

金融研究

2026 年 7 月 / 第 45 卷 第 3 号

ディープフェイク検知モデルの評価・比較：

研究事例を活用する際の留意点と課題……………宇根正志

SHAP の代替的手法の検討：

協力ゲーム理論を用いたアプローチ……平木一浩／石原慎一／篠潤之介

日本銀行金融研究所

日本銀行 金融研究所

Institute for Monetary and Economic Studies (IMES)

Bank of Japan

特別顧問

星 岳雄 東京大学教授

国内顧問

大橋 和彦 一橋大学・東京科学大学教授

粕谷 誠 東京大学教授

神田 秀樹 東京大学名誉教授

塩路 悦朗 中央大学教授

榆井 誠 東京大学教授

海外顧問

Markus K. Brunnermeier

Edwards S. Stanford Professor of Economics, Princeton University

Athanasios Orphanides

Professor of the Practice, Global Economics and Management, Massachusetts Institute of Technology

所 長

渡辺 真吾

本誌に掲載されている論文等の内容や意見は、執筆者個人に属し、日本銀行あるいは金融研究所の公式見解を示すものではありません。

金融研究

2026 年 7 月 / 第 45 卷 第 3 号

目次

- | | | |
|----|---|----------------|
| 1 | ディープフェイク検知モデルの評価・比較：
研究事例を活用する際の留意点と課題 | 宇根正志 |
| 27 | SHAP の代替的手法の検討：
協力ゲーム理論を用いたアプローチ | 平木一浩／石原慎一／篠潤之介 |
| 81 | 金融研究所の概要等 | |

ディープフェイク検知モデルの 評価・比較： 研究事例を活用する際の 留意点と課題

うね まさし
宇根正志

要 旨

スマートフォンを用いた金融取引では、顧客の本人確認手段として、顔の動画や静止画による生体認証が採用されるケースがある。こうした生体認証への脅威として、偽の動画を提示してなりすましを試みる攻撃が想定される。最近では、AI・機械学習によるディープフェイクを用いた攻撃の可能性を示唆する研究成果が複数発表され、現実的な脅威として認識する必要性が高まっている。対策としては、機械学習モデルによってディープフェイクを検知する手法の研究が活発化しており、複数の検知モデルを横並びで評価・比較した研究成果も発表されはじめている。こうした動向を踏まえると、今後、生体認証を採用している金融機関は、ディープフェイクによるリスクを評価し、必要に応じてディープフェイクの検知モデルの採用を検討することが必要となるであろう。その際には、想定されるディープフェイクや検知モデルの評価基準をまず設定し、評価・比較の研究事例を参照しつつ評価基準と合致したものを選択することが望ましい。また、ディープフェイクに関する技術の進歩が非常に速く、対策の有効性やリスクが時間とともに変化するため、技術動向のフォローやリスクの再評価を継続的行うことも重要である。

キーワード： 機械学習、生体認証、セキュリティ、ディープフェイク、リスク管理、AI

.....
本稿の作成に当たっては、市野将嗣准教授（電気通信大学）から有益なコメントを頂戴した。ここに記して感謝したい。ただし、本稿に示されている意見は、筆者個人に属し、日本銀行の公式見解を示すものではない。また、ありうべき誤りはすべて筆者個人に属する。

宇根正志 日本銀行金融研究所参事役（E-mail: masashi.une@boj.or.jp）

1. はじめに

近年、AI や機械学習の研究が急速に進展し、実際には行われていない人間の動作をあたかも行われていたかのようにみせる動画や音声を容易に生成できるようになっている。こうした技術や生成されたメディアは、深層学習を含む機械学習に基づいて開発されてきたことから、ディープフェイクと呼ばれている。例えば、自分の髪型や洋服を選ぶ際に、自分の動画や静止画に特定の髪型や洋服を合成してフィット感を確かめるといったユースケースがよく知られている。また、与えられたテキストを外国語に翻訳し、それに対応する音声を生成する自動翻訳も代表例の1つである。最近では、テキストから画像や音声を生成する生成 AI によるディープフェイクも注目を集めている（坂本・宇根 [2025]；笹原 [2023]；Gaur 2023）。

情報セキュリティの観点では、ディープフェイクは、金融サービスにおいてなりすましに悪用される可能性がある。例えば、モバイル・バンキングの使用時にスマートフォン内蔵のカメラで撮影した顔の動画によって本人確認（生体認証）を行う場合、攻撃者が他人の顔の動画を生成し、それをカメラに提示するなどしてなりすましを試みる攻撃が想定される。こうした攻撃は、金融機関の顧客の個人情報盗まれたり不正な取引が実行されたりするリスクにつながる。金融庁は、AI ディスカッションペーパー（金融庁 [2025]）のなかで、特定の人物になりすました偽の動画や音声による詐欺行為などがディープフェイクによって容易になる可能性を指摘するとともに、新たなリスクとして留意することが重要であるとしている¹。

生体認証の文脈では、人工物をセンサーに提示してなりすましを試みる攻撃シナリオが既によく知られており、さまざまな対策手段が研究されてきた（宇根 [2016; 2024]；Khan and Khan 2025）。例えば、顔の動画を用いた認証の場合、カメラで取得した動画を分析し、血流による肌の色合いの変化（Li *et al.* 2016）、顔の表面の微細な凹凸の有無（Xu, Li and Deng 2015）、目や鼻などの特徴点の三次元構造の有無（Wang *et al.* 2019）などを手掛りに、対象が人間か否かを判定する対策手段が挙げられる。

こうしたアプローチに加え、顔の動画が与えられたときにそれがディープフェイクか否かを機械学習の技術によって判定するアルゴリズム（ディープフェイク検知

.....
1 海外のセキュリティ当局もディープフェイクの脅威に注意を促している。アメリカの国家安全保障局・連邦捜査局・サイバーセキュリティ社会基盤安全保障庁は、ディープフェイクによって、重要人物へのなりすまし、重要な情報への不正アクセスなどが発生しうるとしたうえで、対策の必要性を指摘している（National Security Agency, Federal Bureau of Investigation, and Cybersecurity and Infrastructure Security Agency 2023）。また、アメリカの金融犯罪取締ネットワークは、金融機関の顧客の動画などを生成 AI によって偽造し、他人になりすまして口座開設を試みたりフィッシングや詐欺を実行したりするおそれがあるとの注意喚起を行っている（Financial Crimes Enforcement Network 2024）。

手法)の研究が近年活発化している。ディープフェイク検知手法に訓練用データを適用して検知モデルを生成し、それを生体認証のシステムに組み込んで活用することが考えられる。最近では、大規模言語モデル(LLM: Large Language Model)を活用する手法も相次いで提案されている(Deng *et al.* 2025 ; Lin *et al.* 2025 ; Sony *et al.* 2025)。

顔の動画や音声による生体認証を顧客の本人確認の手段として採用している金融機関は、ディープフェイクによるなりすましのリスクを適切に評価し、リスクが許容レベルを超えていた場合にはリスク低減策を検討する必要がある。リスク評価や対策手段の検討を適切に行うためには、ディープフェイクに関する技術・実装の動向をタイムリーにフォローすることが重要である。

本稿では、顔の動画のディープフェイクに焦点を当てて、検知モデルを評価・比較した最近の研究事例を紹介し、検知モデルを選択するために研究事例を活用する際の留意点や課題を考察する。2節では、顔の動画のディープフェイクの生成手法や、ディープフェイクによるなりすましの可能性を検証する実験の事例を紹介する。3節では、検知モデルを評価・比較した最近の研究事例を紹介する。4節では、研究事例を活用する際の準備や留意点を考察するとともに、研究コミュニティにおける課題についても考察する。

2. 顔の動画を対象とするディープフェイクと実験研究

本節では、人間の顔の動画を対象とするディープフェイクの主な生成手法と、生成されたディープフェイクによるなりすましの可能性を示唆する実験研究を紹介する。

(1) 主な生成手法の類型

ディープフェイクの生成手法は、主に次の3つに分類されることが多い(例えば、Le *et al.* 2025 ; Tariq *et al.* 2025)²。

- なりすましの対象者の顔の静止画を準備し、攻撃者自身の顔の動きに合わせてその静止画を変化させて動画を生成する(リエナクトメント)。
- なりすましの対象者の顔の静止画の一部に、攻撃者の顔の動画の一部³を切り

