IWANAMI
DATA
SCIENCE Vol. 5
岩波データサイエンス

特集「スパースモデリングと多変量データ解析」
池田思朗・伊庭幸人・麻生英樹……4

〈**基礎編**〉

モデル選択超速習……AIC からスパースまで
伊庭幸人……6

スパース性を用いた推定　**池田思朗**……19

スパースモデリングを体験してみる
岩波データサイエンス刊行委員会……39

〈**展開編**〉

依存関係にスパース性を入れる
……グラフィカル lasso の話　**井手 剛**……48

画像処理とスパース　**本谷秀堅**……64

時間遷移のスパース性
……マーケットシェアの遷移を捉えるスパースグラフモデリング
日野英逸……80

〈**応用編**〉

行列データの分解　**麻生英樹**……91

〈コラム〉行列のトレースノルム
麻生英樹……104

行列分解をリコメンデーションに活かす
……Rによる実践例　**尾崎 隆**……106

［**話題**］

ことばの見分け方……テキスト言語判定
中谷秀洋……118

医学研究におけるメタアナリシス
……科学的根拠に基づく医療におけるエビデンス統合の方法
野間久史……130

［**その他・小説・漫画**］

計算機で作る面白いナンプレ⑤　**とん**……143

掌編小説《海に溺れて》❺深い谷を渡る
円城 塔……46

掌編漫画 ①偶然の気配　**panpanya**……116

〈**次巻予告**〉……63

表紙挿画＝蛯名優子
デザイン＝佐藤篤司

特集 スパースモデリングと多変量データ解析

10 年前，海外の統計学，情報理論，信号処理といった理論分野の学会の発表は，スパースモデリングで埋め尽くされていた。中心となるのは絶対値の和で定義された L_1（エルワン）ノルムによる正則化である。圧縮センシングの立役者の一人である Candès は「L_2 ノルムが 20 世紀のノルムだとするならば L_1 ノルムは 21 世紀のノルムだ」といっていた。この物言いはいささか大げさだが，L_1 ノルム正則化を用いることでそれまで諦めていた情報処理が可能となったことは事実である。

理論的枠組みが確立すると，他の分野での問題に応用されるようになる。スパースモデリングも，バイオインフォマティックス，工学をはじめ，ビッグデータを含めたあらゆるデータ解析において，その方法が取り込まれてきた。今となってはスパース性が使われていない分野を探すほうが大変なくらいだ。スパースモデリングは，基礎的な方法としてすっかり定着しているといえる。

日本国内の状況は少し異なる。1980 年代後半に石川眞澄の先駆的な研究はあったが，流行を作るまでには至らず，2010 年代に入ってスパースモデリングが注目されている。現在，科研費プロジェクト「スパースモデリングの深化と高次元データ駆動科学の創成」（新学術領域研究，平 25-29，代表：岡田真人）が行われており，その中では自然科学分野を中心に様々な応用が展開されている。国内のデータ処理の分野では今，まさに普及している時期といえる。

本特集では，こうしたスパースモデリングを概観する。大きな広がりを持つ方法論であるため，すべてを細かく説明することはできないが，基礎的な枠組みを示し，何ができて，どうやればいいのかを解説する。また，応用，実践についてもいくつかの記事を示した。多くの読者の興味の中心であるビジネスへの応用もすでに始まっているが，実際の普及はまだまだこれからだろう。本書がそうした応用を目指すデータサイエンティストのお役に立てれば幸いである。

以下，もう少し細かく内容を見てみよう。スパースモデリングというと数理的な面を強調した解説をよく見るが，ここでは，できるだけ実践的な面を強調するように心がけた。また，関連するテーマとして，リコメンデーションなどビジネスへの応用を意識した行列分解手法の解説も加えた。『岩波データサイエンス』の 1 巻のベイズモデリングや 3 巻の因果推論は，それぞれ多変量データに対する異なる視点を与えているが，本特集もまたひとつの切り口となるとよいと思う。

最初の 3 編が「基礎編」である。最初の伊庭の解説は「スパース以前」のモデル選択をめぐる諸問題を短くまとめている。スパースモデリングの位置付けをするという意図であるが，好みでスキップしても差し支えない。2 つ目の池田の解説が本特集の要となるもので，スパースモデリングの代表格である lasso の入門から，実際に使う上での注意，バリエーション，数値計算手法まで限られたページ数でわかりやすく説明している。3 つ目の解説では，R での実行例を示した。

