研究論文

交互作用効果の計量化および可視化に関する考察

岩沢宏和[1]　　大塚忠義[2]

2023 年 9 月 5 日投稿
2024 年 1 月 17 日受理

Abstract

機械学習モデルは高い予測力を示すが，予測の際の説明力に欠けている．現在，機械学習モデルを解釈可能なものとするための手法として，Interpretable Machine Learning (IML) 分野の研究が行われているが，その手法はアクチュアリーがその携わる業務で機械学習モデルを活用するには不十分である．しかし，予測精度の高いモデルを将来にわたって活用しないことはアクチュアリー業務の発展を阻害する．本稿の目的は，機械学習モデルの説明力をアクチュアリー業務に活用できるレベルに引き上げるために必要な手法および指標を提案し，基礎的な考察を行うことである．

本稿では,交互作用効果を定義したうえで，その諸性質を確認し交互作用効果を計量化するための指標として，交互作用効果割合，交互作用効果量，未解釈割合，交互作用効果重要度を提案した．そして，既存手法である PD と ALE は主効果，交互作用効果のいずれもうまく捉えられていないことを明らかにしたうえで，これらに替わる IML 手法として CE，TCE，SI を提案し，それらの指標を用いて既存の手法および代替手法の特徴を明らかにした．加えて，既存の手法と代替手法を単純なモデルに適用して比較分析することで，既存手法の問題点と代替手法の可能性を示した．また，これらの検討結果を踏まえ，アクチュアリーがその実務に活用するために，主効果項のみを使ってできるだけ多くの部分の解釈ができる手法を活用することを提案する．

Keywords：交互作用（Interaction effect），IML，関数分解（functional decomposition），ALE，PD

1　はじめに

機械学習モデルをはじめとするデータサイエンス技術は，驚異的な速度で発展を続けている．交互作用効果が強いデータに対して機械学習モデルは高い予測力を示しており，その予測精度は伝統的な回帰分析等の手法に基づく予測とは比較にならない．一方で，機械学習モデルはブラックボックスモデルになりがちで，その内容を解釈することは困難である．そもそも機械学習モデルの主たるユーザーのニーズは予測精度の高さにあり，予測結果に対する高い説明力は求められてこなかった．一方で，機械学習モデルを解釈可能なものとするための技術に対する要望もある．機械学習モデルが捉えた交互作用効果を可視化・計量化することによって予測結果の解釈に役立てる手法として，Interpretable Machine Learning (IML) 分野の研究が行われている．特に，交互作用効果を可視化する方法としては Partial Dependence (PD) と Accumulated Local Effects (ALE) が知られている．PD にはさまざまな欠点があ

[1] 早稲田大学大学院会計研究科　東京都新宿区西早稲田 1-6-1, Email: iwahiro@bb.mbn.or.jp
[2] 早稲田大学大学院会計研究科　東京都新宿区西早稲田 1-6-1, Email: otsukata@waseda.jp

ることが知られ，その代替案としてALEが提案された．しかしながら，交互作用効果を可視化するツールとして生まれたPD，ALEのどちらも，交互作用効果を正確に計量化できているかは疑問である．

一方で，現在の交互作用効果の計量化手法では，アクチュアリーがその携わる業務で機械学習モデルを広く活用するには不十分である．アクチュアリーにはプライシングや将来収支分析などの業務を遂行していく過程で取締役会，規制当局，そして契約者に対する説明責任を有しているため，解釈が困難または不十分なモデルをそのまま活用することはできない．筆者が先行研究調査を進めたところ，交互作用効果は，機械学習に関連して用いられる際に専門家の間で広く共通して使われる定義が存在せず，その基本的なことも解明されていない，または少なくとも広く知られていないことが分かった．さらには，交互作用効果以前に，主効果についても計量化した結果が正確であるといえる測定方法が確立されていないことが分かった．このような現状のもとでは，機械学習モデルの予測精度が高くてもアクチュアリー業務に活用できないのはやむを得ない．一方で，予測精度の高いモデルを将来にわたって活用しないことはアクチュアリー業務の発展を阻害することに繋がる．

本稿の目的はこのような現状に風穴を開け，アクチュアリー業務に機械学習モデルを活用するために必要な議論を行うことである．そのために，機械学習モデルの説明力をアクチュアリー業務に活用できるレベルに引き上げるのに必要な手法および指標を提案し，基礎的な考察を行う．

本稿の次節以降の構成は以下のとおりである．第2節では，交互作用効果に明確な定義を与えるとともに，本稿で使用する記号を定義する．加えて，IML分野の先行研究・手法を紹介する．第3節では，主効果および交互作用効果の諸性質を確認したえうで，それらを捉える手法の必要性とPDがそのニーズを満たしたしていないことを明らかにするとともに，主効果および交互作用効果を計量化するため有用な指標として交互作用効果割合，交互作用効果量，未解釈割合，交互作用効果重要度を提案する．第4節では，ALEの問題点について論じ，PD，ALEの代替手法としてConditional Expectation Based Decomposition (CEBD)，True Conditional Expectation Based Decomposition (TCE)，Shapley Interaction Effect Decomposition (SIED)，Minimum Interaction Effect Decomposition (MIED)を提案する．続いて第5節では，第3節で提案した指標を用いて，各手法が捉える主効果，交互作用効果を，事例をもとに計量化・可視化することで，主効果が解析可能であることの重要性および主効果や交互作用効果の基本事項を確認するとともに，　それぞれの手法の解釈可能性に関する比較を行うことで，提案した代替手法が優位であることを主張する．そして，第6節で，分析結果を考察し，本稿の提案の新規性を主張する．そして，アクチュアリーが機械学習モデルをはじめとするブラックボックスモデルを活用するためのモデルの解釈に関する提言を行い，残された課題を整理する．

2　先行手法

本節では，本稿全体に共通して数式内で用いる記号を定義したうえで，予測モデルを解釈する手法である関数分解(functional decomposition）という概念に基づき，先行手法としてPD，ALE，H統計量，SHAPを紹介する．予測モデルは，観測対象の特徴量変数ベクトルを入力すると，予測値を出力する．その入出力関係を関数として捉え，その関数を（その予測モデルの）「予測関数」とよぶ．関数分解では予測関数を主効果項と交互作用効果項に分解し，それぞれから出力される情報を解釈する．

2.1　記号と交互作用効果の定義

IML (Adadi and Berrada [2018]，Biecek and Burzykowski [2021]，Molnar [2022]．アクチュアリー分野では

Lorentzen and Mayer [2020]や Baeder et al. [2021]）の手法のうち，本稿で扱う関数分解については，特に注目すべき先行文献として Apley and Zhu [2016, 2020]があるので，比較参照の便宜のため，以下で使う記号等は，それらに準じられるところはできるだけ準じている．

特徴量変数の添え字の集合を$D := \{1, \dots, d\}$とし，特徴量変数を$x_1, \dots, x_d$とする．変数を確率変数として扱うときは大文字とする．空集合でない$J \subseteq D$について，$\mathbf{x}_J := \{x_j \mid j \in J\}$とし，特に，$\mathbf{x}_D := \{x_1, \dots, x_d\}$である．これらは（値の集合ではなく）変数の集合である．ただし，誤解の余地のないときは，特徴量変数ベクトルとしても用いる．

$J \subseteq D$について，$\setminus J := D \setminus J$（差集合）とする．たとえば，

$$\mathbf{x}_{\setminus\{j,l\}} = \mathbf{x}_D \setminus \mathbf{x}_{\{j,l\}} = \{x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_{l-1}, x_{l+1}, \dots, x_d\}$$

であり，また，特に$\setminus j := D \setminus \{j\}$とする．したがって，たとえば，

$$\mathbf{x}_{\setminus j} = \mathbf{x}_{\setminus\{j\}} = \{x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d\}$$

である．

IML の計算に用いるデータの標本サイズはnとし，

$$X = (x_{ij}) = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix} =: \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}$$

と表す．このデータは，典型的には予測モデルを構築するときの学習データであるが，実用上は，学習データ以外のデータや，学習データの一部をランダムに抽出したものが用いられる場合がある．

予測モデルの予測関数をfとする．予測関数は，典型的にはブラックボックスだが，d個すべての特徴量変数の具体的な値を（実データであってもなくてもよいので）入力すると，すぐに予測値を返してくれるものである．ただし，実用上は，予測関数fそのものでなく，何らかのリンク関数g（たとえば 2 値分類問題に対してロジットリンク関数）による$g \circ f$を以下のfの代わりに用いてもよい．同様に，入力側についても，前処理などによって特徴量変数を変換したものが実際の予測関数の入力になっている場合には，もとの特徴量変数と変換後の変数のうち解釈に用いたいものの方を以下での$x_1, \dots, x_d$と考える．すなわち，実用上は，以下における予測関数の入力および出力は，実際に解釈に用いたいもののことであり，文字通りの特徴量変数や予測値である必要はない．

そうした予測関数fを解釈するために，$f(\mathbf{x}_D)$は，IML の手法によって，定数$f_\emptyset := \mathbb{E}[f(\mathbf{X}_D)]$と適当な 1 変数関数$f_j(x_j)$（$j \in D$）によって次のように近似される場合がある．

$$f(\mathbf{x}_D) \fallingdotseq f_\emptyset + \sum_{j \in D} f_j(x_j)$$

そのときの$f_\emptyset$を「0 次の効果の項」といい，$f_j(x_j)$を「1 次の効果の項」もしくは「主効果項」という．伝統的には，主効果項は予測モデル自身の表現に現れるものとして定義されてきたが，ブラックボックスモデルでは一般にはそのような項はなく IML の手法によって見いだされるものになる．

上式の両辺には一般に差があるが，その差を以下のように表すことができ，これを「交互作用効果」と定義する．

$$\text{交互作用効果} = f(\mathbf{x}_D) - f_\emptyset - \sum_{j \in D} f_j(x_j)$$

この定義は，伝統的なものとは異なる．伝統的には 先に 2 つ（以上）の変数を特定してから，その間の効果として定義され，項として表現される（Jaccard and Turrisi [2003]）．これに対し本稿では，原理的に考えられるそれらの効果の総和を交互作用効果と定義した．このように定義を変えるのは，(i) 伝統的には，（たとえば線形の）モデルの表現が先にあって，その各項について直接論ずることができたのに対して，ブラックボックスモデルでは，項ありきで考えることができず，また，(ii) わかりやすさのためにはブラックボックスモデルをまずは主効果項を用いて解