2 個々の手法に関しては、宇根[2024]において紹介されているので参照されたい。

3 一般に、目、鼻、口、耳などの特徴的な部位とそれらの周辺領域である。

取って埋め込むことによって動画を生成する（リプレースメントまたはフェイス・スワップ）。

- なりすまし対象者の特徴を示すテキストなどを生成 AI モデルに入力してなりすまし対象者の動画を生成する（シンセティック・フェイス・ジェネレーション）。

(2) ディープフェイクによるなりすましの実験

顔の静止画や動画による生体認証システムにおいて、ディープフェイクによるなりすましの実現可能性を検証した 3 件の実験の概要を説明する。

イ. 独自に構築したシステムにおける実験

(イ) 川名らによる実験

川名らは、スマートフォン内蔵のカメラで顔の動画を取得し、別途撮影した運転免許証の静止画と照合する生体認証システムを構築した（川名ほか [2021]）。そのうえで、オープンソースのツールを用いて生成したディープフェイクが誤って受け入れられるか否かを検証した。

実験では、なりすまし対象者の運転免許証を撮影して顔の静止画を取得し、生体認証システムへ送信する。次に、攻撃者の顔の動画をカメラで取得し、それをディープフェイク生成ツールに入力してディープフェイク（なりすまし対象者の顔の静止画に攻撃者の動画の動きを反映させたもの）を生成する。ディープフェイクをディスプレイに表示し、それをスマートフォンのカメラで撮影して取得した動画を生体認証システムへ送信する。生体認証システムは、運転免許証の静止画とディープフェイクを照合し、一致するか否かを判定する。

実験は 1 人の実験協力者（なりすまし対象者）に対して行われ、静止画がディープフェイクと一致すると誤って判定された⁴。

(ロ) Šalko らによる実験

Šalko らは、顔の動画や静止画による生体認証システムを構築し、公開されている顔の動画のデータセット（人間の顔の動画とディープフェイクを格納）を用いてディープフェイクによるなりすましの可能性を検証した（Šalko, Firc, and Malinka

.....
4 川名らは、後続の研究として、この生体認証システムにディープフェイクの検知モデル（4 件）を追加し、ディープフェイクを検知できるか否かも検証している（川名ほか [2024]）。検証の結果、ほとんどの検知モデルがリプレースメントに基づくディープフェイクを 50% 以上の確率で検知したが、リエナクトメントに基づくディープフェイクを 50% 以上の確率で検知できた検知モデルは 1 つだけであった。

2024)。

実験では、データセットから、なるべく正面を向いた個人の顔の動画と、その個人に対応するディープフェイクをそれぞれ選択したうえで、両者を生体認証システムに入力して一致するか否かを判定させるというものである。実験の結果、一致と誤って判定する確率が、異なる個人の顔の動画を照合した際に一致と誤って判定する確率よりも高いことが判明した。Šalko らは、ディープフェイクを考慮せずに生体認証システムの判定しきい値を設定した場合、ディープフェイクを誤って受け入れられる確率が高くなる可能性があると考えしている。

ロ. クラウドで提供されている生体認証サービスにおける実験

Li らは、クラウドで提供されている顔の動画による生体認証のサービス（6 件）を対象に、ディープフェイクが提示された際に受け入れられるか否かを実験によって検証した（Li *et al.* 2022）。

対象となった生体認証サービスでは、いずれも、各サービス利用者が認証に用いる顔の動画をスマートフォン内蔵のカメラで撮影し、撮影したデータをクラウド上のサーバに送信する。サーバは、受信した動画を事前に登録されたものと照合し、同一の個人か否かを判定する。顔の動画の形態はサービスによって異なっており、提示された動画が人間を撮影したものであることを確認するために、顔の向きをランダムに変化させるケースや、ランダムに示した数字を発音させるケースが準備されていた。

実験では、事前に登録された顔の動画を模倣するディープフェイクを生成し、サーバに送信して判定させた。その結果、顔の向きをランダムに変化させるサービスでは、誤って一致と判定した確率が約 80% となったディープフェイクが存在した。また、ランダムに示した数字を発音させるサービスでは、誤って一致と判定した確率が約 60% となったディープフェイクが存在した。

この結果は、実運用されている生体認証サービスにおいても、ディープフェイクによるなりすましに対して脆弱なケースがあることを示している。

3. ディープフェイク検知手法の類型と検知モデルの評価・比較

生体認証システムにおけるディープフェイク対策としては、入力データ（顔の動画や静止画）を登録済みのものと照合する処理に加えて、入力データが「人間の顔を撮影したものか否か」を判定する処理を準備するケースが多い。後者の処理として、生体検知（liveness detection）⁵の手法が広く活用されてきたが、近年では、ディープフェイクか否かを判定する機械学習の手法が数多く提案されている。

以下では、機械学習を用いたディープフェイク検知手法に焦点を当てて、それらの類型や検知モデルの評価・比較研究の事例を紹介する。

(1) 主な類型

ディープフェイク検知手法の類型として、まず、最近注目を集めている LLM を用いる手法と用いない手法に大別できる。そのうえで、LLM を用いない手法については、動画を構成する個々のフレーム内の情報に着目する手法、個々のフレーム間の関係性に着目する手法、電子透かしを用いる手法に分けることができる (Zou *et al.* 2025)。

イ. LLM を用いる手法

この類型は、LLM を用いてディープフェイクか否かを判定する手法である。これらの手法では、テキストに加えて動画も入力することができるマルチモーダル型の LLM (MLLM: Multi-modal LLM)⁶ が用いられる (Komaty *et al.* 2025 ; Song *et al.* 2025)。MLLM では、一般に、動画とともにその処理方法がテキスト (プロンプト) として入力され、処理結果がテキストとして出力される。ディープフェイク検知の文脈では、判定したい動画と「動画がディープフェイクか否かを判定せよ」との主旨のプロンプトが MLLM に入力され、判定結果が出力される (Gani *et al.* 2025)。また、プロンプトに「判定の理由を回答せよ」と付け加えることによって、判定の理由も出力させる手法もある (Zou *et al.* 2025)。ただし、本稿執筆時点では、顔の動画のディープフェイクに焦点を当てて複数の MLLM を評価・比較した研究成果は、筆者が知る限り公開されていないようである⁷。

ロ. LLM を用いない手法

(イ) フレーム内の情報に着目する手法

この類型は、動画の各フレームに含まれる情報を手掛りにフレームごとにディープフェイクか否かを判定し、最後に、各フレームの判定結果を統合して動画全体としてディープフェイクか否かを判定するというものである。フレームごとの判定に

.....
5 生体検知は、生体認証システムにおけるセンサーの対象が人間か否かを判定することを目的とする技術であり、主に、センサーに人工物を提示してなりすましを試みる攻撃 (presentation attack) への対策として研究されてきた。生体検知はディープフェイクを提示する攻撃に対しても対策となりうるが、2 節 (2) ロで紹介した Li らの実験で示されているとおり、生体検知のみではディープフェイクを十分に検知できないケースも想定される。

6 LMM (Large Multi-modal Model) と呼ばれることもある。

7 顔の静止画のディープフェイクに関する評価・比較の論文はいくつか発表されている (例えば、Narayan, VS, and Patel 2026)。

において、例えば、顔の不自然な色や影、顔の形状の歪み、顔の特徴点（目、鼻、口、耳など）の位置関係、頭部の三次元構造を検証する手法が知られている（例えば、Bai, Lin, and Cao 2024）。

（ロ） フレーム間の関係性に着目する手法

この類型は、動画のフレームに含まれる情報の時系列的な推移を主な手掛りとして判定する手法である。例えば、顔の形状、特徴点の位置、色などがフレーム間で整合的か否かを検証する手法が挙げられる（He *et al.* 2024；Yan *et al.* 2024）。