次の 3 編が「展開編」に相当する。井手の解説で登場するガウス型のグラフィカルモデル（GGM）は 3 巻の小原・土谷の連載記事でもその数理的な面が論じられている。3 巻の特集での「因果推論」では「因果の向き」が重要だったが，GGM では矢印の向きは無視して「直接の影響」と「間接的なものを含んだ影響」の違いを考える（3 巻のコラム「グラフ表現超速習」を参照）。続く本谷の解説では「辞書」の概念を用いた画像処理においてスパースの考え方がどのように使われるかが示されている。最後の日野の解説はマルコフ連鎖モデルの遷移行列の推定にスパースモデリングを応用する最近の研究の紹介であるが，ビジネスデータへの応用例としても興味深い内容である。

残りの 2 編は「行列分解」とそのリコメンデーションへの応用に関するものである。麻生の解説は，特異値分解や主成分分析などの古典的話題からはじめて，スパースモデリングや非負値の制約付きの分解までをコンパクトにまとめており，初めての方はもちろん，すでに学ばれた方の知識の整理にもお勧めできる。尾崎の解説は対応する「実践編」となっており，R 言語のパッケージを利用して，リコメンデーションへの応用例が示されている。

（特集担当　池田思朗・伊庭幸人・麻生英樹）

モデル選択超速習
AICからスパースまで

伊庭幸人（統計数理研究所）

この巻の特集では「スパースモデリング」を扱うが，ここでは，そこに行くまでの話を前座としてまとめておく。はじめからメインディッシュに行きたい方は，本特集の池田の解説から読み始めても，ほぼ問題なく読めるはずである。

池田の解説では，回帰分析の具体的な例から話がはじまるので，「最尤法やベイズには馴染みがないが，重回帰分析は使ったことがある」という読者にはむしろわかりやすいかもしれない。そのあとでこの解説を読めば，従来からの流れの中にスパースモデリングを位置付けることができるだろう。

問題の整理

大雑把にいうと「**スパースモデリング**」とは，与えられたデータに応じて，統計モデルの必要な部分を自動的に抽出する技術である。重回帰分析の場合でいえば，多数の説明変数のリストから少数の必要なものを取り出す，ということになる。これに対して，昔からある言葉として，「**モデル選択**」や「**変数選択**」がある。「変数選択」はモデルに取り入れる変数を選ぶ場合に，「モデル選択」はより一般に，たとえば雑音の分布形の選択なども含めて使われる。

これらの用語の使い分けにはいろいろ流儀がありそうだが，この解説の中では「複数のモデルを当てはめた後で，その中から選ぶ」方法を変数選択／モデル選択，後述のL_1正則化(lasso)のように「罰則項を加えて推定することで，パラメータ推定と同時に未知パラメータの数が削減される」方法をスパースモデリングと呼び分けることにしよう。

このあたりの問題を考えるには，次の 3 つのレベルを区別するとよい。

1. なぜ変数が少ないモデルが欲しいのか(根本問題)
2. 具体的にどういう数式を最適にするのか(数理的表現)
3. 最適なモデルをどうやって探すか(アルゴリズム)

スパースモデリングで重要な進展があったのは，このうち 3. の部分である。ただし，その中には「3. がうまくいくために，2. の表式をどう設定したらよいか」という部分も含まれる。この解説では，その前提として，1, 2, 3 を含めた問題の全体像を眺めてみたい。

モデルの当てはまりと予測性能は違う

まず，1. の「根本問題」からはじめよう。「なぜ変数が少ないモデルが欲しいのか」という問いの答はひとつではないが，現代のデータサイエンスに大きな影響を与えたのは「予測」という視点である[1]。統計的機械学習でいう「汎化」もほぼ同じ意味になる。

