Fitting and Testing High-Order-Interaction Models
（高次交互作用モデルの同定と検定）

竹内一郎 （名古屋工業大学/理化学研究所/物質材料研究機構）

1 はじめに

さまざまな科学技術分野において研究対象に関する網羅的な高次元データが計測できるようになった．例えば，生命科学分野では網羅的遺伝情報を計測できるようになり，数千〜数万個の遺伝要因に基づく生命現象のモデリングが行われている．高次元データ分析では，膨大な特徴から有用なものを取捨選択することが不可欠である．機械学習や統計科学分野では，高次元線形モデルとその特徴選択に関する研究が盛んに行われている．一方，現象を説明するのに複数の特徴の組み合わせを考えることが必要な場合がある．例えば，生命科学分野では複数の遺伝的変異の組み合わせが薬剤効果に関連する場合があることが知られている．本研究では高次元線形モデルを拡張した高次交互作用モデルを考察する．

まず，問題の定式化を行う．モデル対象に関する d 次元の特徴ベクトル $\boldsymbol{x} = [x_1, \ldots, x_d]^\top \in \{0,1\}^d$ が与えられているとし，以下のような高次交互作用モデルを考える．

$$f(\boldsymbol{x}) = \sum_{j_1 \in [d]} \alpha_{j_1} x_{j_1} + \sum_{\substack{(j_1,j_2) \in [d]\times[d] \\ j_1 \neq j_2}} \alpha_{j_1,j_2} x_{j_1} x_{j_2} + \ldots + \sum_{\substack{(j_1,\ldots,j_r) \in [d]^r \\ j_1 \neq \ldots \neq j_r}} \alpha_{j_1,\cdots,j_r} x_{j_1} \cdots x_{j_r} \qquad (1)$$

ここで，$r \leq d$ は考慮する高次交互作用項の最大の次数を表す．本研究では，(1) 式の高次交互作用モデルにおいて，大部分の係数 α が零となるようなスパースモデリングを考える．遺伝的変異から薬剤効果を予測する問題を再び考えると，例えば，3 次の交互作用項の係数パラメータが $a_{2,4,7} \neq 0$ であれば，2 番目，4 番目，7 番目の遺伝的変異が同時に起こっている場合に薬剤効果が変わることを示唆している．スパース高次交互作用モデルは特徴 $\boldsymbol{x}$ に関して非線形であるので線形モデルよりも複雑な関係を記述できる．また，上述の遺伝的変異の例のように，複数特徴の組み合わせ要因を発見するためにも有用である．

しかしながら，(1) 式の高次交互作用モデルは $D = \sum_{\kappa \in [r]} \binom{d}{\kappa}$ 個のパラメータを持っており，d や r が小さい場合を除き推定すべきパラメータ数が膨大なものとなってしまう．例えば，$d = 10,000$，$r = 5$ のとき，$D > 10^{17}$ となり既存のスパースモデリングの方法を利用することができない．本講演では，スパース高次交互作用モデルの同定と検定に関する筆者らの最近の研究 [1, 2] を紹介する．我々の方法を用いると，D 個のパラメータを陽に扱うことなく，スパース高次交互作用モデルの同定や検定を行うことができる．文献 [1] において提案したスパース高次交互作用モデルの同定アルゴリズムでは，セーフスクリーニングと呼ばれる凸最適化のアルゴリズム [3, 4, 5] を利用する．また，文献 [2] において提案したスパース高次交互作用モデルの検定では，近年注目を集めている selective inference と呼ばれる枠組 [6] を利用する．

モデル作成用の訓練データを $\{(\boldsymbol{x}^i, y^i) \in \{0,1\}^d \times \mathbb{R}\}_{i \in [n]}$ と表記する．表 1 にデータ例を示す．表記を簡潔にするため，d 次元ベクトル $\boldsymbol{x} \in \{0,1\}^d$ を以下のように D 次元ベクトル $\boldsymbol{z} \in \{0,1\}^D$ に拡張し，訓練データを $\{(\boldsymbol{z}^i, y^i) \in \{0,1\}^D \times \mathbb{R}\}_{i \in [n]}$ と表す．

$$\boldsymbol{z} = [\underbrace{x_1, \ldots, x_d}_{\binom{d}{1}}, \underbrace{x_1 x_2, \ldots, x_{d-1} x_d}_{\binom{d}{2}}, \ldots, \underbrace{x_1 \cdots x_r, \ldots, x_{d-r} \cdots x_d}_{\binom{d}{r}}]^\top \in \mathbb{R}^D,$$