（ハ） 電子透かしを用いる手法

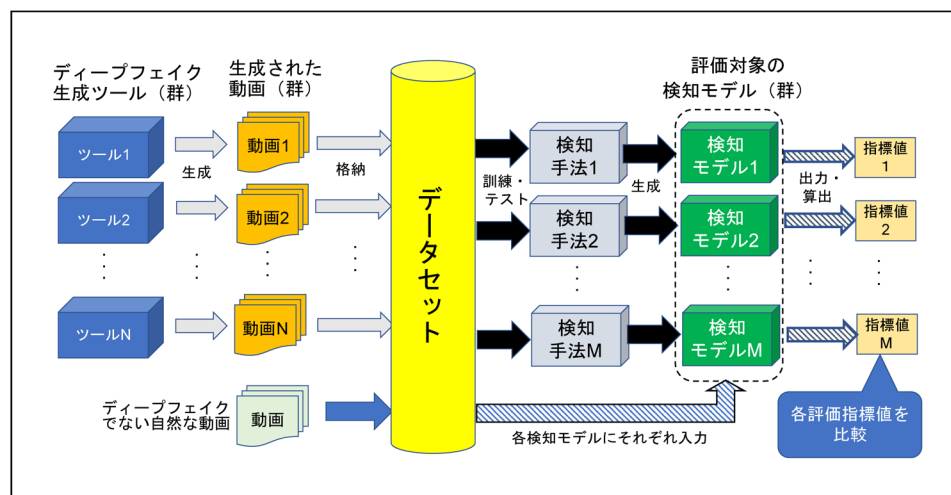
この類型は、動画を作成する際に特定のデータ（電子透かし）を動画の各フレームに埋め込み、ディープフェイクと区別できるようにするものである。ディープフェイクを検知する際には、対象となっている動画から電子透かしを抽出できるか否かを検証する。電子透かしを抽出できた場合、その動画をディープフェイクでないとして判定する（Luo *et al.* 2023；Zhang *et al.* 2023）。

（2） 検知モデルの評価・比較の方法

本節（1）で紹介した各検知手法のうち、電子透かしを用いる手法については、複数の手法を対象に評価・比較した研究成果が少ないことから対象外とする。検知モデルの評価・比較は一般的には次の流れで実施される（図1参照）。

- ① データセットを準備する。
 - データセットには、ディープフェイクと現実のデータの両方が含まれる。ディープフェイクについては、例えば、既存のディープフェイクの生成ツールを用いてディープフェイクを生成したり、インターネットに存在するディープフェイクを収集したりする。
- ② 評価対象となる検知モデル（複数）を生成する。
 - 検知モデルの生成に用いるデータ（訓練用データやテスト用データ）は、データセットから抽出する。検知モデルが生成済みの場合、改めてモデルを生成するのではなく、既存のものを使用するケースがある。
- ③ 評価指標値を算出する。
 - 各検知モデルに評価用データ（データセットの一部）を入力し、判定結果を出力するとともに、評価指標値を算出する。セキュリティの観点では、ディープフェイクをなるべく見逃さないことが重要であり、ディープフェイクが入力されたときにそれを正しく判定した割合（再現率）が高いことが

図1 ディープフェイク検知モデルの評価・比較の流れ



望ましい。また、アプリケーションの可用性の観点から、現実のデータが入力されたときに誤ってディープフェイクと判定した割合（偽陽性率）が低いことが望ましい。そこで、再現率と偽陽性率の関係を表す ROC（Receiver Operating Characteristic Curve）や AUC（Area Under the ROC Curve）⁸ が評価指標として採用される場合がある（図2 参照）。

- 偽陽性率の代わりに、ディープフェイクと判定されたケースにおいてその判定が正しかったものの割合（適合率）を採用し、適合率と再現率が同時に高くなるケースが望ましいとする見方もある。この場合、適合率と再現率の関係を表す F 値（F score、図3 参照）⁹ が評価指標として用いられる。
- このほか、正解率¹⁰ を評価指標とするケースもある。

④ 検知モデル間で評価指標値を比較する。

.....
8 X 軸に偽陽性率、Y 軸に再現率をとって両者の関係をグラフで表したときの曲線が ROC であり、ROC の積分値が AUC である。AUC は 0 から 1 の間の値をとり、値が大きいくほど判定精度が高いと評価される。

9 F 値は、適合率と再現率の調和平均として定義され、0 から 1 の間の値をとる。適合率と再現率が同時に高い値を示すと F 値も大きな値となり、判定精度が高いと評価される。

10 正解率は全試行に対する正しい回答（ディープフェイクが入力された場合にはディープフェイクであると判定し、ディープフェイクでない入力の場合にはディープフェイクでない判定）の数の割合であり、0 と 1 の間の値となる。値が大きいくほど望ましい。

図2 ROC と AUC

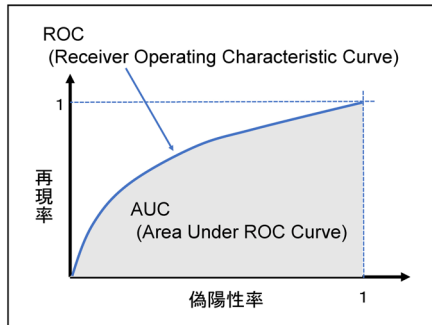
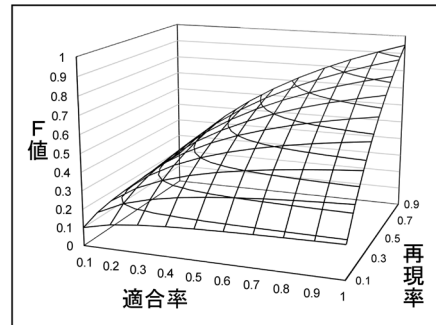


図3 適合率・再現率と F 値



(3) 主な研究事例

顔の動画や静止画のディープフェイクに対する検知モデルを評価・比較した最近の主な研究事例 4 件（LLM を用いないもの 3 件、LLM を用いるもの 1 件）を紹介する（表 1 参照）。ここでは、第三者による評価・比較の論文に絞り、特定の手法の提案やその優位性を主張する論文を対象外とした。また、MLLM の評価・比較については、論文執筆時点において顔の動画のディープフェイクを対象とするものが見当たらなかったため、顔の静止画のディープフェイクを対象とするものを取り上げる。

イ. Yan らの研究

(イ) 概要

Yan らは、顔の動画のディープフェイクを対象とする検知モデル 15 件¹¹を対象に評価・比較した（Yan *et al.* 2023）。具体的には、① AI・機械学習分野におけるトップクラスの学会や論文誌で発表されている、②顔の静止画のディープフェイクの検知にも使用できるように、動画のフレーム内の情報を手掛りとする、③（検知モデルの基になる）検知手法がオープンソースとして公開されているという条件を満たす検知モデルが対象として選ばれた。

評価には、公開されているディープフェイクのデータセット 9 件（ディープフェイクは合計約 13 万件）¹² が用いられ、これらに含まれる動画の顔の部分抽出し、

.....
11 対象となった検知モデルは、MesoNet、MesoInception、CNN-Aug、EfficientNet-B4、Xceptionnet、Capsule Forensics、RECCE、Face X-ray のほか、Xceptionnet をベースとした 7 件（FFD、CORE、DSP-FWA、UCF、F3Net、SPSL、SRM）である。

12 具体的には、FaceForensics++、CelebDF-v1、CelebDF-v2、DeepFake Detection、DeepFake Detection Challenge Preview、DeepFake Detection Challenge、DeeperForensics-1.0、FaceShifter、UADFV である。これらのうち、FaceForensics++、DeepFake Detection、DeepFake Detection Challenge Preview、DeepFake

表 1 ディープフェイク検知モデルを評価・比較した主な研究事例

研究事例 (文献)	データセットと ディープフェイク生成手法	評価対象の 検知モデル	主な評価 指標値	評価指標値が 高い検知モデル
Yan <i>et al.</i> 2023	約 13 万件の動画（公開済みの データセットから抽出）	15 件 ・ LLM を用いない ・ 独自にモデルを生 成	AUC 再現率 適合率 正解率	SPSL
Bei <i>et al.</i> 2024	約 42 万件の動画（16 件の手法 を用いて独自に生成）	7 件 ・ LLM を用いない ・ 独自にモデルを生 成	AUC	Exposing
Le <i>et al.</i> 2025	・ 約 200 件の動画（7 件の手法 を用いて独自に生成） ・ 約 1,400 件の動画（動画プ ラットフォームから抽出）	16 件 ・ LLM を用いない ・ 既存のモデルを使 用	AUC F 値 再現率 適合率 正解率	FTCN AltFreezing
Narayan, VS, and Patel 2026	・ 300 件のプロンプト（独自に 生成） ・ 300 件の静止画（公開済みの データセットから抽出）	28 件（MLLM）	正解率	GPT-4o

統一的なフォーマットに加工・修正したものが使用された。ディープフェイクの生成手法は、主に、リエナクトメントまたはリプレイスメントに分類されるものである。各検知モデルは、9 件のデータセットのうちの 1 件に含まれる動画（群）によって生成された。