釈するのが自然であり，その場合の残差に名前をつけておくことが便利だからである．

この交互作用効果がさらに，IML の手法によって，2 変数関数$f_{\{j,l\}}(x_j, x_l)$ ($\{j,l\} \subseteq D$) の和$\sum_{\{j,l\}\subseteq D} f_{\{j,l\}}(x_j, x_l)$で近似される場合もある．そのときの各項（2 変数関数）を「2 次の交互作用効果の項」という．

この考えを発展させると，予測関数fの関数分解（functional decomposition）として次のようなものを考えることができる．

$$f(\mathbf{x}_D) = \sum_{J\subseteq D} f_J(\mathbf{x}_J) = f_\emptyset + \sum_{j\in D} f_j(x_j) + \sum_{\{j,l\}\subseteq D} f_{\{j,l\}}(x_j, x_l) + \cdots + \sum_{j\subseteq D} f_{\setminus j}(\mathbf{x}_{\setminus j}) + f_D(\mathbf{x}_D)$$

ここで，$f_J(\mathbf{x}_J)$は$|J|$次の（交互作用）効果を表すと考えられる．後述するとおり，既存の手法についていえば，1 次元 PD はここでいう 1 次以下の効果，2 次元 PD はここでいう 2 次以下の効果，1 次元 ALE はここでいう 1 次（のみ）の効果，2 次元 ALE はここでいう 2 次（のみ）の効果に対応するものとして理解できる．ただし，以下ではこの分解のアプローチを用いるため，PD に対しても，その導入をする際に，1 次のみの効果や 2 次のみの効果の式も与える．

2.2 Partial Dependence (PD)

PD は，Friedman [2001] の中で提案された．関数自体は，3 次元以上でも定義されるが可視化のためには使い難く，実用上は 1 次元 PD と 2 次元 PD のみが扱われ，本稿でもその 2 つだけ扱う．1 次元（2 次元）PD は，予測関数における各特徴量変数（特徴量変数 2 個の各組）の 1 次（2 次）以下の効果を可視化するのに用いられる．そのプロットを PDP (Partial Dependence Plot という意味）という．

1 次元（2 次元）PD 自体は，個々の観測対象iに対して，着目している特徴量変数の値x_j（2 つの特徴量変数の値の組（x_j, x_l)）だけを（とりうる値の）全域に変化させ，そのほかの特徴量変数の値はそのままとして得られる予測値をプロットした曲線（曲面）の関数を平均して得られる 1 変数（2 変数）関数である．

理論的には，1 次元 PD の関数は$\mathrm{PD}_j(x_j) \coloneqq \mathbb{E}[f(x_j, \mathbf{X}_{\setminus j})]$と定義され，2 次元 PD の関数は$\mathrm{PD}_{\{j,l\}}(x_j, x_l) \coloneqq \mathbb{E}[f(x_j, x_l, \mathbf{X}_{\setminus\{j,l\}})]$と定義される．実用上は，$\mathrm{PD}_j(x_j)$の各点の値は，$\widehat{\mathrm{PD}}_j(x_j) = \frac{1}{n}\sum_{i=1}^n f(x_j, \mathbf{x}_{i,\setminus j})$で推定され，$\mathrm{PD}_{\{j,l\}}(x_j, x_l)$の各点の値は，$\widehat{\mathrm{PD}}_{\{j,l\}}(x_j, x_l) = \frac{1}{n}\sum_{i=1}^n f(x_j, x_l, \mathbf{x}_{i,\setminus\{j,l\}})$で推定される．

しかし本稿では，予測関数の分解というアプローチをとるため，次のように，1 次のみの項$f_{j,\mathrm{PD}}(x_j)$と 2 次のみの項$f_{\{j,l\},\mathrm{PD}}(x_j, x_l)$も定義する．

$$f_{j,\mathrm{PD}}(x_j) \coloneqq \mathrm{PD}_j(x_j) - \mathbb{E}[\mathrm{PD}_j(X_j)],$$
$$f_{\{j,l\},\mathrm{PD}}(x_j, x_l) \coloneqq \mathrm{PD}_{\{j,l\}}(x_j, x_l) - \mathbb{E}[\mathrm{PD}_{\{j,l\}}(X_j, X_l)] - f_{j,\mathrm{PD}}(x_j) - f_{l,\mathrm{PD}}(x_l).$$

ここで，2 次のみの項は，後述の H 統計量において暗黙に想定されているものと整合的になるように定義したものである．

PD の大きな欠点は計算コストがきわめて高いことである．それゆえ代替手法の存在が強く望まれる．また，性質上の問題でよく指摘されるのは，特徴量変数どうしが独立でない限り，結果として得られる関数は非現実的なデータに基づいて求められる可能性があるため，特徴量変数どうしの相関が強い場合に不適切だと考えられることである．

2.3 Friedman の H 統計量

Friedman の H 統計量（以下，「H 統計量」とよぶ）は，Friedman and Popescu [2008] で提案された．また，「正規化しない」ものが Inglis et al. [2022]で提案されている．もとのものは，着目している 2 つの特徴量変数の間に交互作用効果があるかないかの尺度であった．それに対して，正規化しないものは，予測全体に対する交互作用効果の

相対的な大きさを計量化するものである．

H 統計量は，「2 次元 PD は 2 次以下の効果をすべて捉えており，かつ，1 次元 PD は 1 次以下の効果をすべて捉えている」と想定して，2 次元 PD と（2 つの）1 次元 PD の合計との差をもとに尺度を作るものである．Friedman のオリジナルの尺度は次のとおりである．

$$H_{jl}^2 := \frac{\sum_{i=1}^n f_{\{j,l\},\mathrm{PD}}\left(x_{ij}, x_{il}\right)^2}{\sum_{i=1}^n \left(f_{\{j,l\},\mathrm{PD}}\left(x_{ij}, x_{il}\right) + f_{j,\mathrm{PD}}\left(x_{ij}\right) + f_{l,\mathrm{PD}}\left(x_{il}\right)\right)^2}$$

これに対して，本稿が関心をもつのは，Inglis et al. [2022]で提案された次の尺度である．

$$H_{jl}^* := \sqrt{\frac{1}{n}\sum_{i=1}^n f_{\{j,l\},\mathrm{PD}}(x_{ij}, x_{il})^2}$$

この尺度は，the square-root of the average unnormalized (numerator only) version of Friedman's H^2 である．後述するように，この考え方は，PD 以外の手法にも自然と応用できる．

2.4 Accumulated Local Effects (ALE)

ALE は，Apley and Zhu [2016, 2020] で PD に代わるものとして提案された．1 次元（2 次元）ALE は，予測関数における各特徴量変数（特徴量変数 2 個の各組）の寄与を 1 変数（2 変数）関数で表現するものであり，それを可視化したものは ALEPlot とよばれる．ただし，PD が 1 次（2 次）以下の効果を表現するものだったのに対して，ALE は 1 次（2 次）のみの効果を表現するものである．

1 次元（2 次元）ALE を求めるには，まず，着目している特徴量変数の値x_j（2 つの特徴量変数の値の組(x_j, x_l)）だけ（全域でなく）微小に変化させて得られる予測値の変化を表す関数を平均したものを局所（local）ごとに求め，それを累積（accumulate）して，最後に，ある種の中心化を施す．ここでいう「中心化」とは，1 次元 ALE については，平均の値が 0 となるように定数を加減し，また，2 次元 ALE については，平均の値が 0 となるだけでなく，2 つの特徴量変数のどちらの主効果も 0 となるように 2 つの 1 変数関数を加減するものである．特に，1 次元の方は，定数を加減するだけのことであり，本稿では，あとでその具体的計算例も出てくる．それに対し，2 次の効果に対する中心化については，一般的な形のものを「2 変数関数に対する中心化」として Appendix A.1 で定式化するが，以下では詳細が必要となる計算例は登場しない．

理論的には（予測関数が十分に微分可能であれば），1 次元の ALE の関数$f_{j,\mathrm{ALE}}\left(x_j\right)$（$x_j$の関数）は，$f^j\left(x_j, \mathbf{x}_{\setminus j}\right) := \frac{\partial f\left(x_j, \mathbf{x}_{\setminus j}\right)}{\partial x_j}$として，

$$\int_{\min\{x_j\}}^{x_j} \mathbb{E}\left[f^j\left(X_j, \mathbf{X}_{\setminus j}\right) \middle| X_j = z_j\right] dz_j$$

に中心化を施したものであり，2 次元の ALE の関数$f_{\{j,l\},\mathrm{ALE}}\left(x_j, x_l\right)$（$x_j, x_l$の関数）は，$f^{\{j,l\}}\left(x_j, x_l, \mathbf{x}_{\setminus\{j,l\}}\right) := \frac{\partial^2 f\left(x_j, x_l, \mathbf{x}_{\setminus\{j,l\}}\right)}{\partial x_j \partial x_l}$として，

$$\int_{\min\{x_l\}}^{x_l} \int_{\min\{x_j\}}^{x_j} \mathbb{E}\left[f^{\{j,l\}}\left(X_j, X_l, \mathbf{X}_{\setminus\{j,l\}}\right) \middle| (X_j, X_l) = (z_j, z_l)\right] dz_j\, dz_l$$

に中心化を施したものである．ここで，$\min\{x_j\}$および$\min\{x_l\}$はそれぞれ，変数x_jおよびx_lの取りうる値の実質的な下限近くから適当に選んだ値（たとえば学習データから得られた最小値）とする．ただし，計算結果に中心化を施すため，それらの値の選び方は，取りうる値の実質的な範囲を含むようにしている限り，ALE の計算結果に影響は及