表 1: 本稿で対象とするデータのイメージ（二値表現された d 個の遺伝的変異の有無に基づいて薬剤効果を予測する問題の例）

	x_1 (gene 1)	x_2 (gene 2)	$\cdots$	x_d (gene d)	y (drug effect)
patient 1	0	1	...	0	16.5
patient 2	1	0	...	1	22.0
patient 3	1	1	...	0	18.0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	
patient n	0	1	...	1	24.5

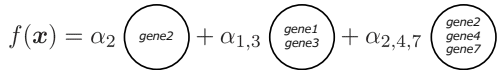

図 1: 高次交互作用モデルのスパース推定結果のイメージ : 表 1 のデータに高次交互作用モデルのスパース推定を適用した結果のイメージ図で，遺伝子 2 の変異，遺伝子 1, 3 の変異の組み合わせ，遺伝子 2, 4, 7 の変異の組み合わせが薬剤効果に影響を与えることを示唆している．

パラメータ α も同様に並べた D 次元係数ベクトル $\boldsymbol{\beta} \in \mathbb{R}^D$ を用いると，式 (1) の高次交互作用モデルは拡張データ z に関する D 次元線形モデル $f(\boldsymbol{z}) = \boldsymbol{\beta}^\top \boldsymbol{z} = \sum_{j\in[D]} \beta_j z_j$ と解釈できる．図 1 に高次交互作用モデルのスパース推定結果のイメージを例示する．

2　セーフスクリーニングに基づくスパース高次交互作用モデルの同定

以下では，文献 [1] で筆者らが提案したスパース高次交互作用モデルの同定アルゴリズムを簡単に紹介する．スパース推定のため，以下のような L_1 正則化を考える．

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^D} \mathcal{P}_\lambda(\boldsymbol{\beta}) = \sum_{i\in[n]} (y^i - \boldsymbol{\beta}^\top \boldsymbol{z}^i) + \lambda \|\boldsymbol{\beta}\|_1. \tag{2}$$

ここで，$\mathcal{P}_\lambda(\cdot)$ は正則化パラメータ λ における最適化問題の目的関数である．式 (2) は，D 次元の係数ベクトル $\boldsymbol{\beta}$ に関する最適化問題であるため，既存の凸最適化アルゴリズムを用いて解くのは困難である．

スパース推定においては係数が非零となる特徴をアクティブな特徴と呼ぶ．アクティブな特徴の集合を $\mathcal{A}^* = \{j \in [D] \mid \beta_j^* \neq 0\}$ と表す．凸最適化理論 [7] では，最適解 $\boldsymbol{\beta}^*$ がアクティブでない特徴に依存しないことが知られている．近年注目を集めているセーフスクリーニングと呼ばれるアプローチを用いると，最適化問題を解く前にアクティブになりうる特徴の集合 $\mathcal{A} \supseteq \mathcal{A}^*$ を見つけることができる．

セーフスクリーニングの考え方を説明するため，式 (2) の双対問題を考える．

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^n} \ \mathcal{D}_\lambda(\boldsymbol{\gamma}) = \frac{1}{2}\|\boldsymbol{\gamma} - \lambda^{-1}\boldsymbol{y}\|_2^2 \quad \text{s.t. } |\boldsymbol{z}_j^\top \boldsymbol{\gamma}| \leq 1 \ \forall\, j \in [D], \tag{3}$$

ここで，$\mathcal{D}_\lambda(\cdot)$ は双対問題の目的関数，$\boldsymbol{\gamma} \in \mathbb{R}^n$ は双対変数，$\boldsymbol{y} = [y^1, \ldots, y^n]^\top \in \mathbb{R}^n$，$\boldsymbol{z}_j = [x_j^1, \ldots, x_j^n] \in \mathbb{R}^n, \forall j \in [D]$ である．凸最適化理論に基づくと，双対問題 (3) の最適解 $\boldsymbol{\gamma}^*$ を用いて，

$$|\boldsymbol{z}_j^\top \boldsymbol{\gamma}^*| < 1 \ \Rightarrow \ \beta_j^* = 0, j \in [D]. \tag{4}$$