（ロ） 評価・比較

Yan らは、検知モデルの訓練に使用したデータセットを除く 8 件のデータセットに含まれるデータをそれぞれ各検知モデルに入力し、評価指標値（AUC など）を算出した。その結果、いずれの検知モデルにおいても AUC の値がデータセットによって大きく変化し¹³、すべてのデータセットに対して偏りなく高い判定精度を示す検知モデルは存在しなかった。

また、Yan らは、生成手法に基づいて（すべてのデータセットの）ディープフェイクを 4 つのグループに分けたうえで、そのうちの 1 つのグループに属するディープフェイクのみを用いて各検知モデルを生成し、残りの 3 つのグループのディープ

.....
Detection Challenge には、複数の異なる手法によって生成されたディープフェイクが混在している。他のデータセットのディープフェイクはそれぞれ同じ手法で生成されている。

13 例えば、SPSL（Spatial-Phase Shallow Learning）は、AUC の平均値（約 0.79）が最も高かったものの、データセットによって AUC の値が 0.64 から 0.95 の間で変化した。SPSL は、動画のフレーム内の情報を手掛りとする手法の 1 つであり、顔の形状、特徴点の位置などの情報に加え、画素値の系列を周波数領域の系列に変換したデータを用いて検知する（Liu *et al.* 2021）。

フェイクを各検知モデルにそれぞれ入力して評価指標値を算出した（クロス・バリデーション）。その結果、いずれの検知モデルにおいても、入力したディープフェイクのグループによって AUC の値が大きく変化した¹⁴。

Yan らは、今回対象とした検知モデルでは、訓練用データに含まれていないディープフェイクを高い確率で検知することが難しいとの見方を示した。

ロ. Bei らの研究

（イ） 概要

Bei らは、顔の動画や静止画のディープフェイクのデータセットを構築するとともに、顔の動画や静止画のディープフェイクを対象とする検知モデルを評価・比較した（Bei *et al.* 2024）。動画のディープフェイクを対象とする検知モデルとして、提案論文などから比較的高い性能を期待できる 7 件を対象とした¹⁵。

データセットは、顔の動画・静止画のディープフェイクを生成する手法 34 件によって独自に生成されたディープフェイク（約 35 万件の静止画、約 42 万件の動画）などから構成されている。動画のディープフェイクは 16 件の手法によって生成され、リエナクトメントおよびリプレースメントに属するものに加え、シンセティック・フェイス・ジェネレーションに属する手法も用いられている¹⁶。

（ロ） 評価・比較

Bei らは、クロス・バリデーションによって各検知モデルの評価指標値（AUC）を算出した。その結果、いずれのモデルも、入力されたディープフェイクの種類によって AUC の値が大きく変化した¹⁷。

また、Bei らは、各検知手法において、シンセティック・フェイス・ジェネレーションによるディープフェイクで生成した検知モデル（モデル S）と、リプレースメントやリエナクトメントによるディープフェイクで生成した検知モデル（モデル R）をそれぞれ準備し、モデル S に対してリプレースメントやリエナクトメントによるディープフェイクを入力した際の判定精度と、モデル R に対してシンセティッ

14 各検知モデルにおける AUC の値は 0.50 から 0.80 の間の値であった。

15 対象となった検知モデルは、CViT、EfficientNet-B0、Exposing、F3Net、MesoNet、SLADD、Xceptionnet である。

16 リエナクトメントに属する生成手法が 4 件（ATVG-Net、FOMM、Motion-cos、Talking Head Video）、リプレースメントに属する手法が 7 件（BlendFace、DeepFakes、DSS、FaceShifter、FSGAN、MMReplacement、SimSwap）、シンセティック・フェイス・ジェネレーションに属する手法が 5 件（AnimateDiff-Lightning、AnimateLCM、Hotshot、MagitTime、Zeroscope）である。

17 例えば、AUC の値が比較的高かった Exposing では、モデル生成に用いられていないタイプのディープフェイクを入力すると、AUC の値が 0.59 から 0.90 の間で変化した。Exposing は、各フレーム内の特徴点の周辺領域の情報を手掛りにディープフェイクが否かを判定する手法であり（Ba *et al.* 2024）、顔の左右の対称性（形状、位置、皮膚の色など）、目・口・歯の輪郭の明確性、これらのフレーム間の整合性などを検証する（複数のフレーム間の関係性を手掛りにする手法の類型に属する）。

ク・フェイス・ジェネレーションによるディープフェイクを入力した際の判定精度をそれぞれ算出・比較した。その結果、モデル S の判定精度がモデル R の判定精度よりも低くなる傾向がみられた¹⁸。

今回対象とした検知モデルについて、Bei らは、さまざまなタイプのディープフェイクを偏りなく高い確率で検知することが難しく、検知モデルの訓練用データに含まれていないディープフェイクが入力されると判定精度が低下する可能性があるとの見方を示している。そのうえで、実際の生体認証に適用する際には、想定されるディープフェイクの種類を考慮して検知モデルを選択する必要がある旨を指摘している。

ハ. Le らの研究

(イ) 概要

Le らは、顔の動画のディープフェイクのデータセットを構築するとともに、検知モデル 16 件¹⁹ を評価・比較した (Le *et al.* 2025)。対象とする検知モデルは、2019 年から 2023 年にかけて主要な学会・論文誌において提案論文が発表されている 51 件のなかから、複数のタイプのディープフェイクを検知可能である旨が提案論文で主張されているとともに、検知モデルが生成済みのものを選択した (生成済みの検知モデルを用いて評価)。なお、各検知モデルの訓練用データは統一されていなかった。

Le らは、7 種類の手法²⁰ によってそれぞれ生成したディープフェイク (約 200 件)、および、インターネット上の動画プラットフォーム²¹ で公開されているディープフェイク (約 1,400 件) を含むデータセットを構築した。動画プラットフォームのディープフェイクの生成手法は不明であった。

(ロ) 評価・比較

Le らが独自に生成したディープフェイクをそれぞれ各検知モデルに入力した

18 この傾向について、Bei らは、モデル R がフレーム内の特徴点に着目して判定するのに対して、モデル S は特徴点以外の顔や背景の領域も考慮して判定しているためと考察している。モデル S は、リブレイスメントやリエナクトメントによるディープフェイク (特徴点周辺のみが合成対象) において、合成されていない領域 (特徴点周辺以外の顔の部分や背景) もチェック対象とするため、ディープフェイクであると判定しづらくなる。モデル R は、シンセティック・フェイス・ジェネレーションによるディープフェイクにおいて、特徴点周辺のみをチェック対象とする。Bei らはチェック対象の違いが判定精度の差異につながったとしている。

19 検知モデルは、ADD、AltFreezing、CADDMM、Capsule Forensics、CCViT、CLRNet、EfficientNet-B4、FTCN、ICT、LGrad、LipForensics、LRNet、MAT、MCX-API、SBIs、Xceptionnet である。

20 生成手法はオープンソースとして提供されているものであり、DeepFaceLab、Dfaker、Faceswap、FOM-Animation、FOM-Faceswap、FSGAN、LightWeight であった。いずれもリエナクトメントまたはリブレイスメントに分類される手法によるものであった。

21 動画プラットフォームは、Bilibili、Reddit、TikTok、YouTube であった。これらの動画のうち、AI によって生成された旨のラベルが付与されているものがディープフェイクとして使用された。

ところ、2 件の検知モデル (FTCN、AltFreezing)²² の判定精度がいずれのタイプのディープフェイクに対しても比較的高い値となった²³。Le らは、これらの検知モデルがいずれも複数のフレーム間における顔の形状などの時系列的な整合性を手掛りとしていることから、今回評価に用いたディープフェイクに対して、複数のフレーム間の関係性を手掛りとする手法が有効であったとしている。

動画プラットフォームから入手したディープフェイクを各検知モデルに入力したところ、ディープフェイクをまったく検知できなかった検知モデルが 2 件存在していたほか、その他の検知モデルにおいても判定精度が低下した²⁴。判定精度の低下について、Le らは、各検知モデルの訓練用データに含まれていなかったタイプのディープフェイクが動画プラットフォームに存在し、それらを検知することができなかった可能性があるとしている。

