれる事後分布に関する期待値を解析的に求めることが困難な場合でも，マルコフ連鎖モンテカルロ法を適用することによって数値的に求めることができるという利点を有する．そのため，医学統計，生物統計，経済学等極めて幅広い分野で応用されている (たとえば，Egger *et al.* (2002), Yu (2004)). Lasso タイプによって推定されたモデルの評価に DIC を適用した研究としては Tateishi *et al.* (2010) が挙げられる.

謝辞

査読者の方には貴重なご意見ご指摘をいただきました．ここに記して御礼申し上げます.

参 考 文 献

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd Inter. Symp. Information Theory* (eds. by B. N. Petrov and F. Csaki), Akademiai Kiado, Budapest, 267–281. (Reproduced

in *Breakthroughs in Statistics*, Vol. I, Foundations and Basic Theory (eds. S. Kotz and N. L. Johnson), Springer-Verlag, New York, (1992) 610–624.)

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Autom. Contr.*, **AC-19**, 716–723.

Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks, *J. Statist. Plann. Inference*, **138**, 3616–3633.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations (with discussion), *J. Am. Statist. Assoc.*, **96**, 939–967.

Antoniadis, A., Gijbels, I. and Nikolova, M. (2010). Penalized likelihood regression for generalized linear models with non-quadratic penalties, *Ann. Inst. Statist. Math.* (to appear).

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning, *J. Mach. Learn. Res.*, **9**, 1179–1225.

Bishop, C. M. (1991). Improving the generalization properties of radial basis function neural networks, *Neural Networks*, **3**, 579–588.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford, Oxford University Press.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *Ann. Statist.*, **24**, 2350–2383.

Buhmann, M. D. (2009). *Radial Basis Functions: Theory and Implementations*, Cambridge, Cambridge University Press.

Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Variable selection for multivariate failure time data, *Biometrika*, **92**, 303–316.

de Boor, C. (2001). *A Practical Guide to Splines*, Revised Edition. New York, Springer.

Denison, D., Mallick, B. and Smith, A. (1998). Automatic Bayesian curve fitting, *J. Roy. Statist. Soc. Ser. B*, **60**, 333–350.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, **81**, 425–455.

Donoho, D. L. and Johnston, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *J. Roy. Statist. Soc. Ser. B*, **90**, 1200–1224.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Ann. Statist.*, **7**, 1–26.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation, *J. Am. Statist. Assoc.*, **78**, 316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule?, *J. Am. Statist. Assoc.*, **81**, 461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation, *J. Am. Statist. Assoc.*, **99**, 619–642.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion), *Ann. Statist.*, **32**, 407–499.

Egger, M., May, M., Chêne, G., Phillips, A. N., Ledergerber, B., Dabis, F., Costagliola, D., Monforte, A. D., de Wolf, F., Reiss, P., Lundgren, J. D., Justice, A. C., Staszewski, S., Leport, C., Hogg, R. S., Sabin, C. A., Gill, M. J., Salzberger, B., Sterne, J. A. C. and the ART Cohort Collaboration (2002). Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies, *Lancet*, **360**, 119–129.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties (with discussion), *Statist. Sci.*, **11**, 89–121.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd ed., New York, Marcel Dekker.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*, London, Chapman & Hall.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Statist. Assoc.*, **96**, 1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's propotional hazards model and frailty model, *Ann. Statist.*, **30**, 74–99.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *J. Am. Statist. Assoc.*, **99**, 710–723.

Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters, *Ann. Statist.*, **32**, 928–961.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109–148.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion), *Ann. Statist.*, **19**, 1–141.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization, *Ann. Appl. Statist.*, **1**, 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2009). Regularization paths for generalized linear models via coordinate descent, *Technical Report*, Stanford University.

Fu, W. (1998). Penalized regression: the bridge versus the lasso, *J. Comput. Graph. Statist.*, **7**, 397–416.

Fujii, T. and Konishi, S. (2006). Nonlinear regression modeling via regularized wavelets and smoothing parameter selection, *J. Multivariate Anal.*, **97**, 2023–2033.

Fujikoshi, Y., Noguchi, T., Ohtaki, M. and Yanagihara, H. (2003). Corrected versions of cross-validation criteria for selecting multivariate regression and growth curve models, *Ann. Inst. Statist. Math.*, **55**, 537–553.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, London, Chapman & Hall.

Gu, C. (2002). *Smoothing Spline ANOVA Models*, New York, Springer.

Hall, P. and Patil, P. (1996). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear wavelet methods, *J. Roy. Statist. Soc. Ser. B*, **58**, 361–377.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, London, Chapman & Hall.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models, *J. Roy. Statist. Soc. Ser. B*, **55**, 757–796.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed., New York, Springer.

日野幹雄 (1977). 「スペクトル解析」朝倉書店.

Horel, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

市田浩三, 吉本富士市 (1979). 「スプライン関数とその応用」教育出版.

井元清哉, 小西貞則 (1999). 「情報量規準に基づく B-スプライン非線形回帰モデルの推定」『応用統計学』**27**, 137–150.

Imoto, S. and Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria, *Ann. Inst. Statist. Math.*, **55**, 671–687.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC, *Ann. Inst. Statist. Math.*, **49**, 411–434.

Kawano, S. and Konishi, S. (2007). Nonlinear regression modeling via regularized Gaussian basis functions, *Bull. Inform. Cybern.*, **39**, 83–96.

Kim, Y., Choi, H. and Oh, H. (2008). Smoothly clipped absolute deviation on high dimensions, *J. Am. Statist. Assoc.*, **103**, 1665–1673.

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks, *Biometrika*, **91**, 27–43.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection, *Biometrika*, **83**, 875–890.