ぼさない．なお，理論的考察をする場合など，下限が$-\infty$となる場合には，1 次元の場合で例示すれば，取りうる値の範囲に属する値a_jを$\min\{x_j\}$の値として定義した場合の ALE の関数を$f_{j,\mathrm{ALE},a_j}(x_j)$として，$f_{j,\mathrm{ALE}}(x_j) \coloneqq \lim_{a_j \to -\infty} f_{j,\mathrm{ALE},a_j}(x_j)$と定義する．

実際には，ブラックボックスモデルの予測関数の偏微分はできないので，偏微分の代わりに適当な区分に基づく差分を用い，積分の代わりに全区分についての総和を用いる．その具体的な定式は，やや煩雑になる．実用上は，そうした関数の各点は，$\mathbf{x}_D$が取りうる値の範囲内の一部の点に対する予測値のデータをもとにした推定によって求めることになるが，その定式はさらに煩雑である．また，以上のみの説明では，特徴量変数にカテゴリー変数があった場合（特に，順序がつかない名義変数の場合）の取り扱いについては明らかではないが，本稿で具体的に示す計算例では登場しないので説明を省略する（原論文では説明されている）．

歴史上は，PD がもつ理論的な欠点を解消しようとして提案されたのが ALE であったが，ALE にも疑問がある．というのも，ALE では，局所的に条件付平均をとったものを累積するのであるが，そのときに得られるものが大域的に見て何であるかは明らかでない．原論文でも安易に大域的に解釈することを“wrong way”といって警告している．しかし，もしそうした「解釈」を許さないものだとしたら， ALE は，ほしかったわかりやすい解釈を与えるものにそもそもなっていないといえる．

2.5 SHapley Additive exPlanations (SHAP)

SHAP は，Lundberg and Lee [2017] で提案された．その中で，具体的な近似計算方法の 1 つとして KernelSHAP も提案された．計算方法はほかにも多数提案されている．たとえば，実用上で広く使われている TreeSHAP は，元々は Lundberg et al. [2018] で提案され，その後，Lundberg et al. [2020] で少し異なる内容のものが改めて提案されており，現在，Python の shap パッケージや R の treeshap パッケージに実装されているのは後者のものである．これらは，各観測対象iの予測値$f(\mathbf{x}_i)$を，$\phi_0 \coloneqq \mathbb{E}[f(\mathbf{X}_D)]$と，$j = 1, \ldots, d$に対する特徴量変数$x_j$の貢献$\phi_{ij}$（正も負もありうる）に分解して，

$$f(\mathbf{x}_i) = \phi_0 + \sum_{j=1}^{d} \phi_{ij}$$

とするものである．これにより，局所的な情報が多量に得られるので，それらを組み合わせたり集計したりすることで，さまざまな可視化が可能であり，また，本稿が着目する使用法でいえば，集計によって，（個々の予測値の分解だけでなく）予測関数全体の分解にも利用可能である．

SHAP の発想自体は非常にシンプルである．それは，各貢献を，ゲーム理論でいう Shapley 値で評価しようというものである．その原理上の値は次式で与えられる．

$$\phi_{ij} \coloneqq \sum_{J \subseteq (\setminus j)} \frac{|J|!\,(d - |J| - 1)!}{d!} \left(g_i(J \cup \{j\}) - g_i(J)\right)$$

ただし，SHAP の原理上の定義において，この式のg_iを何とすべきかについては議論がある（Janzing et al. [2020]）．

原論文（Lundberg and Lee [2017]）では$g_i(J) = \mathbb{E}\left[f(\boldsymbol{X}) | \mathbf{X}_J = \mathbf{x}_{iJ}\right]$（条件付期待値）と考えられており，それに基づけば，SHAP の定義は次のとおりである．

$$\phi_{ij} \coloneqq \sum_{J \subseteq (\setminus j)} \frac{|J|!\,(d - |J| - 1)!}{d!} \left(\mathbb{E}\left[f(\mathbf{X}_D) | \mathbf{X}_{J \cup \{j\}} = \mathbf{x}_{i,(J \cup \{j\})}\right] - \mathbb{E}\left[f(\mathbf{X}_D) | \mathbf{X}_J = \mathbf{x}_{iJ}\right]\right)$$

これに対して，$g_i(J) = \mathbb{E}\left[f\left(\mathbf{x}_{iJ}, \boldsymbol{X}_{\setminus J}\right)\right]$（期待値）とする考えがある．ただし，「本来あるべきもの」として SHAP の

定義を期待値とする場合もある一方で，「条件付期待値の近似」として期待値をとる場合もある．いずれにせよ，このg_iを使った場合の SHAP の定義は次のとおりである．

$$\phi_{ij} := \sum_{J \subseteq (\setminus j)} \frac{|J|!\,(d-|J|-1)!}{d!} \left(\mathbb{E}\left[f\left(\mathbf{x}_{i,(J \cup \{j\})}, \boldsymbol{X}_{\setminus (J \cup \{j\})} \right) \right] - \mathbb{E}\left[f\left(\mathbf{x}_{iJ}, \boldsymbol{X}_{\setminus J} \right) \right] \right)$$

いろいろと提案されている実用上の計算方法は，ともかくもどちらかの近似となっている．たとえば，KernelSHAP は（提案者は理論値としては前者を考えているが，実際の近似計算上は）後者の近似をするものであり，TreeSHAP は Lundberg et al. [2018] の提案では前者の近似をするものであり，Lundberg et al. [2020] の提案では，どちらの近似のアルゴリズムも提示されている．

上で示した原理上の Shapley 値の計算は，計算時間がいわゆる指数関数時間になるので非現実的である．それゆえ近似計算がとられるが，それでも計算コストは高い．また，計算結果は，誤解を招きやすいものになる場合があることも指摘されている（Molnar [2022]）．いずれにせよ SHAP は，元々は個々の予測値の解釈用であり，直接に関数分解に用いることができる手法とは単純な比較はできない．しかし，理論値どおりに計算した SHAP を関数分解に利用したとした場合に対応する手法は，次節での議論でとりあげる．

3　交互作用効果を計量化するための指標の提案

3.1　主効果および交互作用効果を捉える手法の必要性

次の予測関数をもつ線形モデルを考える．

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

このモデルにおいて$\beta_1 x_1$は（IML 手法による関数分解の結果ではないので本論文の定義での主効果項とは必ずしも一致しないが）形式上はx_1の主効果項に見える．それでは，このモデルは，X_1とX_2が独立で$\beta_1 > 0$であるとき，「X_1が大きくなるほど目的変数yは大きくなる傾向がある」ということを含意しているであろうか．

答えは否である．反例は，$\beta_{12}\mathbb{E}[X_2]$が負で，その絶対値が十分に大きい場合を考えてみればよい．

IML の観点からすれば，X_1とX_2が独立なら，fは，各項に中心化（平均が 0 となるようにする）を施して，次のとおりに分解すべきである．

$$f_\emptyset = \mathbb{E}[f(X_1, X_2)], \quad f_1 = (\beta_1 + \beta_{12}\mathbb{E}[X_2])(x_1 - \mathbb{E}[X_1]), \quad f_2 = (\beta_2 + \beta_{12}\mathbb{E}[X_1])(x_2 - \mathbb{E}[X_2]),$$
$$f_{\{1,2\}} = \beta_{12}(x_1 - \mathbb{E}[X_1])(x_2 - \mathbb{E}[X_2])$$

事実，PD ではこのように分解するし，ALE による分解も（また，本論文が提案するどの手法による分解も），特徴量変数どうしが独立なときは PD の結果と一致する．この分解は，モデリングの手順において，モデルを構築する前に特徴量変数を標準化するというアプローチを正当化するものでもある．ただし，特徴量変数どうしが独立でない場合にどう分解すべきかは議論が必要である．実際，手法次第で分解結果は一般に異なる．

いずれにせよ，この簡単な事例の教訓は，特徴量変数どうしが独立であったとしても，線形予測子に現れる見た目の主効果項は，一般には，本来の主効果を表すものではない，ということである．

次に交互作用効果項を理解するために，見た目が交互作用効果項のみからなる次の線形モデルを考える．

$$f(x_1, x_2) = x_1 x_2$$

1 つ前に見た例と違って，今度は，X_1とX_2は独立とは限らないものとする．また，$\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$だったとする．その場合，各変数に中心化はすでに施されている．すると，この場合は「主効果は（少なくとも平均的に見れば）0 である」といえるであろうか．

答えは否である．実際，たとえば，X_2の主効果は平均的には0である一方で，X_1が大きくなるほど目的変数yが大きくなる強い傾向があるという事例がありうる（図1参照）．この事例は後の5.1で詳しく見る．そこで見るとおり，この事例では実際に（大きな）主効果があると考えられる．

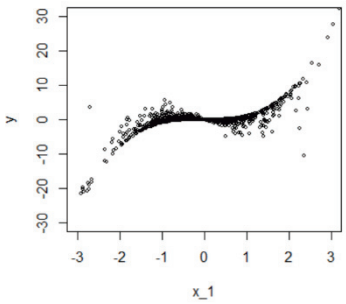

図1

それにもかかわらず，この事例の場合，PDによれば主効果は0であると判定される．線形モデルに対しては，PDはつねに，中心化を施したあとの（見た目の）主効果項のみをそのまま主効果と捉えてしまうからである．このことからPDはIML手法としては不適切であると主張することができる．それでは，どのような手法であれば主効果をうまく捉えられるであろうか．あとで見るように，ALEも主効果をうまく捉えられていないようである．したがって，既存のいずれの手法でも主効果をうまく捉えられない．しかし，うまく捉えられていないことはわかったとしても，代替として主効果をどう捉えるべきかについては明らかではなく，大いに議論が必要だと筆者は考えている．また，交互作用効果は，主効果では捉えきれないものとして定義されるので，その議論は，交互作用効果をどう捉えるかにとっても決定的に重要である．