二. Narayan らの研究

(イ) 概要

Narayan らは、顔の静止画を対象とする各種タスクを実行する MLLM²⁵ を評価・比較するためのデータセット (FaceXBench) を構築するとともに、それぞれのタスクに関して複数の MLLM を評価・比較した結果を発表した (Narayan, VS, and Patel 2026)。ここでは、顔の静止画のディープフェイクを検知するタスクに関する結果に絞って紹介する。

データセットとして、公開済みのデータセット 2 件²⁶ から抽出されたディープフェイク 300 件と、ディープフェイク検知用のプロンプト 300 件 (独自に作成) が準備された。300 件のプロンプトには以下の 3 種類が含まれていた²⁷。

- 複数のディープフェイクを提示しつつ、タスクの内容を説明しないで、「これらのうち、どの静止画がディープフェイクですか？」という質問のみを含むプロンプト (ゼロショット・タスク)

.....
22 FTCN (Fully Temporal Convolution Network) は複数のフレームにおいて顔の形状や特徴点の位置などが整合的か否かを重視する検知モデルである (Zheng *et al.* 2021)。AltFreezing は、動画の各フレーム内の情報、および、複数のフレーム間の関係性に基づいて別々のニューラル・ネットワークを生成し、それらを組み合わせた検知モデルである (Wang *et al.* 2023)。

23 FTCN と AltFreezing の評価指標値は、平均すると、AUC の値がともに約 0.98、F 値がそれぞれ約 0.93、約 0.94 であった。また、再現率はそれぞれ約 0.95、約 0.94 であった。

24 各検知モデルにおける AUC の値は 0.39 から 0.69 の間であった。FTCN と AltFreezing においても AUC の値がそれぞれ約 0.58、約 0.60 となった。

25 ここでの MLLM は、テキストと静止画が入力として与えられ、テキストによって示されるタスクの回答をテキストで出力するものである。

26 データセットとして CelebDF と FaceForensics++ が選択された。

27 300 件のプロンプトのうち、3 種類のプロンプトがそれぞれいくつ含まれていたかについては、論文に記載されていない。

- タスクの内容を説明したうえで、複数のディープフェイクを提示しつつ、「どの静止画がディープフェイクですか？」という質問を行うプロンプト（インコンテキスト・タスク）
- タスクの内容を説明したうえで、ディープフェイクを提示しつつ、「ディープフェイクですか？ 質問への回答をステップ・バイ・ステップで考えてください。回答を出力するだけでなく、その理由や考え方も示してください。」という指示を含むプロンプト（チェーンオブソート・プロンプティング）

評価対象とした MLLM については、オープンソースとして公開されているもの 26 件²⁸ と、商用サービスとして提供されているもの 2 件（GPT-4o、GeminiPro 1.5）が選択された。

（ロ） 評価・比較結果

Narayan らは、各 MLLM に対してプロンプトをそれぞれ入力して正解率を算出した。すべてのプロンプトを対象とした場合の正解率は、いずれの MLLM においても低い値となった²⁹。また、3 種類の MLLM を対象に、プロンプトの種類別に正解率を算出・比較したところ、正解率はほとんど変化しなかった³⁰。

この結果を踏まえ、Narayan らは、今回対象とした MLLM における判定精度が低く、現時点ではディープフェイク検知に有効な手段とは言い難いほか、プロンプトの種類を変更しても判定精度の向上を期待しづらいとの見方を示した。

（4） 小括

本節（3）で紹介した研究事例をまとめると以下のとおりである。

- Yan らの事例、Bei らの事例、Le らの事例では、データセット、ディープフェ

28 対象となったオープンソースの MLLM は、Chameleon-7b、CogVLM2-19b、Eagle-X4-8B-Plus、Idedics-9b-Instruct、Idedics-80b-Instruct、InternVL-Chat-v1.5、InternVL2-8b、InternVL2-76b、LLaVA-OneVision-0.5b-OV、LLaVA-OneVision-7b-OV、LLaVA-OneVision-7b-SI、LLaVA-OneVision-72b-OV、LLaVA-NeXT-Interleave-7b、LLaVA-v1.5-7b、LLaVA-v1.5-13b、Mantis-SIGLIP-8b、MiniCPM-Llama3-v2.5、Monkey-Chat、PaliGemma、Phi-3.5-Vision、Qwen2-VL-7b-Instruct、Qwen2-VL-72b-Instruct、VILA 1.5-3b、VILA 1.5-13b、VILA 1.5-40b である。

29 正解率はいずれの MLLM においても約 0.05 から約 0.30 の間の値となった。最も高い正解率（約 0.30）となったのは GPT-4o であった。

30 対象となった MLLM は、InternVL2-8b、Phi-3.5-Vision、Qwen2-VL-7b-Instruct であり、ゼロショット・タスク、インコンテキスト・タスク、チェーンオブソート・プロンプティングの場合の正解率が、それぞれ、約 0.27 から約 0.30 の間の値、約 0.27 から約 0.32 の間の値、約 0.24 から約 0.32 の間の値となった。

イク生成手法、検知モデルの種類や準備方法（独自生成、または、既存のものを使用）などが区々であり、事例間の比較が困難である。また、検知モデルの訓練用データに含まれていないディープフェイクに対して、いずれの検知モデルも判定精度が低下する傾向がみられた。

- 上記の3つの事例において評価指標値が相対的に高かった検知モデルは区々であり、高い判定精度を期待できる検知手法のタイプや特徴を特定するには至っていない。
- Narayan らの事例では、いずれの検知モデルにおいても正解率が低く、現時点では現実的な対策として有効とはいえない。

4. 考察

本節では、ディープフェイク検知モデルの活用を検討する主体（ユーザ）が、リスク・ベースでの意思決定を志向する場合、検知モデルの評価・比較の研究事例をどのように活用することができるかを考察する。また、評価・比較を実施する研究コミュニティ側での主な課題についても考察する。

(1) ユーザ側の準備：検知モデルに関する要求事項

ここでは、顔の動画を用いた生体認証において、ディープフェイクによるなりすましのリスクを低減するために有効な検知モデルを選択することを目標とする。ユーザは、悪用される可能性があるディープフェイクの生成手法をリストアップするとともに、検知モデルの評価基準を設定しておくことが望ましい。

イ. ディープフェイク生成手法のリストアップ

悪用される可能性があるディープフェイクの生成手法として、公開済みの手法すべてが候補となる。もっとも、すべてをリストアップして検討することは現実的とはいえないため、対象を絞り込んだうえで優先順位を付けることが考えられる。例えば、以下のようなタイプの生成手法を対象とすることが考えられる。

- オープンソースとして公開されており容易に生成モデルを入手・使用できるなど、ディープフェイク生成に要する特別なノウハウやコストが不要な手法
- ディープフェイクによるなりすましの成功確率が比較的高いと評価されている手法

- ディープフェイクを高速・リアルタイムで生成できる手法

こうした考慮点を列挙し、より多くの考慮点に合致するものに高い優先順位を付けてリストアップすることが考えられる。

ディープフェイク生成手法は今後も進化する可能性があるため、生成手法のリストを継続的にメンテナンスすることが望ましい。想定される生成手法のリストを作成しても、時間の経過とともに既存手法の改良や新手法の出現によってリストは陳腐化する。ディープフェイク生成手法の研究動向をフォローし、新手法の追加や陳腐化した手法の削除などを定期的実施することが重要である。

ロ. ディープフェイクの検知モデルの評価基準の設定

(イ) 許容できるなりすまし成功確率の上限

ユーザは、生体認証によって保護するアプリケーションにおいて、ディープフェイクによるなりすましの被害額を見積もったうえで、なりすましの成功確率がどの程度であれば許容できるかを明らかにしておくことが望ましい。例えば、以下のステップでこれを行うことが考えられる（宇根 [2024]）。

① 一定期間におけるなりすましによる被害額（X）を見積もる。

➤ X は、一定期間におけるなりすましの試行回数の上限³¹ と、1 回のなりすまし成功によって生じうる被害額の上限³² をそれぞれ見積もり、両者を掛け合わせて算出する。

② 許容できる被害額の上限（Y）を見積もる。

③ $Y \div X = Z$ を計算し、許容できるなりすまし成功確率の上限とする。

なりすまし成功確率は、ディープフェイクが入力として与えられた際に正規のサービス利用者と誤って判定する確率であり、その確率は「1－再現率」に相当する。したがって、上記の手順で Z の値が決定されると、再現率が「1－Z」以上の値となる検知モデルが採用の候補となる。「1－Z」の値は、許容できる再現率の下限といえる。