が成り立つことが知られている．セーフスクリーニングの主なアイデアは，最適化問題を解くことなく，各特徴 $j \in [D]$ に対して，$\boldsymbol{z}_j^\top \boldsymbol{\gamma}^*$ の上界 $\mathrm{UB}(|\boldsymbol{z}_j^\top \boldsymbol{\gamma}^*|)$ を効率的に求めることである．この上界を用いると，$\mathrm{UB}(|\boldsymbol{z}_j^\top \boldsymbol{\gamma}^*|) < 1 \Rightarrow \beta_j^* = 0$ が言え，非アクティブな特徴を事前に同定できる．

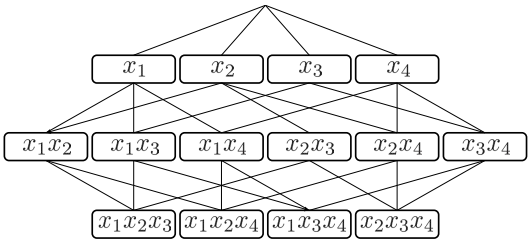

図 2: $d = 4$, $r = 3$ の場合の高次交互作用モデルの木構造 $\mathcal{T}$

しかしながら，(1) 式の高次交互作用モデルでは，D 個のパラメータ $\beta_1, \ldots, \beta_D$ それぞれに対して，上界 $\mathrm{UB}(|\boldsymbol{z}_j^\top \boldsymbol{\gamma}^*|)$ を計算できない．そこで，高次交互作用モデルの各項が図 2 のような木構造 $\mathcal{T}$ で表現できることに着目する．図 2 の木構造 $\mathcal{T}$ では，高次交互作用モデルの各項が各ノードに対応しているため，ノードのインデックスも $j \in \{1, \ldots, D\}$ と表記する．あるノード $j \in [D]$ に対し，その部分木を $\mathcal{T}_{\mathrm{sub}}(j) \subseteq [D]$ と表記する．このとき，以下が成り立つ．

定理 1 主問題 (2) の任意の実行可能解 $\tilde{\boldsymbol{\beta}}$ と双対問題 (3) の任意の実行可能解 $\tilde{\boldsymbol{\gamma}}$ が与えられたとき，任意の $j' \in \mathcal{T}_{\mathrm{sub}}(j)$ に対して次のルールが成り立つ．

$$\mathrm{SPR}(j) = \max\left\{ \sum_{i: z_j^i \tilde{\gamma}_i > 0} z_j^i \tilde{\gamma}_i, - \sum_{i: z_j^i \tilde{\gamma}_i < 0} z_j^i \tilde{\gamma}_i \right\} + \lambda^{-1} \sqrt{2(\mathcal{P}_\lambda(\tilde{\boldsymbol{\beta}}) - \mathcal{D}_\lambda(\tilde{\boldsymbol{\gamma}})) \sum_{i \in [n]} (z_j^i)^2} < 1 \Rightarrow \beta^*_{j'} = 0.$$

定理 1 では，図 2 の任意のノード j において $\mathrm{SPR}(j)$ を計算し，その値が 1 未満であれば，その子孫ノードに対応する特徴が最適解においてアクティブでないことを保証できることを示唆している．この性質を活用すれば，図 2 の木構造を探索し，$\mathrm{SPR}(j) < 1$ となるようなノードで枝刈りを実施していけば，枝刈りされずに残ったノードに対応する特徴のみが最適解においてアクティブになりうることを意味している．

3 Selective Inference によるスパース高次交互作用モデルの検定

以下では，文献 [2] で筆者らが提案したスパース高次交互作用モデルの同定アルゴリズムを簡単に紹介する．L_1 正則化などによって特徴選択が行われた後，アクティブな特徴のみを用いて作成したモデルに対しては，選択バイアスのため，通常の統計的推測を行うことができない．この問題は古くから指摘されていたが，最近まで有効な対処方法は提案されていなかった．2013 年に arXiv に投稿された Lee らの論文 [6] をきっかけに Selective Inference と呼ばれるアプローチがモデル選択後の統計的推測に有効であることがわかり，注目を集めている．

特定の特徴選択アルゴリズムによって選択された特徴の集合を $S \subseteq [D]$ とする．Lee らのアプローチを用いると，その特徴選択アルゴリズムが $\mathcal{S}$ を選択したイベントが，$\boldsymbol{y}$ の線形イベントして記述できる，すなわち，ある行列 A とベクトル $\boldsymbol{b}$ を用いて，$A\boldsymbol{y} \leq \boldsymbol{b}$ と書けるならば，S のみを用いた線形モデルのパラメータの標本分布を厳密に求めることができる．紙面の都合上詳細は割愛するが，前節で考察した L_1 正則化（LASSO），Maginal Screening (MS), Orthogonal Matching Pursuit (OMP) など多くの特徴選択アルゴリズムにおいて，特徴選択イベントを線形イベントとして記述できることが確認できる．