ここでは簡単な例として，多項式回帰の次数選択を考える。まず，次の式で模擬データ $\{(x_i, y_i)\}$, $i=1,...,N$, を生成しよう。

$$y_i = \beta_2 x_i^2 + \beta_1 x_i + \eta_i \tag{1}$$

ここで，η_i は正規分布 $N(0, \sigma^2)$ に従う i ごとに独立な雑音であり，説明変数 $\{x_i\}$ の値は適当に選ぶとする(あとの数値例では $[0, 1]$ の区間を等分割した)。そして，次の 2 つの統計モデルによる当てはめを考える。

$$y = \beta_2 x^2 + \beta_1 x + \eta \tag{2}$$

$$y = \beta_1 x + \eta \tag{3}$$

η は正規分布に従う観測ごとに独立な雑音とする。式(2)は模擬データを作ったのと同じ 2 次式のモデル，(3)は 2 次の項をはぶいた直線当てはめのモデルである。ここでは，定数項(切片)はゼロであることがわかっているとした。

最初に，与えた模擬データ(訓練データ)そのものに対する当てはまりを見よう。モデル(2)については，β_1, β_2 の最小 2 乗推定値を $\hat{\beta}_1, \hat{\beta}_2$ とすると，訓練データとの差をあらわす 2 乗和(残差平方和)は

$$\sum_i (y_i - \hat{\beta}_2 x_i^2 - \hat{\beta}_1 x_i)^2$$

スパース性を用いた推定

池田思朗（統計数理研究所）

スパース性を利用した推定方法の代表格である lasso が提案されたのは 1996 年[8]。それから 20 年，機械学習がメジャーになり，いくつもの方法が生まれている中で，スパース性を使った方法の浸透，定着ぶりは群を抜いている。lasso の方法自体が複数の分野で用いられて結果を出しているのはもちろんのこと，「スパース性」を用いた方法はカーネル法や深層学習の中でもあちこちで顔を出す。もはや「スパース性」の流行は一過性ではない。

この流行は応用分野はもちろん，理論分野にも及んでいる。統計学や機械学習の分野以外でも，情報理論，凸最適化理論といった複数の分野に面白い問題を提供し，そうした分野の成果によって基盤が強化されてきた。ここでは，lasso を中心にスパース性に基づく方法を説明していこうと思う。

1 lasso

背景

まず lasso（Least Absolute Shrinkage and Selection Operator）について説明する。基本は次の線形回帰モデルである。

$$y_i = \beta_1 x_i^{(1)} + \cdots + \beta_p x_i^{(p)} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

y_i は従属変数，$x_i^{(1)}, \cdots, x_i^{(p)}$ は説明変数，推定したいパラメータは p 個の係数 β_j である。i はサンプルのインデックスで N 個のサンプルがあるとする。以下では y と $x^{(j)}$ の平均は 0 に調整してあり，切片は $\beta_0 = 0$ として省く。

ノイズ ϵ が正規分布にしたがうとき，最尤推定によって $\boldsymbol{\beta}$ を求めるには以下の

問題を解けばよい。

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_i^{(j)} \right)^2 \tag{2}$$

統計学では，$N \gg p$ とし，推定値 $\widehat{\beta}_j$ が $N \to \infty$ のとき漸近的にどう振舞うかを議論する。これに対し，近年，頻繁に直面するのが $N < p$ のケースだ。不思議に感じるかもしれないが，サンプルの共変量を数多く調べればこうした状況は容易に生じる。例えば，多くの患者の血液を調べて疾患とそれにかかわる遺伝子の関係を推定したいとする。血液のサンプルをとる人数 N はそれほど多くできないが，技術が進み，血液から調べられる遺伝子の種類 p は非常に大きい。サンプルを詳細に調べると，その分だけ p が大きくなる。こうした場合，統計学ではどう扱えばいいのだろう。漸近論による議論は使えないし，そもそもパラメータの推定ができるかさえ怪しい。

もし，N に比べて p が大きくても β_j のほとんどが 0 だとわかっているならば，これは大きなヒントとなる。遺伝子の例であれば，特定の疾患にかかわる遺伝子は少なく，多くの β_j は 0 であるという仮定にあたる。これは妥当なものに思える。あとは信じるに足る何らかの方法で，β_j の多くが 0 となるような推定ができればよい。lasso はまさにこの目的のために提案された方法である。$N < p$ であってもスパースな(0 の多い)パラメータベクトルを推定する。さらに都合のいいことに，lasso の方法は縮小推定の方法であり，モデル選択の方法としても，ベイズ統計の MAP 推定としても妥当な方法となっている。統計学が困っていた問題に対して，ちょうどよい強力な道具が現れたのだ。