.....
31 なりすましの試行回数の上限として、例えば、生体認証システムのサービス利用者数に、一定期間中にサービス利用者が生体認証システムにアクセスできる回数の上限を乗じて算出することが考えられる。また、既に生体認証システムが運用されており、なりすましなどの不正行為が観察されている場合、一定期間内の不正行為の回数をなりすまし試行回数とすることも考えられる。

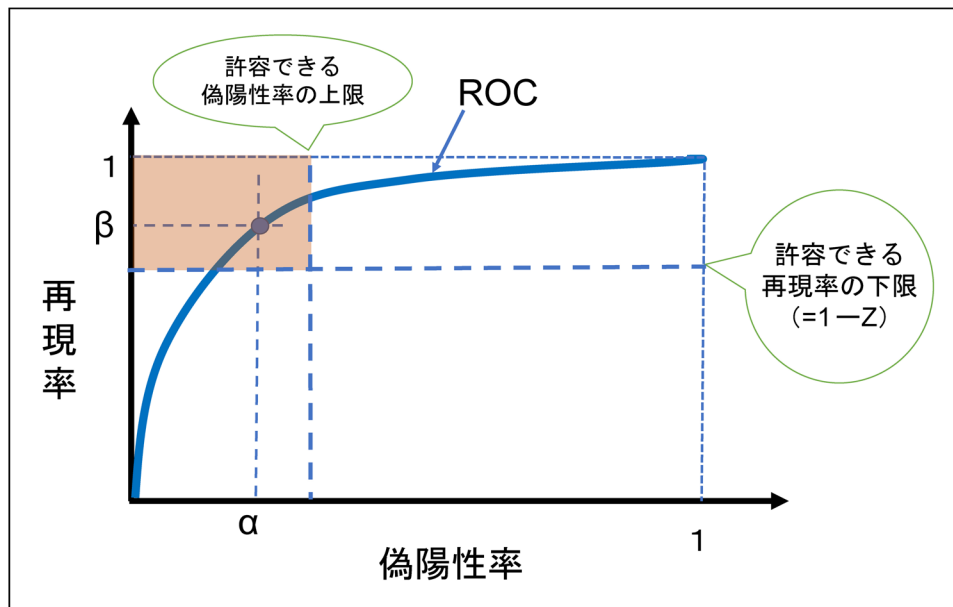
32 被害額の上限に関しては、1 回のなりすましによって実行可能な取引額の上限や、不正に入手した情報資産の価値の上限によって見積もることが考えられる。

(ロ) 許容できる偽陽性率の上限

再現率などの評価指標値は検知モデルの判定しきい値によって変動する。例えば、ある判定しきい値のもとで、再現率が「 $1-Z$ 」の値よりも大きくなったとしても、偽陽性率も高い値となる可能性がある³³。偽陽性率が高いと、正規のサービス利用者が生体認証をパスできずアプリケーションを使用できないケースが頻発する。ユーザは、そのアプリケーションにおいて許容できる偽陽性率の上限を設定しておくことが望ましい。

許容できる偽陽性率の上限が決まれば、ユーザは検知モデルの評価結果（ROC）を参照し、偽陽性率と再現率の条件が同時に達成される判定しきい値が存在するか否かを確認する（図4参照）。そうした判定しきい値が存在すれば、採用の候補となる。

図4 評価基準を満たす再現率と偽陽性率



(2) 評価・比較の研究事例を活用する際の留意点

想定されるディープフェイク生成手法のリストアップ・優先順位付け、および、

.....
³³ 一般的に、高い再現率を達成するために判定しきい値を低く設定すると、偽陽性率も高くなる傾向がある。

検知モデルの評価基準の設定を行った後、検知モデルの評価・比較の研究事例を活用して検知モデルを選択する際の留意点を考察する。

イ. ディープフェイク生成手法

ユーザは、研究事例で使用されているデータセットに含まれているディープフェイク生成手法を確認し、それらのなかにリストアップしたものが含まれているか否かを確認することが望ましい。含まれているならば、評価・比較の結果を参照する意義がある。

3 節 (3) で紹介した Bei らの事例と Le らの事例では、データセットを構成するディープフェイクが独自に生成され、その際に用いられた生成手法が公開されていることから、準備したリストと照らし合わせることができる。Yan らの事例では、ディープフェイクの生成に使用されたデータセットが公開されていることから、各データセットの情報を入手してディープフェイク生成手法を確認し、準備したリストと照合することができる。

一方、Le らの事例のうち、動画プラットフォームから抽出したディープフェイクによる評価に関しては、ディープフェイク生成手法が不明であり、準備したリストと照合することが困難である³⁴。

ロ. 検知モデルの生成に用いられたデータ

評価対象の検知モデルの生成にどのようなデータが用いられたかを確認しておくことは、評価の妥当性を検討するうえで重要である。また、こうした情報は、評価・比較の結果として選択した検知モデルを生体認証システムに実装する際に、評価対象であった検知モデルを再現する（または検知モデルを改善する）うえで有用である。確認事項として主に次の点に留意することが望ましい。

- リストアップした生成手法によるディープフェイクが検知モデルの生成に使用されているか。
 - 使用されていないとすれば、リストアップした生成手法によるディープフェイクが検知モデルに入力された際に、判定精度が低下する可能性がある。
- 検知モデルの訓練用データにおけるディープフェイクとそれ以外のデータの量やバランスが妥当か。
 - 訓練用データの量ができるべく多い方が望ましい。ディープフェイクの量が

34 こうしたディープフェイクであっても、各検知モデルに入力して判定精度を算出・比較することは可能である。例えば、リストアップしたディープフェイクに関する評価結果が同一となる検知モデルが複数存在し、それらの優劣を決定する必要がある場合に、生成手法が不明なディープフェイクに関する評価結果を参照するという方法がありうる。

少ないと、本来の判定精度が発揮されず低い精度となる可能性があることから、評価・比較の事例として参考にできない場合がある。

- ディープフェイクの量と現実の動画のデータの量の比率に偏りがあると、モデルの出力にも偏りが生じる可能性がある³⁵。
- 複数の検知モデルの評価・比較において訓練用データが統一されていたか。
 - 各検知モデルの生成において異なる訓練用データが用いられていた場合、判定精度を同一条件で比較することが困難となり、評価指標値を比較する意義が失われる可能性がある。

3 節 (3) における Yan らの事例と Bei らの事例では、検知モデルが独自に生成されており、モデルの訓練用データの詳細をそれぞれの論文執筆者に問い合わせることができる。一方、Le らの事例では、既に生成されていた検知モデルが評価対象となっており、各モデルの訓練用データを把握するためには、各モデル生成者に問い合わせる必要がある。また、訓練用データが各モデルで異なり、統一されていない可能性が高いとみられる。

Narayan らの事例の対象となっている MLLM の場合には、LLM をベースにそれぞれ生成されており、モデルの訓練用データの詳細を把握することは難しい。そのため、MLLM をブラックボックスとして評価結果を解釈する必要がある。

ハ. 評価指標

研究事例で示される評価指標値には、再現率と偽陽性率、および、これらの関係を示す ROC が含まれていることが望ましい。これらが示されているならば、ユーザは許容できるなりすまし成功確率との関係を考慮し、生体認証システムへの実装の対象となりうるか否かを検討することができる。

3 節 (3) で紹介した Yan らの事例、Bei らの事例、Le らの事例では、いずれも AUC が評価指標となっており、AUC の値を算出する際に用いられた ROC から再現率と偽陽性率の関係を知らることができる³⁶。一方、Narayan らの事例では、正解率のみが示されており、AUC が示されていないことから、再現率と偽陽性率の関係を確認することができない可能性がある。

.....
35 Layton らは、訓練用データやテスト用データに含まれるディープフェイクの量と現実の動画の量の比率を変化させながら、それぞれ検知モデルを生成して判定精度を算出したところ、比率の変化によって判定精度が有意に変化したと報告している (Layton et al. 2024)。Layton らは、検知モデル運用時の入力全体に占めるディープフェイクの割合を推定し、それに合致するように訓練用データに占めるディープフェイクの割合を決定する必要があるとしている。

36 研究事例の論文に AUC の記載があるものの、ROC が記載されていないというケースもありうるが、論文執筆者に ROC のデータを問い合わせることによって再現率と偽陽性率の関係を確認することができると思われる。