選択された特徴 S のみから成る線形モデルの最小二乗推定量を $\hat{\boldsymbol{\beta}}_S$ とする．もしも特徴がデータに依存せずに選ばれたものであれば，帰無仮説 $\beta^*_{S,j} = 0, j \in S$ のもと，有意水準 α における第一種の過誤を制御する，すなわち，$\Pr(\hat{\beta}_{S,j} \notin [\ell_{\alpha/2}, u_{\alpha/2}]) \leq \alpha$ を満たすような棄却限界値 $\ell_{\alpha/2}, u_{\alpha/2}$ を求めることが

できる．一方，特徴がデータに依存して選ばれた状況では，以下のような条件付きの第一種の過誤を制御することで選択バイアスを補正した統計的推測が可能となる．

$$\Pr(\hat{\beta}_{S,j} \notin [\tilde{\ell}_{\alpha/2}, \tilde{u}_{\alpha/2}] \mid \{S \text{ is selected}\}) \leq \alpha \tag{5}$$

Lee らのアプローチを用いると，特徴選択イベント $\{S \text{ is selected}\}$ が線形イベントとして記述できる，すなわち，式 (5) が

$$\Pr(\hat{\beta}_{S,j} \notin [\tilde{\ell}_{\alpha/2}, \tilde{u}_{\alpha/2}] \mid A\boldsymbol{y} \leq \boldsymbol{b}\}) \leq \alpha \tag{6}$$

と表されるとき，補正された棄却限界値 $\tilde{\ell}_{\alpha/2}, \tilde{u}_{\alpha/2}$ を陽に求めることができる．式 (6) の補正棄却限界値 $\tilde{\ell}_{\alpha/2}, \tilde{u}_{\alpha/2}$ は，それぞれ，n 次元空間の多面体 $\{\boldsymbol{y} \in \mathbb{R}^n \mid A\boldsymbol{y} \leq \boldsymbol{b}\}$ によって定義される切断正規分布の下側 $\alpha/2$ 点，上側 $\alpha/2$ 点となる．

しかしながら，本研究で対象とする高次交互作用モデルでは，特徴選択イベントが膨大な数の線形制約によって記述される．すなわち，n 次元空間の多面体 $\{\boldsymbol{y} \in \mathbb{R}^n \mid A\boldsymbol{y} \leq \boldsymbol{b}\}$ を構成する線形制約の数が d や r に対して指数的に増加する．したがって，Lee らの Selective Inference のアプローチを利用して，補正棄却限界値 $\tilde{\ell}_{\alpha/2}, \tilde{u}_{\alpha/2}$ を求めることが現実的に難しい．筆者らが文献 [2] で提案した方法では，図 2 の木構造を再び利用することにより，検定統計量の補正された棄却限界値 $\tilde{\ell}_{\alpha/2}, \tilde{u}_{\alpha/2}$ に影響を与えない線形制約を枝刈りによって除去することができる．紙面の都合上詳細は割愛するが，筆者らの提案法を用いると，スパース高次交互作用モデルの係数の統計的有意性を，式 (5) の第一種の過誤を α 未満に制御するという意味において，正確に検定することができる．

4　おわりに

本講演においては，スパース高次交互作用モデルの同定法 [1] と検定法 [2] の基本アイデアを説明し，数値実験などを通してこれらの方法の有効性を示す．

参考文献

[1] Kazuya Nakagawa, Shinya Suzumura, Masayuki Karasuyama, Koji Tsuda, and Ichiro Takeuchi. Safe pattern pruning: An efficient approach for predictive pattern mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1785–1794. ACM, 2016.

[2] Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, and Ichiro Takeuchi. Selective inference for sparse high-order interaction models. In *International Conference on Machine Learning*, pages 3338–3347, 2017.

[3] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization*, 2012.

[4] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *Advances in Neural Information Processing Systems*, pages 811–819, 2015.

[5] Atsushi Shibagaki, Masayuki Karasuyama, Kohei Hatano, and Ichiro Takeuchi. Simultaneous safe screening of features and samples in doubly sparse modeling. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1577–1586, 2016.

[6] Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[7] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.