定式化

まず，Tibshirani が提案した lasso の問題を示す。

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_i^{(j)} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{3}$$

(2)式に対し，条件 $\sum_j |\beta_j| \leq t$ が加わっている。この条件は $\boldsymbol{\beta}$ の凸領域を与え，最適化の目的関数は $\boldsymbol{\beta}$ の凸関数である。図 1 にあるように，凸領域には「角(かど)」があり，最適化したい関数は滑らかな等高線をもつ。楕円の中心が凸領域の外にある場合，等高線が角や辺に接した点が最適値を与える。図は N も p も 2 次元

行列分解をリコメンデーションに活かす
R による実践例

尾崎隆（データサイエンティスト）

本書で取り上げられている，行列分解を初めとするさまざまなスパース化手法は，一般には「次元削減」ないし「特徴量削減」として捉えられることが多いようである。しかしながら，これを「ある少数の軸に絞って集約化する」と捉えることによって，「共通項を見出す」という新たな活用法につながる。この考え方に基づき，近年では多くの web サービスで行列分解がリコメンデーション（推薦）機能を実現するために用いられるようになっている。

このセクションでは，いくつかの行列分解手法を題材として，実際にどのようにしてリコメンデーション機能として適用されているかを R コードによる実践を通じて見ていく。なお，実際のリコメンデーション機能はむしろ Python や Java, C++ などの本番運用システム実装に適したプログラミング言語によって実装されることが多いが[1]，ここでは実践例の簡便さを重視して R と{PMA}, {NMF}パッケージを用いている。

1　リコメンデーションとは

一般に「リコメンデーション」と言った場合に，読者の多くが想像するものとして「EC サイトでのお薦め商品表示」や「各種ポータルサイトでのお薦めコンテンツ表示」といったものが挙げられよう。これは言い換えると「ユーザーが次に興味を持つであろう情報を与える」という行為であり，それを過去のユーザーの行動データから推定する試みこそがリコメンデーションであるとも言える[1]。

リコメンデーションに際してよく用いられる手法として，協調フィルタリング

(collaborative filtering)がある[1]。これは，購買を行った個々のユーザーごとにそれぞれのアイテムを購入したか否か(もしくは個々のアイテムごとにそれぞれのユーザーが購入したか否か)から成るベクトルを作り，その上でベクトル同士の類似度を算出し，この類似度に基づいてアイテムごとの評価値を算出し「あなたと同じ商品を買った人はこんな商品も買っている」「この商品を買った人はこんな商品も買っている」という推薦を行う，というものである。

ところが，一般に購買データはスパースなものになりがちとされる。理由はシンプルで，そもそもショッピングの際，一度に多くの種類の物を購入する人は少ない上に，ユーザー同士で全く同じ商品の組み合わせを購入することは極めて稀なためである。例えば，揃ってスポーツに関心の高いユーザー 2 人がいても，往々にして購入するものは{テニス雑誌，ラバーチューブ，テーピング},{サッカー雑誌，消炎スプレー}というように，商品単位で言えば重ならないリスト同士になってしまったりする。そのため，購買データに対してそのまま協調フィルタリングを適用しようとすると，そもそも「ユーザー同士で互いに購入したことのない商品だらけ」ということになってしまい，ほとんど何も推薦できないということになりやすいことが知られている。

この問題を避けるため，次元削減を行うことで「互いに購入したことのない商品であっても同じ軸でまとめられるものは一まとめにして扱」い，その集約された「軸」に基づいて改めてリコメンデーションを行うという考え方がある。そこで用いられるのが，本書で取り上げられている行列分解系の手法である。

2　行列分解とリコメンデーション

素の購買データセットは「ユーザー同士で互いに購入したことのない商品」だらけになりがちな，いわば「スパース」なデータであるとも言える。これをより少数の「軸」に集約するためには，次元削減を効率良く行う必要がある。上記のスポーツに関心の高いユーザーの例で言えば，{スポーツ雑誌，スポーツ傷害向けグッズ}というように集約できる可能性があり，これを(可能であれば)特に事前知識なく数理的に行えることが求められる。そこで登場するのが行列分解である。

ここでは本特集の別の解説で取り上げられている手法を 3 点だけ，概要を取り上げるに留める。1 つめは特異値分解(Singular Value Decomposition: SVD)で