(3) 研究コミュニティにおける課題

3 節 (3) で紹介したように、評価の前提・内容・指標が研究事例の間で区々であり、第三者が複数の研究事例の結果を比較することは困難である。個々の研究事例のなかで、それぞれが対象としている検知モデルを評価・比較することができるに過ぎない。

対応として、研究者が、評価・比較研究の成果を公開する際に、データセットを構成するディープフェイクの生成手法、検知モデルの生成に用いられた訓練用データなどを公開し、別の研究者もそれらを可能な限り参照・使用できるようにすることが望ましい³⁷。評価・比較の追試も可能となり、結果の信頼性の向上にもつながると期待できる。

また、ユーザがなりすましによるリスクを考慮して検知モデルの評価基準（許容できる再現率の下限と偽陽性率の上限）を設定するとすれば、評価指標として、再現率と偽陽性率の関係（ROC）が公表されることが望ましい。この点についても各研究事例において統一されれば、研究事例間での比較可能性が高まると期待できる。

5. おわりに

本稿では、顔の動画を用いた生体認証への脅威としてディープフェイクによるなりすましを想定し、対策手段の 1 つであるディープフェイク検知に焦点を当てて、主な検知手法の類型、検知モデルの評価・比較に関する主な研究事例を紹介した。また、ユーザが研究事例を活用する際の留意点、研究コミュニティにおける主な課題について考察した。

顔の動画を用いた生体認証は、パソコンやスマートフォンに標準装備されているカメラや画像処理のソフトウェアを用いて実施可能である。そのため、生体認証を実施するうえで改めてカメラやソフトウェアを準備する必要がない（実装のためのコストが低い）ほか、パスワードを覚えたり認証のための特別なデバイス（ワンタイム・パスワード・トークンなど）を安全に保持したりする必要がない（生体

.....
37 こうした対応は既に一部の学会で始まっている。例えば、セキュリティ分野のトップ・カンファレンスの 1 つである USENIX Security Symposium では、投稿論文の査読の公正性を担保することに加え、研究成果の再現性（replicability）と再生産性（reproducibility）を促進するために（Open Science Policy）、投稿論文の執筆者に対して、執筆や研究に使用したデータやツールを投稿時点で公開するとともに、それらのリストや格納場所を投稿論文に明記することを求めている（The 35th USENIX Security Symposium Organizers 2025）。

認証の利用者の利便性が高い)。こうしたメリットを享受するという観点からは、ディープフェイクによるなりすましのリスクを他の技術を用いてどのように低減するかが論点となる。

対策としてディープフェイクの検知モデルを採用する際に、ユーザは、リスク管理の観点から、なりすましに悪用されうるディープフェイクの生成手法をリストアップし、検知モデルの評価基準（許容できる再現率の下限や偽陽性率の上限）を設定することが望ましい。そのうえで、公開されている研究事例を参照しつつ、評価基準に合致するものを検討することが考えられる。なりすましのリスクの評価などが困難な場合には、音声による認証など、他の認証手段と組み合わせるマルチモーダル認証を採用する（多重防御）、または、認証成功によって実行できる取引の金額の上限を低く設定するといった対応を検討することが有用である。

生体認証を活用している金融機関は、リスク管理の一環としてディープフェイクに関する技術動向をフォローし、現在採用しているなりすましへの対策手段の有効性やリスクのレベルを確認することが望ましい。ディープフェイクに関する研究は非常に活発であり、新しい生成技術や検知技術が次々に提案されている。既存の対策手段が陳腐化するスピードも速いことから、技術動向のフォロー、リスクの再評価、対策手段の見直しを継続的に行うことが重要である。

参考文献

- 宇根正志、「生体認証システムにおける人工物を用いた攻撃に対するセキュリティ評価手法の確立に向けて」、『金融研究』第35巻第4号、日本銀行金融研究所、2016年、55～90頁
- 、「スマートフォンによる顔認証のセキュリティ：ディープフェイクによる脅威と対策」、『金融研究』第43巻第4号、日本銀行金融研究所、2024年、95～122頁
- 川名のん・大島敬志・鈴木 茜・吉野雅之、「Deepfake 動画を用いた eKYC に対するなりすまし攻撃の検知手法の検討」、『第38回人工知能学会全国大会論文集』、4M3-GS-10-03、人工知能学会、2024年、1～4頁
- ・長沼 健・吉野雅之・太田原千秋・富樫由美子・笹 晋也・山本恭平、「Deepfake を用いた e-KYC に対するなりすまし攻撃と対策の検討」、『第35回人工知能学会全国大会論文集』、1F2-GS-10a-02、人工知能学会、2021年、1～4頁
- 金融庁、「AI ディスカッションペーパー（第1.0版）—金融分野における AI の健全な利活用の促進に向けた初期的な論点整理—」、金融庁、2025年（<https://www.fsa.go.jp/news/r6/sonota/20250304/aidp.pdf>、2026年2月26日）
- 坂本静生・宇根正志、『AI・量子コンピュータにかかわるリスク管理』、オーム社、2025年
- 笹原和俊、『ディープフェイクの衝撃：AI 技術がもたらす破壊と創造』、PHP 研究所、2023年
- Ba, Zhongjie, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. “Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection.” *Proceedings of the 38th AAAI Conference on Artificial Intelligence* 38, no. 2 (2024): 719–28. <https://doi.org/10.1609/aaai.v38i2.27829>.
- Bai, Jianfa, Man Lin, and Gang Cao. 2024. “AI-Generated Video Detection via Spatio-Temporal Anomaly Learning.” arXiv: 2403.16638. Accessed February 26, 2026. <https://arxiv.org/pdf/2403.16638>.
- Bei, Yijun, Hengrui Lou, Jinsong Geng, Erteng Liu, Lechao Cheng, Jie Song, Mingli Song, and Zunlei Feng. 2024. “A Large-Scale Universal Evaluation Benchmark for Face Forgery Detection.” arXiv: 2406.09181v2. Accessed February 26, 2026. <https://arxiv.org/pdf/2406.09181>.
- Deng, Jingyi, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Zhe Peng, Qian Wang, and Chao Shen. 2025. “A Survey of Defenses against AI-Generated Visual Media: Detection, Disruption, and Authentication.” arXiv: 2407.10575v3. Accessed February 26, 2026. <https://arxiv.org/pdf/2407.10575>.
- Financial Crimes Enforcement Network. 2024. “FinCEN Alert on Fraud Schemes Involv-

- ing Deepfake Media Targeting Financial Institutions.” Financial Crimes Enforcement Network. Accessed February 26, 2026. <https://www.fincen.gov/sites/default/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf>.
- Gani, Hanan, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan. 2025. “VANE-Bench: Video Anomaly Evaluation Benchmark for Conversational LMMs.” arXiv: 2406.10326v2. Accessed February 26, 2026. <https://arxiv.org/pdf/2406.10326>.
- Gaur, Loveleen. *DeepFakes: Creation, Detection, and Impact*. CRC Press. 2023.
- He, Peisong, Leyao Zhu, Jiaying Li, Shiqi Wang, and Haoliang Li. 2024. “Exposing AI-Generated Videos: A Benchmark Dataset and a Local-and-Global Temporal Defect Based Detection Method.” arXiv: 2405.04133v1. Accessed February 26, 2026. <https://arxiv.org/pdf/2405.04133>.
- Khan, Farrukh Aslam, and Muhammad Khurram Khan. “Generative AI and Deepfake Detection in Biometric Systems.” *Cognitive Computation* 17, no. 112 (2025): 1–21. <https://link.springer.com/article/10.1007/s12559-025-10469-3>.
- Komaty, Alain, Hatem Othmani, Shahreza, Anjith George, and Sébastien Marcel. 2025. “Exploring ChatGPT for Face Presentation Attack Detection in Zero and Few-Shot in-Context Learning.” arXiv: 2501.08799v1. Accessed February 26, 2026. <https://arxiv.org/pdf/2501.08799>.
- Layton, Seth, Tyler Tucker, Daniel Olszewski, Kevin Warren, Kevin Butler, and Patrick Traynor. “SoK: The Good, The Bad, and The Unbalanced: Measuring Structural Limitations of Deepfake Media Datasets.” *Proceedings of the 33rd USENIX Security Symposium* (2024): 1027–44. <https://www.usenix.org/system/files/usenixsecurity24-layton.pdf>.
- Le, Binh M., Jiwon Kim, Simon S. Woo, Kristen Moore, Alsharif Abuadba, and Shahroz Tariq. 2025. “SoK: Systematization and Benchmarking of Deepfake Detectors in a Unified Framework.” arXiv: 2401.04364v4. Accessed February 26, 2026. <https://arxiv.org/pdf/2401.04364>.
- Li, Xiaobai, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikäinen. “Generalized Face Anti-Spoofing by Detecting Pulse from Face Videos.” *Proceedings of the 2016 23rd International Conference on Pattern Recognition* (2016): 4239–44. <https://ieeexplore.ieee.org/document/7900300>.
- Li, Changjiang, Li Wang, Shouling Ji, Xuhong Zhang, Zhaohan Xi, Shanqing Guo, and Ting Wang. “Seeing is Living? Rethinking the Security of Facial Liveness Verification in the Deepfake Era.” *Proceedings of the 31st USENIX Security Symposium* (2022): 2673–90. <https://www.usenix.org/system/files/sec22-li-changjiang.pdf>.