そこで次項では，主効果や交互作用効果をどう捉えるべきかの議論を行うための道具立てを用意する．

3.2 交互作用効果割合，交互作用効果量，未解釈割合，交互作用効果重要度

予測モデルに基づく交互作用効果は，定義により

$$f(\mathbf{x}_D) - f_\emptyset - \sum_{j \in D} f_j(x_j)$$

で表される．したがって，単純にいえば，この値が0に近ければ交互作用効果を予測モデルに組み込むことによる影響度は相対的に小さく，0から離れていれば影響度が大きい．そこでそうした相対的な影響度を測る自然な指標として，

$$\widehat{\mathrm{IER}} := \frac{\sum_{i=1}^{n}\left(f(\mathbf{x}_{iD}) - f_\emptyset - \sum_{j \in D} f_j(x_{ij})\right)^2}{\sum_{i=1}^{n}(f(\mathbf{x}_{iD}) - f_\emptyset)^2}$$

というものが考えられる．これを「交互作用効果割合（Interaction Effect Ratio）」とよぶことにする．これに対応する理論値は，

$$\mathrm{IER} := \frac{\mathbb{E}\left[\left(f(\mathbf{X}_D) - f_\emptyset - \sum_{j \in D} f_j(X_j)\right)^2\right]}{\mathbb{E}[(f(\mathbf{X}_D) - f_\emptyset)^2]}$$

であり，これもまた「交互作用効果割合」とよぶ．

実際の場面では，$f_j(x_j)$は何らかの推定によって求められる．その結果，この割合の推定値にはおそらく正のバイアスがある．したがって，この割合は，未知の真の値が小さくても推定値が十分に小さくなるとは限らないが，推定値がすでにかなり小さい値であれば，真の割合も十分に小さいと判断することができる．

この割合は，予測モデル自体に依存するとともに，予測関数の関数分解手法にも依存する．同じ予測モデルに対しては，ほかの条件で優劣の差がないとすれば，その値が小さい関数分解手法ほど有用と考えることができる．また，関数分解手法を比較する際には割合を用いる必要はないので，

$$\widehat{\mathrm{IEQ}} := \frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_{iD}) - f_{\emptyset} - \sum_{j\in D} f_j(x_{ij})\right)^2$$

あるいは，その理論値

$$\mathrm{IEQ} := \mathbb{E}\left[\left(f(\mathbf{X}_D) - f_{\emptyset} - \sum_{j\in D} f_j(X_j)\right)^2\right]$$

を指標とすることができる．これらをどちらも「交互作用効果量（Interaction Effect Quantity)」とよぶ.

実用上の関数分解では，一部のf_Jだけ（たとえば，$|J| \leq 2$であるような$J \subseteq D$についてのみ）特定できているという状態が現実的である．そこで，f_Jが特定できているJの集合を$\mathcal{J} \subseteq (2^D \setminus \emptyset)$ (= Dの部分集合全体の集合から空集合を除いたもの）とすれば,

$$\widehat{\mathrm{UR}} := \frac{\sum_{i=1}^{n}\left(f(\mathbf{x}_{iD}) - f_{\emptyset} - \sum_{J\in\mathcal{J}} f_J(\mathbf{x}_{iJ})\right)^2}{\sum_{i=1}^{n}(f(\mathbf{x}_{iD}) - f_{\emptyset})^2}$$

や，その理論値

$$\mathrm{UR} := \frac{\mathbb{E}\left[\left(f(\mathbf{X}_D) - f_{\emptyset} - \sum_{J\in\mathcal{J}} f_J(\mathbf{X}_J)\right)^2\right]}{\mathbb{E}[(f(\mathbf{X}_D) - f_{\emptyset})^2]}$$

は，予測モデルのうちで，いわば解釈がまだできていない部分の割合を示している．これらを「未解釈割合 (Uninterpreted Ratio)」とよぶ.

ここまではモデルや手法に対する指標を提案したが，2 次の交互作用効果について考察するには，特徴量変数の組み合わせ$\{x_j, x_l\}$がもつ交互作用効果の予測における影響度（重要度）の指標も手法ごとに与えておくべきである．その先行手法は，（正規化なしの）H 統計量である．しかしながら，H 統計量は，PD をもとにしている点ですでに問題があり，さらに，予測関数の分解の観点でも，2 次元 PD と（2 つの）1 次元 PD の合計との差を 2 次の交互作用効果を表すものとしていることは不適切であると考える．後者のことが問題なのは，PD は ALE のような中心化を施していないので，2 次元 PD と（2 つの）1 次元 PD の合計との差には，一般に 1 次の効果が含まれてしまっているからである．したがって，関数$f_{\{j,l\},\mathrm{PD}}(x_{ij}, x_{il})$の代わりに，2 次未満の効果がないように中心化を施した関数$f_{\{j,l\},\mathrm{PD}^*}(x_{ij}, x_{il})$を定義し，それを用いて定義した次式を（正規化なしの）H 統計量の代わりに採用すべきである.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} f_{\{j,l\},\mathrm{PD}^*}(x_{ij}, x_{il})^2}$$

こうして修正した H 統計量を求めるには，注目している 2 つの特徴量変数の組がもつ 2 次の交互作用効果を表す関数の値の 2 乗$f_{\{j,l\},\mathrm{PD}^*}(x_j, x_l)^2$を求め，その結果を，計算対象とするデータ（典型的には学習データ）について平均し，その平方根をとる．これに対応する尺度は，PD 以外の手法にも自然と定義できる．具体的には，関数分解手法により，すでに中心化されている$f_{\{j,l\}}(x_j, x_l)$が得られているならば，予測における交互作用効果の影響の大きさの指標としては,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} f_{\{j,l\}}(x_{ij}, x_{il})^2}$$

を用いるのが自然である．中心化されているのであれば，実用上は，$\frac{1}{n}\sum_{i=1}^{n} f_{\{j,l\}}(x_{ij}, x_{il}) = 0$となるように定義されているので，この指標は，要するに，$f_{\{j,l\}}(x_j, x_l)$の標本標準偏差である.

しかしながら，本稿の文脈では，それらの指標は，最後に平方根をとったものとする必要はない．というのも，本

稿で提案している，交互作用効果割合，交互作用効果量，未解釈割合は，平方根をとらずに扱うのがふさわしいので，それに合わせた方が好都合だからである．そこで，交互作用効果の重要度の大きさの指標は，平方根をとらず，

$$\widehat{\mathrm{IEI}}_{\{j,l\}} := \frac{1}{n}\sum_{i=1}^{n} f_{\{j,l\}}(x_{ij}, x_{il})^2$$

または，その理論値

$$\mathrm{IEI}_{\{j,l\}} := \mathbb{E}\left[f_{\{j,l\}}(X_j, X_l)^2\right]$$

を用いるものとする．これらを「交互作用効果重要度（Interaction Effect Importance)」とよぶ．

4 ALE の問題点と代替手法の提案

4.1 ALE の不適切さ

ALE は，PD と違って，予測関数の見た目に主効果項がない場合でも主効果を捉えることができる．しかし，偏りがある．実際，特徴量変数どうしに単純な相関だけがある場合でさえ偏りがある．

たとえば，$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$であるときに，

$$f(x_1, x_2) = x_1 x_2$$

という予測関数を考えてみる．この事例では，$f_{1,\mathrm{ALE}}(x) = f_{2,\mathrm{ALE}}(x) = \frac{\rho}{2}(x^2 - 1)$となる．しかし， 5.3 で見るように，望ましい IML 手法による主効果項は，ρが小さいときにはどちらも$\rho(x^2 - 1)$（もしくは，少なくともそれに近いもの）であるべきだと考えられる．もしそれが正しければ，ALE は主効果を（率としては）大幅に過小評価しているということである．

ALE によるこの「過小評価」の現象が見て取りやすい事例を示す．具体的には，$U_1, \dots, U_d \overset{i.i.d.}{\sim} \mathrm{N}(0,1)$として，

$X_1 = U_1$で，$j = 2, \dots, d$について，$X_j = \rho U_1 + \sqrt{1 - \rho^2} U_j$であり，予測関数が，

$$f(x_1, \dots, x_d) = \frac{1}{d-1}\sum_{j=2}^{d} x_1 x_j$$

であるという事例を考える．この事例では，一方で，$d \to \infty$とすると，$f(X_1, \dots, X_d) \overset{a.s.}{\to} \rho X_1^2$となるので，$d$を大きくすると，$f(x_1, \dots, x_d)$の主効果は，$x_1$以外は消えていく．他方で，$f_{1,\mathrm{ALE}}$は，$d$によらず$\frac{\rho}{2}(x_1^2 - 1)$である．したがって，$d$が大きくなると，両者の乖離が際立ってくる．図 2 は，$\rho = 0.2$，$d = 1000$の場合の図である．横軸がx_1で縦軸が予測値$y = f(x_1, \dots, x_d)$であり，図に示されていない$x_2, \dots, x_d$の値によってyがいろいろな値をとるため，散布図で描いている．

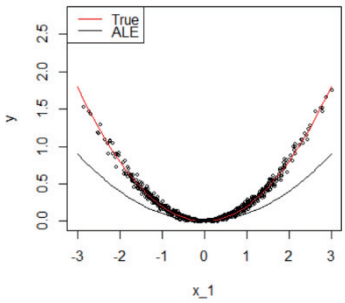

図 2

この例が示すように，ALE は，特徴量変数が多次元正規分布に従っていて，特徴量変数間に単純な相関があるだけの場合にも，ALE がもつ偏りは率としては大きいため，一般にも，ALE が主効果を適切に捉えることができているかは大いに疑問がある．

4.2 代替手法の提案

本項では，PD と ALE に替わる関数分解手法を提案する．それらは，実際には 2 次以上の効果についての分解を与えるようにも定式化できるが，本稿では単純化のために，あとの考察に必要最小限である 1 次の効果への分解のみを定義する．

4.2.1 Conditional Expectation Based Decomposition (CEBD)

本稿では，いくつかの手法をまとめて Conditional Expectation Based Decomposition (CEBD) とよび，ALE などと比較するときは，単に CE とよぶ．

最も単純な CE では，あらかじめ各特徴量変数x_jのとりうる値の中から適当な代表値 (representative value) $\mathrm{Rep}[X_j]$ を 1 つずつ決定しておく．代表値としては，期待値や中央値や最頻値などが考えられる．