小西貞則, 北川源四郎 (2004). 「情報量規準」朝倉書店.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*, New York, Springer.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statist.*, **22**, 79–86.

Lang, S. and Brezger, A. (2004). Bayesian P-splines, *J. Comput. Graph. Statist.*, **13**, 183–212.

Leitenstorfer, F. and Tutz, G. (2007). Knot selection by boosting techniques, *Comput. Statist. Data Anal.*, **51**, 4605–4621.

Lin, Y. and Zhang, H. (2006). Component selection and smoothing in smoothing spline analysis of variance models, *Ann. Statist.*, **34**, 2272–2297.

Loader, C. R. (1999). *Local Regression and Likelihood*, New York, Springer.

Marx, B. and Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood, *Comput. Statist. Data Anal.*, **28**, 193–209.

Matsui, H. and Konishi, S. (2009). Variable selection for functional regression model via the L_1 regularization, MI Preprint Series, Kyushu University, MI2009–3.

McCaffrey, D. F., Ellner, S., Gallant, A. R. and Nychka, D. W. (1992). Estimating the lyapunov exponent of a chaotic system with nonparametric regression, *J. Am. Statist. Assoc.*, **87**, 682–695.

Meinshausen, N. and Bühlmann, P. (2006). High-dimenional graphs and variable selection with the lasso, *Ann. Statist.*, **34**, 1436–1462.

Miyata, S. and Shen, X. (2005). Free-knot splines and adaptive knot selection, *J. Japan Statist. Soc.*, **35**, 303–324.

Moody, J. and Darken, C. J. (1989). Fast learning in networks of locally-tuned proccessing units, *Neural Comput.*, **1**, 281–294.

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models, *Electr. J. Statist.*, **2**, 605–633.

Osborne, M. R., Presnell, B. and Turlach, B. A. (1998). Knot selection for regression splines via the lasso, *Comput. Sci. Statist.*, **30**, 44–49.

Park, M. Y. and Hastie, T. (2007). L_1 regularization path algorithm for generalized linear models, *J. Roy. Statist. Soc. Ser. B*, **69**, 659–677.

Park, T. and Cassella, G. (2008). The Bayesian lasso, *J. Am. Statist. Assoc.*, **103**, 681–686.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models, *J. Roy. Statist. Soc. Ser. B*, **71**, 1009–1030.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge, Cambridge University Press.

Rosset, S., Zhu, J. and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier, *J. Mach. Learn. Res.*, **5**, 941–973.

坂本亘，井筒理人，白旗慎吾 (2008).「罰則付きスプラインによる非線形回帰構造の推測」『計算機統計学』**21**, 55–94.

佐藤義治 (1996).「統計モデルとしてのニューラルネットワーク」『統計数理』**44**, 85–98.

Schumaker, L. L. (1993). *Spline Functions: Basic Theory*, Melbourne, Florida, Krieger.

Shen, X. and Ye, J. (2002). Adaptive model selection, *J. Am. Statist. Assoc.*, **97**, 210–221.

Shen, X., Huang, H.-C. and Ye, J. (2004). Adaptive model selection and assessment for exponential family models, *Technometrics*, **46**, 306–317.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, New York, Springer.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression, *Statist. Comput.*, **14**, 199–222.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *J. Roy. Statist. Soc. Ser. B*, **64**, 583–639.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution, *Ann. Statist.*, **9**, 1135–1151.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. B*, **36**, 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. Roy. Statist. Soc. B*, **36**, 44–47.

Tateishi, S., Matsui, H. and Konishi, S. (2010). Nonlinear regression modeling via the lasso-type regularization, *J. Statist. Plann. Inference*, **140**, 1125–1134.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso, *Biostatistics*, **9**, 18–29.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso, *J. Roy. Statist. Soc. Ser. B*, **67**, 91–108.

Wager, C., Vaida, F. and Kauermann, G. (2007). Model selection for P-spline smoothing using Akaike information criteria, *Austral. N. Z. J. Stat.*, **49**, 173–190.

Wahba, G. (1983). Bayesian "Confidence Intervals" for the cross-validated smoothing spline, *J. Roy. Statist. Soc. Ser. B*, **45**, 133–150.

Wahba, G. (1990). *Spline Models for Observational Data*, Philadelphia, SIAM.

Wand, M. P. (1999). On the optimal amount of smoothing in penalised spline regression, *Biometrika*, **86**, 936–940.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, London, Chapman & Hall.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data, *Bioinformatics*, **23**, 1486–1494.

Weisberg, S. (1980). *Applied Linear Regression*, New York, Wiley.

Wood, S. N. (2003). Thin plate regression splines, *J. Roy. Statist. Soc. Ser. B*, **65**, 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models, *J. Am. Statist. Assoc.*, **99**, 673–686.

Yanagihara, H., Tonda, T. and Matsumoto, C. (2006). Bias correction of cross-validation criterion based on Kullback-Leibler information under a general condition, *J. Multivariate Anal.*, **97**, 1965–1975.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection, *J. Am. Statist. Assoc.*, **93**, 120–131.

Yu, J. (2004). On leverage in a stochastic volatility model, *J. Econ.*, **127**, 165–178.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *J. Roy. Statist. Soc. Ser. B*, **68**, 49–67.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator, *J. Roy. Statist. Soc. Ser. B*, **69**, 143–161.

Zou, H. (2006). The adaptive lasso and its oracle properties, *J. Am. Statist. Assoc.*, **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *J. Roy. Statist. Soc. Ser. B*, **67**, 301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models, *Ann. Statist.*, **36**, 1509–1533.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters, *Ann. Statist.*, **37**, 1733–1751.

Zou, H., Hastie, T. and Tibshirani, R. (2006). On the "degrees of freedom" of the lasso, *Ann. Statist.*, **35**, 2173–2192.