CE は，x_jの各値における平均的な効果を測る際，予測関数の値どうしを$x_j = \mathrm{Rep}[X_j]$のときと比べた差を用いる．その「平均的な効果」を求める際に条件付期待値 (conditional expectation) をとるためにこのような名称としている．また，ALE とまったく同様の中心化も施す．ここでは 1 次元の場合のみを考えるので，その中心化は，平均が 0 となるように定数を加減するだけのものである．

理論的には，CE の関数$f_{j,\mathrm{CE}}(x_j)$は，

$$\mathbb{E}\left[f(\mathbf{X}_D) - f\left(\mathrm{Rep}[X_j], \mathbf{X}_{\setminus j}\right) \middle| X_j = x_j\right]$$

に中心化を施したものである．実用上は，これらの関数の各点は，$\mathbf{x}_D$が取りうる値の範囲内の一部の点に対する予測値のデータをもとにした推定によって求められる．その定式は，ALE ほどではないがやや煩雑であり，また，本稿の議論では理論的な面だけとりあげるので，アルゴリズムの詳細は紹介しない．

単純な CE では，各特徴量変数に対して代表値をただ 1 つずつのみ定めることにしていた．しかし，ただ 1 つの代表値とすると特異なことが起こる場合がある．そこで一般には，x_jに対して，ある数$k_j \geq 1$を決めて，代表値$\mathrm{Rep}_1[X_j], \ldots, \mathrm{Rep}_{k_j}[X_j]$を選び，$f\left(\mathrm{Rep}[X_j], \mathbf{X}_{\setminus j}\right)$の代わりに，適当な重み$w_1, \ldots, w_{k_j}$, s.t. $\sum_{k=1}^{k_j} w_k = 1$による

$$\sum_{k=1}^{k_j} w_k f\left(\mathrm{Rep}_k[X_j], \mathbf{X}_{\setminus j}\right)$$

を用いることも考えられる．たとえば，レベル数がk_jのカテゴリー変数であれば，そのすべての値を「代表」とし，実際のデータに各レベルが占める割合を重みとする，といったことも考えられるかもしれない．あるいは，事実上，「代表」は設けず，$f\left(\mathrm{Rep}[X_j], \mathbf{X}_{\setminus j}\right)$の代わりに，すべてのデータに基づいて計算する

$$\frac{1}{n}\sum_{i=1}^{n} f\left(x_{ij}, \mathbf{X}_{\setminus j}\right)$$

を用いることも考えられるかもしれない．

CE がよい方法と考えられる理由は，以下のとおりである．PD の欠点の 1 として指摘される「非現実的なデータに基づいて求められる可能性がある」という点は，PD が，平均を評価するときに条件付期待値を用いていないことによって生じる．そこで，CE では，条件付期待値で平均を評価する．ALE も条件付期待値を用いているが，ALE の場合は，偏微分によって局所的に計測したものを大域的に累積している点に問題が孕んでいる．実のところ，特徴量変数どうしの相関が強い場合には，ALE を大域的に解釈してはならないというのは，原論文 (Apley and Zhu [2016, 2020]) も “the wrong way to interpret the ALE plot” と表現して注意を促していることであり，Molnar [2022] でも指摘されていることである．したがって，予測関数の分解を大域的に意味のあるものとしたいなら，狭い区分内の差分を累積して差分を求めるのでなく，最初から広い区間をまたいだ差分を用いるという代替案が考え

られるのである．

以下で CE というときは，特に断りがない限り，代表値を 1 つだけとる最も単純なもののことであり，代表値は平均値とする．

4.2.2 True CE (TCE)

CE は計算コストなどの点で実用上ではすぐれているかもしれないが，代表値などをとることから ad hoc な印象がある．また，その方法では（PD よりはずっとよいにしても）非現実的な特徴量変数の組み合わせを十分に排除できているという保証はない．

ALE や CE で条件付期待値を考えることは，非現実的な特徴量変数の値の組み合わせを排除することに寄与するはずであった．その点を（計算コストには配慮せずに）徹底すれば，下記のような PD の代替手法が考案でき，それを True CE (TCE) とよぶことにする．

$X_j^{\perp} \sim X_j$，$X_j^{\perp} \perp \mathbf{X}_{\setminus j}$である確率変数$X_j^{\perp}$を考えると，PD による主効果の項は，

$$f_{j,\mathrm{PD}}(x_j) = \mathbb{E}\left[f(x_j, \mathbf{X}_{\setminus j})\right] - \mathbb{E}\left[f(X_j^{\perp}, \mathbf{X}_{\setminus j})\right] = \mathbb{E}\left[f(x_j, \mathbf{X}_{\setminus j}) - f(X_j^{\perp}, \mathbf{X}_{\setminus j})\right]$$

と書ける．TCE は，この条件なしの期待値の代わりに条件付期待値をとる発想に近い．ただし，非現実的な組み合わせを排除するために，実際には

$$\mathbb{E}\left[f(x_j, \mathbf{X}_{\setminus j}) - f(X_j^*, \mathbf{X}_{\setminus j}) \middle| X_j = x_j\right] = \mathbb{E}\left[f(\mathbf{X}_D) - f(X_j^*, \mathbf{X}_{\setminus j}) \middle| X_j = x_j\right]$$

というものを考える．ここで，X_j^*は，$(X_j^*, \mathbf{X}_{\setminus j}) \sim \boldsymbol{X}_D$，$X_j^* | \mathbf{X}_{\setminus j} \perp X_j | \mathbf{X}_{\setminus j}$を満たす確率変数である．そして，これに中心化を施したものを，$f_{j,\mathrm{TCE}}(x_j)$として定義する．

4.2.3 Shapley Interaction Effect Decomposition (SIED)

SHAP の理論値と整合的な関数分解も考えられる．SHAP の理論値によれば，対象iについて$\mathbf{X}_{iD} = \mathbf{x}_D$としたとき，予測値$f(\mathbf{x}_D)$における$x_j$の貢献$\phi_j(x_j)$は次のとおりである．

$$\phi_j(x_j) = \sum_{J \subseteq (\setminus j)} \frac{|J|!\,(d - |J| - 1)!}{d!} \left(\mathbb{E}\left[f(\mathbf{X}_D) | \mathbf{X}_{J \cup \{j\}} = \mathbf{x}_{J \cup \{j\}}\right] - \mathbb{E}\left[f(\mathbf{X}_D) | \mathbf{X}_J = \mathbf{x}_J\right]\right).$$

そこで，$f_{j,\mathrm{SI}}(x_j)$を，母集団全体で見たときの$X_j = x_j$の平均的な貢献だと考えれば，次のとおりとなる．

$$\begin{aligned}
f_{j,\mathrm{SI}}(x_j) &:= \mathbb{E}\left[\phi_j(X_j) | X_j = x_j\right] = \sum_{J \subseteq (\setminus j)} \frac{|J|!\,(d - |J| - 1)!}{d!} \left(\mathbb{E}\left[\mathbb{E}\left[f(\mathbf{X}_D) | \mathbf{X}_{J \cup \{j\}}\right] | X_j = x_j\right] - \mathbb{E}\left[\mathbb{E}\left[f(\mathbf{X}_D) | \mathbf{X}_J\right] | X_j = x_j\right]\right) \\
&= \sum_{J \subseteq (\setminus j)} \frac{|J|!\,(d - |J| - 1)!}{d!} \mathbb{E}\left[f(\mathbf{X}_D) | X_j = x_j\right] - \sum_{J \subseteq (\setminus j)} \frac{|J|!\,(d - |J| - 1)!}{d!} \mathbb{E}\left[f(\mathbf{X}_{\setminus J}^*, \mathbf{X}_J) | X_j = x_j\right] \\
&= \mathbb{E}\left[f(\mathbf{X}_D) | X_j = x_j\right] - \sum_{J \subseteq (\setminus j)} \frac{|J|!\,(d - |J| - 1)!}{d!} \mathbb{E}\left[f(\mathbf{X}_{\setminus J}^*, \mathbf{X}_J) | X_j = x_j\right].
\end{aligned}$$

ここで，$\mathbf{X}_{\setminus J}^*$は，$(\mathbf{X}_{\setminus J}^*, \mathbf{X}_J) \sim \boldsymbol{X}_D$，$\mathbf{X}_{\setminus J}^* | \mathbf{X}_J \perp \mathbf{X}_{\setminus J} | \mathbf{X}_J$であるような確率変数ベクトルである．

このように定義される主効果項の集合$\{f_{j,\mathrm{SI}}\}_{j \in D}$を与える分解を，「Shapley 交互作用効果分解（Shapley Interaction Effect Decomposition）」とよび，SIED もしくは SI と略記する．実データに対して，これらの主効果項の関数値を推定するためには計算コストがかかるが，単純なモデルの場合には解析的に決定できる場合もあり，理論的考察に役立つ．

4.2.4 Minimum Interaction Effect Decomposition (MIED)

交互作用効果量を最小化する主効果項の集合$\{f_{j,\mathrm{MI}}\}_{j\in D}$を与える分解を，「最小交互作用効果分解（Minimum Interaction Effect Decomposition)」とよび，MIED もしくは MI と略記する．ただし，規約として，$\mathbb{E}[f_j(X_j)]=0$とする．そのような規約をしたとしても，一般には，条件を満たす主効果項の集合が一意に決まるとは限らないが，単純なモデルの場合には解析的に決定できる場合もあり，理論的考察に役立つ．

5 PD，ALE と代替手法の比較

5.1 見た目は主効果項がないが主効果が大きい例

3.1 で，線形モデルにおいて，式の見た目には 2 次の効果の項しかないが，実際には大きな主効果がある例に言及した．その事例は，実は，

$$U,V\overset{i.i.d.}{\sim}\mathrm{N}(0,1),\ X_1=U,\ X_2=U^2-V^2,\ f(x_1,x_2)=x_1x_2.$$

というものであった．この事例に対する主効果を PD，ALE，CE，TCE，SI，MI のそれぞれについて計算すると，どの手法でも$f_2(x_2)=0$である一方で，

$$f_{1,\mathrm{PD}}(x_1)=0,\qquad f_{1,\mathrm{ALE}}(x_1)=\frac{x_1^3}{3}-x_1,$$

$$f_{1,\mathrm{CE}}(x_1)=f_{1,\mathrm{TCE}}(x_1)=f_{1,\mathrm{SI}}(x_1)=f_{1,\mathrm{MI}}(x_1)=x_1^3-x_1$$

となる．図 3 は，横軸がx_1で縦軸が予測値$y=f(x_1,x_2)$である．図に示されていないx_2の値によってyがいろいろな値をとるため，散布図で描いている．また，各手法に基づくf_1の曲線も描いている．この図から，この事例では，PD だけでなく，ALE も本来の主効果が捉えられていないことが示唆される．

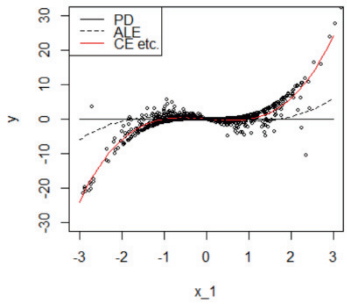

図 3

各手法の計算過程は特に示さないが，MI（最小交互作用効果分解）のf_1の計算だけ確かめておく．交互作用効果量

$$\mathrm{IEQ}:=\mathbb{E}\left[\left(f(\mathbf{X}_D)-f_\emptyset-\sum_{j\in D}f_j(X_j)\right)^2\right]=\mathbb{E}[(X_1X_2-f_\emptyset-f_1(X_1)-f_2(X_2))^2]$$

を最小化するf_1は，$f_1(x_1)=\mathbb{E}[f(X_1,X_2)-f_\emptyset-f_2(X_2)|X_1=x_1]$であることと，対称性などからどのような関数分解でも$f_2(x_2)=0$が成り立つはずであることから，MI のf_1は，

$$f_1(x_1)=\mathbb{E}[f(X_1,X_2)-f_\emptyset-f_2(X_2)|X_1=x_1]=\mathbb{E}[X_1X_2|X_1=x_1]=x_1^3-x_1$$

となる．

この事例に対する交互作用効果量 IEQ を各手法について求めると，

$$\mathrm{IEQ}_{\mathrm{PD}}=12,\qquad \mathrm{IEQ}_{\mathrm{ALE}}=\frac{26}{3}\fallingdotseq 8.7,\qquad \mathrm{IEQ}_{\mathrm{CE}}=\mathrm{IEQ}_{\mathrm{TCE}}=\mathrm{IEQ}_{\mathrm{SI}}=\mathrm{IEQ}_{\mathrm{MI}}=2$$

となる．したがって，PD と ALE に比べて， CE ほかの値は圧倒的に小さい．

交互作用効果量が小さいということは，主効果だけで予測関数をよく説明できているということであり，主効果で説明できていない割合を未解釈割合 UR と見なして計算すれば，

$$\mathrm{UR}_{\mathrm{PD}}=100\%,\qquad \mathrm{UR}_{\mathrm{ALE}}=\frac{13}{18}\fallingdotseq 72\%,\qquad \mathrm{UR}_{\mathrm{CE}}=\mathrm{UR}_{\mathrm{TCE}}=\mathrm{UR}_{\mathrm{MI}}=\mathrm{UR}_{\mathrm{SI}}=\frac{1}{6}\fallingdotseq 17\%$$

となる．

したがって，CE ほかが捉えているものをこの事例における主効果と考えるとよい．また，実用上の観点からいうと，主効果のみを考慮したときの未解釈割合が小さい手法ほど便利だと思われる．そのことを強くとると，定義上，MI (最小交互作用効果分解) によって捉えられた主効果が「正解」ということになるが，少なくとも現時点ではそこまでは主張しない．しかし，交互作用効果量が MI に対するものよりもずっと大きくなっている手法はうまくいっていないと断じることは可能であろう．

5.2 交互作用効果の可視化・計量化

5.1 で見た事例から，手法の選択を誤ると交互作用効果の可視化・計量化が不適切になることが示唆された．そのことを具体的に見るため，5.1.の例から主効果（と思われる要素）を除いた

$$U, V \overset{i.i.d.}{\sim} \mathrm{N}(0,1),\ X_1 = U,\ X_2 = U^2 - V^2,\ f(x_1, x_2) = x_1 - x_1^3 + x_1 x_2.$$

という事例を考えてみる．この事例を図示すると図 4，5，6 のとおりである．

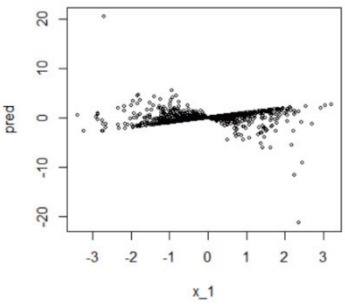

図 4

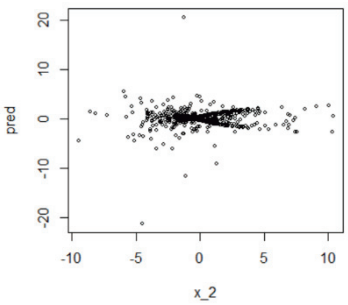

図 5

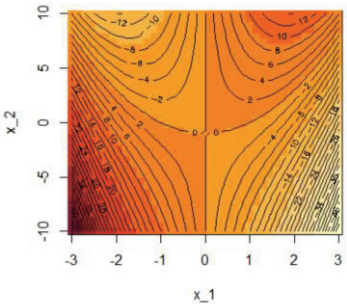

図 6

図 4 は横軸がx_1で縦軸が予測値yであり，図に示されていないx_2の値によってyがいろいろな値をとるため，散布図で描いている．図 5 は横軸がx_2で縦軸が予測値yであり，散布図としている理由は図 4 と同様である．図 6 は横軸がx_1で縦軸がx_2でありyの値の違いを色で表している（値の大きさにはついては等高線を参照)．どの図を見ても，何らかの強い交互作用効果の存在が示唆される一方で，特段の主効果はないように見える．

この予測関数に PD，ALE，CE ほかをあてはめてみると，どの手法でもx_2の主効果は$f_2(x_2)$ =0 となる一方で，x_1の主効果は次のとおりとなる．

$$f_{1,\mathrm{PD}}(x_1) = x_1 - x_1^3, \qquad f_{1,\mathrm{ALE}}(x_1) = -\frac{2x_1^3}{3}, \qquad f_{1,\mathrm{CE}}(x_1) = f_{1,\mathrm{TCE}}(x_1) = f_{1,\mathrm{SI}}(x_1) = f_{1,\mathrm{MI}}(x_1) = 0$$

すなわち，CE ほかではf_1は 0 となるが，PD と ALE では予測関数にはっきりした主効果があると計算される．また，この事例は，5.1 で見た事例に対して主効果を変更したものにすぎないので，交互作用効果量の PD，ALE，CE ほかに対する値はそれぞれ12, 8.7, 2であって先の事例と同じであり，CE ほかの値が，本項で示した関数分解手法のなかでは最小である．

この事例の予測関数は 2 変数関数であり，$f_\emptyset = 0$である．各手法のf_1とf_2はすでに示したとおりである．したがって，

$$f_{\{1,2\}}(x_1, x_2) = f(x_1, x_2) - f_\emptyset - f_1(x_1) - f_2(x_2)$$

とすることで，2 次の交互作用効果項も求められる．PD と ALE は，

$$f_{\{1,2\},\mathrm{PD}}(x_1,x_2)=x_1x_2,\qquad f_{\{1,2\},\mathrm{ALE}}(x_1,x_2)=x_1-\frac{x_1^3}{3}+x_1x_2$$

となる．CE ほかの$f_{\{1,2\}}$は，予測関数そのものと等しく，次のとおりとなる．

$$f_{\{1,2\}}(x_1,x_2)=f(x_1,x_2)=x_1-x_1^3+x_1x_2$$

したがって，この事例については，CE ほかによる分解が正しいとすると，主効果はなく，2 次の効果は，予測関数そのものである．それに対し，この予測関数を PD, ALE それぞれによって分解した結果に基づいて，x_1の主効果と 2 次の効果を可視化すると図 7 のようになり，CE ほかと大きく異なっていることが目に見える．2 次の効果の方は違いがわかりにくいかもしれないが，0 の等高線の形を見れば，はっきりした違いが見て取れる．

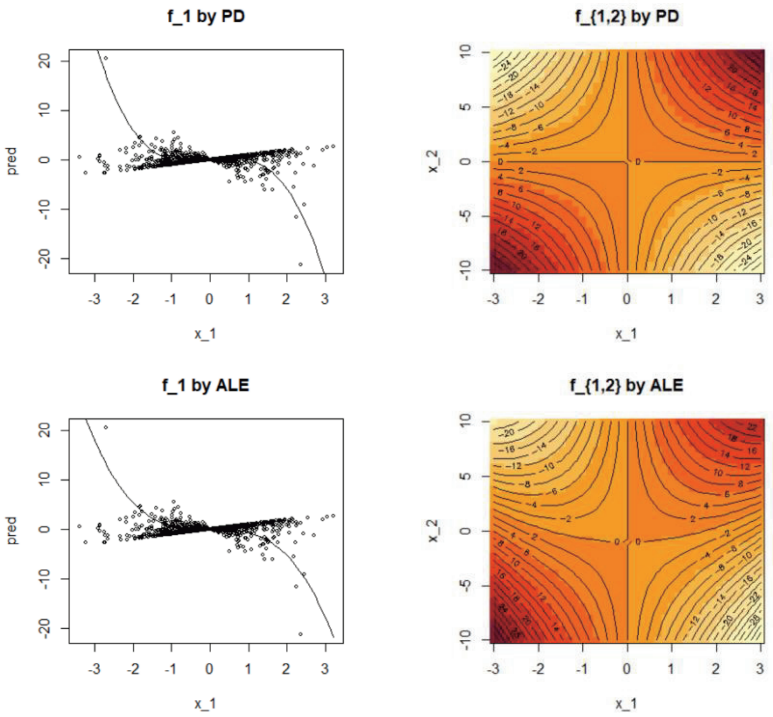

図 7

計量化した場合の違いは，より明らかである．この事例の予測関数の変数は 2 つだけだったので，変数x_1とx_2に対する（交互作用効果）重要度$\mathrm{IEI}=\mathbb{E}\left[f_{\{1,2\}}(X_1,X_2)^2\right]$は，モデル全体の交互作用効果量と一致し，

$$\mathrm{IEI}_{\{1,2\},\mathrm{PD}}=12,\qquad \mathrm{IEI}_{\{1,2\},\mathrm{ALE}}\fallingdotseq 8.7,\qquad \mathrm{IEI}_{\{1,2\},\mathrm{CE}}=\mathrm{IEI}_{\{1,2\},\mathrm{TCE}}=\mathrm{IEI}_{\{1,2\},\mathrm{MI}}=\mathrm{IEI}_{\{1,2\},\mathrm{SI}}=2$$

となり，重要度の評価が大きく異なってしまう．

5.3 単純な相関の影響

単純な線形予測子においては，見た目の「2 次の交互作用効果」項は，たとえばx_1x_2というように 2 つの変数の単なる積で表され，product term とよばれる．本稿のアプローチに基づけば，一般に product term は主効果を含んでいる．本項では，product term に含まれている主効果はどのようなものであるかを調べる．前項では，特徴量変数どうしに特殊な従属性がある場合を取り上げたが，本項では単純な相関がある場合を扱う．

具体的には，$\begin{pmatrix}X_1\\X_2\end{pmatrix}\sim \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}\right)$であり，予測関数が次の場合をとりあげる．

$$f(x_1,x_2)=x_1x_2$$

このとき，$f_\emptyset=\rho$であり，どの関数分解手法においても $f_1=f_2$である．f_1とf_2は次のとおりとなる．

$$f_{1,\mathrm{PD}}(x) = f_{2,\mathrm{PD}}(x) = 0, \qquad f_{1,\mathrm{ALE}}(x) = f_{2,\mathrm{ALE}}(x) = \frac{\rho}{2}(x^2-1), \qquad f_{1,\mathrm{CE}}(x) = f_{2,\mathrm{CE}}(x) = \rho(x^2-1),$$

$$f_{1,\mathrm{TCE}}(x) = f_{2,\mathrm{TCE}}(x) = (\rho-\rho^3)(x^2-1), \qquad f_{1,\mathrm{SI}}(x) = f_{2,\mathrm{SI}}(x) = \left(\rho - \frac{1}{2}\rho^3\right)(x^2-1),$$

$$f_{1,\mathrm{MI}}(x) = f_{2,\mathrm{MI}}(x) = \frac{\rho}{1+\rho^2}(x^2-1)\ .$$

ここで，$\frac{\rho}{1+\rho^2} = \rho - \rho^3 + \rho^5 - \cdots$であることからすると，CE と TCE と SI は，MI の近似になっていると見ることができる．

交互作用効果量 IEQ はそれぞれ

$$\mathrm{IEQ}_{\mathrm{PD}} = 1+\rho^2, \qquad \mathrm{IEQ}_{\mathrm{ALE}} = 1-2\rho^2+\rho^4, \qquad \mathrm{IEQ}_{\mathrm{CE}} = 1-3\rho^2+4\rho^4,$$

$$\mathrm{IEQ}_{\mathrm{TCE}} = 1-3\rho^2+4\rho^4-4\rho^6+4\rho^8, \qquad \mathrm{IEQ}_{\mathrm{SI}} = 1-3\rho^2+4\rho^4-3\rho^6+\rho^8,$$

$$\mathrm{IEQ}_{\mathrm{MI}} = \frac{1-2\rho^2+\rho^4}{1+\rho^2} = 1-3\rho^2+4\rho^4-4\rho^6+4\rho^8-\cdots$$

である．

上記のいずれの$f_1(x)$も$f_2(x)$も，ρに応じて決まる何らかのαによって$\alpha(x^2-1)$と書けるので，そのαを比較する図を示せば図 8 のとおりである．また，それぞれの交互作用効果量 IEQ および未解釈割合 UR はそれぞれ図 9 および図 10 のとおりである．

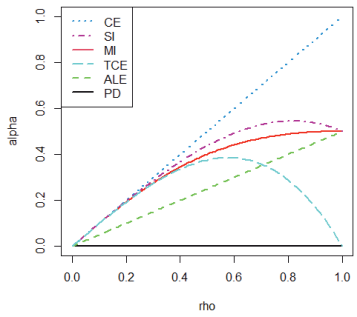

図 8

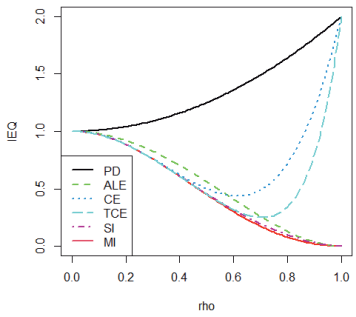

図 9

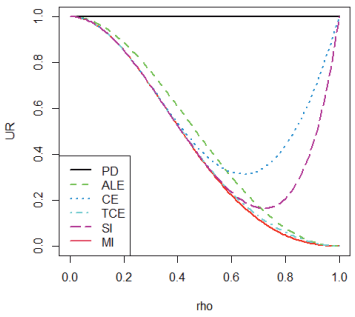

図 10.

図 8 から見て取れるように，MI を基準とすれば，f_1やf_2を，PD と ALE と TCE はつねに過小評価しており，CE と SI はつねに過大評価している．ただし，SI は一貫して MI にかなり近い．特に，交互作用効果量や未解釈割合で測れば，図 9，10 では一貫して MI の値とほぼ一致しており，少なくともこの事例においては，SI は原理的には理想にかなり近い手法であることが示唆される．

どの図を見ても，PD は他と比べて MI からの乖離が大きいが，ALE は，ρが小さいときは，TCE や SI や CE と比べると MI の値からかなり離れているものの，ρが大きくなるに従って MI の値に近づいていき，$\rho = 1$のときには一致する．

CE を，交互作用効果量や未解釈割合で測ると，ρが小さいときは，ALE よりもかなり MI の値に近いが，$\rho = 0.4$を超えたあたりから目に見えて MI の値との差が大きくなりはじめ，$\rho = \frac{1}{\sqrt{3}} \fallingdotseq 0.58$を超えると，ALE よりも MI の値から遠くなり，その後，急激に離れていき，$\rho = 1$のときには PD の値と一致してしまう．

TCE を，交互作用効果量や未解釈割合で測ると，ρが小さいときは，SI よりもさらに MI に近いが，$\rho = 0.6$を超えたあたりから目に見えて MI の値との差が大きくなりはじめ，特に$\rho = 0.8$を超えたあたりから急激に離れていき，

$\rho = 1$のときにはPDの値と一致してしまう．

5.4 多重共線性などがある場合

多重共線性がある場合などは特徴量変数の値がほかの特徴量変数の値によって決まってしまうが，それに近い状況のときには，どういうことがいえるであろうか．前項の事例でρが大きい場合の結果からすると，そのような状況では，たとえば，線形モデルのproduct termに類似した要素に対する関数分解手法として，CEは不適切だがALEは使用可能であるといえるであろうか．たしかに，その場合にCEを使うのは不適切であるというのは正しいが，ALEは使用可能であるというのは論点を外している．以下で論じるように，そこでいうべきことは，「そのような状況になっているモデルは，解釈可能な予測モデルとして使うこと自体が不適切である」ということである．

そのような状況の場合，ほかの特徴量変数で表すことができる変数があるならば，変数を減らしたモデルに作り直すべきである．少なくとも解釈という点からすると，そもそもモデルに多重共線性はあってはならない．したがって，解釈をはじめる前に特徴量変数の集合が十分に絞られているか，あるいは，適切な特徴量加工などによって，十分に質の高い特徴量変数の集合に変換されているかを確認すべきである．

より理論的に，IMLの観点からコメントを加えれば，以下のとおりである．2つの予測関数があって，現実のデータについて，同じ入力値に対する出力値が2つの関数どうしでつねに一致しているときは，その2つは同じ予測関数である．より正確にいえば，$\boldsymbol{x}_D$のとりうる（ベクトル）値の範囲を$\mathcal{X} \subseteq \mathbb{R}^d$と書くとして，2つの予測関数$f_A$と$f_B$に関して，任意の$\boldsymbol{x}_D \in \mathcal{X}$について$f_A(\boldsymbol{x}_D) = f_B(\boldsymbol{x}_D)$ならば，その2つの関数は「$\mathcal{X}$について同一のブラックボックス関数」であるといい，「$f_A \equiv_{\mathcal{X}} f_B$」と書くことにする．

たとえば，

$$\mathcal{X} = \{(x, x) | x \in \mathbb{R}\}, \qquad U \sim \mathrm{N}(0,1), \qquad X_1 = X_2 = U, \qquad f_A(x_1, x_2) = x_1 x_2, \qquad f_B(x_1, x_2) = x_1^2$$

とすると，$f_A \equiv_{\mathcal{X}} f_B$である．このとき，$f_A$に対しては

$$f_{1,\mathrm{PD}}(x) = f_{2,\mathrm{PD}}(x) = 0, \qquad f_{1,\mathrm{ALE}}(x) = f_{2,\mathrm{ALE}}(x) = \frac{1}{2}(x^2 - 1), \qquad f_{1,\mathrm{CE}}(x) = f_{2,\mathrm{CE}}(x) = x^2 - 1$$

であるが，f_Bに対しては，

$$f_{1,\mathrm{PD}}(x) = x^2, \qquad f_{1,\mathrm{ALE}}(x) = x^2 - 1, \quad f_{1,\mathrm{CE}}(x) = x^2 - 1, \qquad f_{2,\mathrm{PD}}(x) = f_{2,\mathrm{ALE}}(x) = f_{2,\mathrm{CE}}(x) = 0$$

となり，分解がまったく異なる．しかし，この2つの予測関数はブラックボックス関数としては同一なのだから，IMLの趣旨からすれば，異なる分解が得られるのは不適切である．つまり，ここでは解釈はうまくいっていないと理解すべきだと思われる．

一般に，一部の特徴量変数の値がほかの特徴量変数の値によって決まってしまう場合（もしくは多重共線性が疑われる場合）には，同様のことが起こる．すなわち，ブラックボックス関数としてはほぼ同一ながら異なる分解が与えられてしまう複数の予測関数が必ず存在する．したがって，そのような状況のときは，解釈へと進むのではなく，予測モデルを再構築すべきである．

6 おわりに

本稿における分析で，交互作用効果を把握するのに使われる手法であるPDとALEはどちらも主効果をうまく捉えることができないことを明らかにした．このことから，交互作用効果もうまく捉えられているとは考えられない．本稿では，交互作用効果を定義したうえで，その諸性質を確認し交互作用効果を計量化するための指標として，交互

作用効果割合，交互作用効果量，未解釈割合，交互作用効果重要度を提案した．そして，PD と ALE に替わる IML 手法として CE，TCE，SI を提案し，それらの指標を用いて既存の手法および代替手法の特徴を明らかにした．加えて，主効果を計量する際に基準となる理論的な主効果項として MI という概念を提案した．そして，既存の手法と代替手法（理論的な比較対象として提案した MI を含む）を，単純なモデルに適用して比較分析した．これにより，既存の手法の問題点と代替手法の可能性を示すことができたと思料する．これらの分析・検討をとおして，交互作用効果の可視化・計量化のための具体的な技術に関して一定の知見を示すことで本稿の目的を満たすことができたと考える．

その一方で，本稿で示した比較だけでは，より一般的な場合におけるこれらの手法の違いが十分に解明できたわけではない．特に，どのような IML 手法が望ましいか，または各 IML 手法の優越性はどのような条件の下で示されるかについては，より詳しい基礎研究が必要だと思われる，そうした研究により，さらに新たな手法が考案できる可能性もある．また，提案した代替手法のいずれについても，一般的な予測モデルに適用するための具体的な方法および計算負荷までは提示していない．それらの手法を本格的な実用に向けて提案するのだとしたら，実際的なアルゴリズムの提案や，具体的な実装による数値実験，次元を大きくした場合の計算量の増大などに取り組む必要があると考える．

今後，機械学習モデルをはじめとするブラックボックスモデルが多くの場面で広く活用されていくと想定されるなか，アクチュアリーにとってモデルの解釈しやすいほうがよいことは論ずるまでもない．予測モデリングでは，IML は主に 2 つの利用方法がある．1 つは，採用したモデルを解釈することであり，もう 1 つは，モデリングの過程で予測精度が高い機械学習モデルを解釈することによって，より解釈しやすいモデル（典型的には，GLM や GAM）を構築するための参考にすることである．どちらの場合も，解釈しやすいほうがよいのは当然である．特に後者では，解釈しやすい，すなわち主効果項に基づく解釈の占める割合が高いほど，モデルを構築しやすい．本稿の検討結果を踏まえると，筆者は交互作用効果を計量化する前に主効果項のみを使ってできるだけ多くの部分の解釈を行うことを提案する．主効果項のほうが理解しやすいので，主効果項に基づく解釈の占める割合が高い IML のほうが，そうでない IML よりも解釈しやすい．そして，構築したモデルも解釈しやすいほど，適切なモデルであるかどうかの検証もしやすい．さらには，アクチュアリー業務としての実用に際しては，IML 手法を使うことによって機械学習モデルを参考にして，たとえば分類問題に対してロジスティック回帰モデル（GLM の一種）を構築するというアプローチも考えられる．その場合には，目的変数の代わりに目的変数にロジット変換を施したものを用いて IML 手法を適用することが考えられる．逆に，交互作用項に基づく解釈の占める割合が高く解釈しにくいと，適切なモデル構築は困難となり，完成したモデルは不適切なものになる，または不適切であることを解釈できない可能性がある．

機械学習モデルは高い予測力を示す一方で，ブラックボックスモデルになりがちで，その内容を解釈することは困難である．元々機械学習モデルのニーズは予測精度の高さにあり，研究の競争は予測精度の高いモデルを提供することに割かれてきた．そして，我々アクチュアリーは説明力の低さゆえ，機械学習モデルを広くその業務に活用することなく過ごしてきた．筆者は，アクチュアリーがこのような制約により有用なモデルを活用しない現状を解消すべきであると主張する．我々アクチュアリーに機械学習モデルを解釈可能なものとするための技術に対するニーズがあるのなら，アクチュアリー自身がそのようなニーズを満たす技術を開発するべきだと思料する．理論的な観点でも実用上の観点でも，本研究は端緒についたばかりであり，取り組むべき課題は多い．それでも，本稿によって得られた基礎的な結果のみでも，アクチュアリーが機械学習モデルを利用するために必要なことを議論するきっかけには十分になると思われる．これにより，アクチュアリーが自らの責務である説明責任を果たすため，活発な議論が行われることを期待する．加えて，これらの一連の研究に関しては，研究者，実務者にとどまらず，アクチュアリー組織，

できれば監督当局も議論に加わってくれることを期待する．

Appendix A.1　2 変数関数に対する中心化

着目する関数分解手法（たとえば ALE）によって，fからf_jを$f_j \coloneqq \mathcal{H}_j(f)$というように決定する線形演算子$\mathcal{H}_j$を考える．$I$を恒等演算子とし，$\mathbb{E}$を，任意の関数$g(\mathbf{x}_D)$について，その関数に応じて，$\mathbb{E}(g) \coloneqq \mathbb{E}[g(\mathbf{X}_D)]$となるものとし，演算子の合成を$\circ$で表す．すると，たとえばALE の場合は，$\mathcal{L}_j(f)(x_j) = \int_{\min\{x_j\}}^{x_j} \mathbb{E}\left[f^j\left(X_j, \mathbf{X}_{\setminus j}\right) \middle| X_j = z_j\right] dz_j$という演算子$\mathcal{L}_j$を考えると，$\mathcal{H}_j = (I - \mathbb{E}) \circ \mathcal{L}_j$と書ける．一般にも，ここで考える$f_j = \mathcal{H}_j(f)$は中心化されている（平均すると 0 となるようになっている）ので，何らかの$\mathcal{L}_j$を使って$\mathcal{H}_j = (I - \mathbb{E}) \circ \mathcal{L}_j$と書ける．このとき，2 変数関数$g(x_j, x_l)$に対して，$I - \mathbb{E} - \mathcal{H}_j - \mathcal{H}_l$を施すことを 2 変数関数$g(x_j, x_l)$に対する中心化という．

参考文献

Adadi, A., and Berrada, M. [2018], “Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)”, *IEEE access,* Vol. 6, pp. 52138-52160.

Apley, D. W., and Zhu, J. [2016], “Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models”, *arXiv preprint arXiv:1612.08468*.

Apley, D. W., and Zhu, J. [2020], “Visualizing the effects of predictor variables in black box supervised learning models”, *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* Vol. 82, No. 4, pp. 1059-1086.

Baeder, L., Brinkmann, P., andandand Xu, E. [2021], *Interpretable Machine Learning for Insurance*, Society of Actuaries.

Biecek, P., and Burzykowski, T. [2021], *Explanatory model analysis: Explore, explain and examine predictive models*, Chapman and Hall/CRC.

Friedman, J. H. [2001], “Greedy function approximation: a gradient boosting machine”, *Annals of statistics*, pp. 1189-1232.

Friedman, J. H., and Popescu, B. E. [2008], “Predictive learning via rule ensembles”, *The annals of applied statistics*, pp. 916-954.

Inglis, A., Parnell, A., and Hurley, C. B. [2022], “Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models”, *Journal of Computational and Graphical Statistics*, pp. 1-13.

Jaccard, J., and Turrisi, R. [2003], *Interaction effects in multiple regression*, Sage.

Janzing, D. M. [2020], “Feature relevance quantification in explainable AI: A causal problem” in *International Conference on artificial intelligence and statistics* (pp. 2907-2916). PMLR.

Lorentzen, C., and Mayer, M. [2020], “Peeking into the black box: An actuarial case study for interpretable machine learning” *available at SSRN 3595944*.

Lundberg, S. M., and Lee, S. I. [2017], “A unified approach to interpreting model predictions”, *Advances in neural information processing systems,* Vol. 30, pp. 4768-4777.

Lundberg, S. M., Erion, G. G., and Lee, S. I. [2018], “Consistent individualized feature attribution for tree ensembles”, *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal,

N., and Lee, S. I. [2020], “From local explanations to global understanding with explainable AI for trees”, *Nature machine intelligence,* Vol. 2, No. 1, pp. 56-67.

Molnar, C. [2022], *Interpretable machine learning, 2nd Ed.* Lulu. com. Continuously updated online version is available at https://christophm.github.io/interpretable-ml-book/.

Measuring and Visualizing Two-way Interaction Effects

Hirokazu Iwasawa　　Tadayoshi Otsuka

Abstract

Machine learning models show high predictive power but lack interpretability when making predictions. Currently, research on establishing a method for making machine learning models interpretable, is being conducted as the field of Interpretable Machine Learning (IML). However, the current methods are considered inadequate for actuaries to utilize machine learning models in their work. On the other hand, not using models with highly predictive accuracy will hinder the development of actuarial work in the future. The purpose of this paper is to propose methods and indicators necessary to raise the interpretability of machine learning models up to the level that can be utilized in actuarial works.

Firstly, we define the interaction effect and propose the Interaction Effect Ratio, Interaction Effect Quantity, Uninterpreted Ratio, and Interaction Effect Importance as indicators to confirm its various properties and to quantify the interaction effect. As the next step, we clarify that existing methods PD and ALE do not capture either main effects or interaction effects well. Consequently, we propose CE, TCE and SI as alternative IML methods, and clarify the characteristics of existing methods and alternative methods using proposed indicators. In addition, by applying existing methods and alternative methods to a simple model and conducting a comparative analysis, we showed the problems of existing methods and the possibilities of alternative methods. Finaly, we propose that, in order actuaries to use machine learning models in their work, methods that allows actuaries to interpret as many parts as possible using only the main effect term.