- Lin, Li, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. 2025. "Detecting Multimedia Generated by Large AI Models: A Survey." arXiv: 2402.00045v7. Accessed February 26, 2026. <https://arxiv.org/pdf/2402.00045>.
- Liu, Honggu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. "Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain." *Proceedings of the 2021 IEEE/CVE Conference on Computer Vision and Pattern Recognition* (2021): 772–81. https://openaccess.thecvf.com/content/CVPR2021/papers/Liu_Spatial-Phase_Shallow_Learning_Rethinking_Face_Forgery_Detection_in_Frequency_Domain_CVPR_2021_paper.pdf.
- Luo, Xiyang, Yinxiao Li, Huiwen Chang, Ce Liu, Peyman Milanfar, and Feng Yang. "DV-Mark: A Deep Multiscale Framework for Video Watermarking." *IEEE Transactions on Image Processing* 34, (2023): 4371–85. <https://ieeexplore.ieee.org/document/10086041>.
- Narayan, Kartik, Vibashan VS, and Vishal M. Patel. 2026. "FaceXBench: Evaluating Multimodal LLMs on Face Understanding." arXiv: 2501.10360v3. Accessed February 26, 2026. <https://arxiv.org/pdf/2501.10360>.
- National Security Agency, Federal Bureau of Investigation, and Cybersecurity and Infrastructure Security Agency. 2023. "Contextualizing Deepfake Threats to Organizations." Department of Defense. Accessed February 26, 2026. <https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF>.
- Šalko, Milan, Anton Firc, and Kamil Malinka. "Security Implications of Deepfakes in Face Authentication." *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing* (2024): 1376–84. <https://dl.acm.org/doi/pdf/10.1145/3605098.3635953>.
- Song, Xiufeng, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. 2025. "On Learning Multi-Modal Forgery Representation for Diffusion Generated Video Detection." arXiv: 2410.23623v3. Accessed February 26, 2026. <https://arxiv.org/pdf/2410.23623>.
- Sony, Redwan, Parisa Farmanifard, Hamzeh Alzwaery, Nitish Shukla, and Arun Ross. 2025. "Benchmarking Foundation Models for Zero-Shot Biometric Tasks." arXiv: 2505.24214v1. Accessed February 26, 2026. <https://arxiv.org/pdf/2505.24214>.
- Tariq, Shahroz, David Nguyen, M.A.P. Chamikara, Tingmin Wu, Alsharif Abuadbba, and Kristen Moore. 2025. "LLMs Are Not Yet Ready for Deepfake Image Detection." arXiv: 2506.10474v1. Accessed February 26, 2026. <https://arxiv.org/pdf/2506.10474>.
- The 35th USENIX Security Symposium Organizers. 2025. "USENIX Security '26 Call for Papers." USENIX Association. Accessed February 26, 2026. <https://www.usenix.org/conference/usenixsecurity26/call-for-papers>.

- Wang, Zhendong, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. “Alt-Freezing for More General Video Face Forgery Detection.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023): 4129–38. https://openaccess.thecvf.com/content/CVPR2023/papers/Wang_AltFreezing_for_More_General_Video_Face_Forgery_Detection_CVPR_2023_paper.pdf.
- Wang, Zezheng, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, Guojun Qi, Jun Wan, and Zhen Lei. 2019. “Exploiting Temporal and Depth Information for Multi-Frame Face Anti-Spoofing.” arXiv: 1811.05118v3. Accessed February 26, 2026. <https://arxiv.org/pdf/1811.05118>.
- Xu, Zhenqi, Shan Li, and Weihong Deng. “Learning Temporal Features Using LSTM-CNN Architecture for Face Anti-Spoofing.” *Proceedings of 2015 3rd IAPR Asian Conference on Pattern Recognition* (2015): 141–5. <https://ieeexplore.ieee.org/document/7486482>.
- Yan, Zhiyuan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. “DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection.” *Proceedings of the 37th International Conference on Neural Information Processing Systems* 201 (2023): 4534–65. https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf.
- , Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, and Li Yuan. 2024. “Generalizing Deepfake Video Detection with Plug-and-Play: Video-Level Blending and Spatiotemporal Adapter Tuning.” arXiv: 2408.17065v2. Accessed February 26, 2026. <https://arxiv.org/pdf/2408.17065>.
- Zhang, Yulin, Jiangqun Ni, Wenkang Su, and Xin Liao. “A Novel Deep Video Watermarking Framework with Enhanced Robustness to H.264/AVC Compression.” *Proceedings of the 31st ACM International Conference on Multimedia* (2023): 8095–104. <https://dl.acm.org/doi/abs/10.1145/3581783.3612270>.
- Zheng, Yinglin, Jinmain Bao, Dong Chen, Ming Zeng, and Fang Wen. “Exploring Temporal Coherence for More General Video Face Forgery Detection.” *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021): 15044–54. https://openaccess.thecvf.com/content/ICCV2021/papers/Zheng_Exploring_Temporal_Coherence_for_More_General_Video_Face_Forgery_Detection_ICCV_2021_paper.pdf.
- Zou, Yueying, Peipei Li, Zekun Li, Huaibo Huang, Xing Cui, Xuannan Liu, Chenghanyu Zhang, and Ran He. 2025. “Survey on AI-Generated Media Detection: From Non-MLLM to MLLM.” arXiv: 2502.05240v2. Accessed February 26, 2026. <https://arxiv.org/pdf/2502.05240>.

SHAP の代替的手法の検討： 協力ゲーム理論を用いたアプローチ

ひらきかずひろ いしはらしんいち の じゅんのすけ
平木一浩／石原慎一／篠 潤之介

要 旨

説明可能な人工知能（XAI）の手法の1つである AFA は、機械学習モデルの予測値を特徴量の貢献度に分解し、各特徴量が予測に与える影響を可視化する手法である。特に、シャープレイ値（協力ゲーム理論の解概念の1つ）に基づく AFA である SHAP は、近年、金融・経済データへの適用が急速に進んでいる。本稿では、SHAP および Hiraki, Ishihara and Shino (2024) で提示された AFA を、導出過程を含めて平易に紹介するとともに、シャープレイ値以外の解概念を用いた AFA を新たに定式化する。そして、これらの AFA をわが国の長期金利および失業率に適用し、分解パターンの違いや計算コストの観点から比較分析する。分析の結果、(I) 協力ゲーム理論の解概念である残余均等配分解 (ES) やそれに類似した解概念 (ENSC) に基づく AFA については、SHAP との間で分解パターンに明確な差異がみられること、(II) それ以外の AFA の間の差異は相対的に小さいものの、差異のパターンは各 AFA に対応するカーネル関数の形状を反映していること、(III) ES と ENSC を按分した AFA は、計算コストや SHAP との差異が小さく、SHAP の近似計算の手法として優れた性質を持つこと、が明らかになった。

キーワード： AFA、LIME、SHAP、XAI、カーネル、機械学習、協力ゲーム理論
JEL classification: C45, C71, C63, G17

.....
本稿の作成に当たっては、大坪陽一氏（神戸大学）、近郷匠氏（福岡大学）、船木由喜彦氏（早稲田大学）、武藤滋夫氏（東京工業大学）ならびに日本銀行のスタッフ等から有益なコメントを頂いた。本稿に示されている意見は、筆者個人に属し、日本銀行、国際通貨基金、同理事会、同マネジメントの公式見解を示すものではない。また、ありうべき誤りはすべて筆者個人に属する。

平木一浩 国際通貨基金 (E-mail: khiraki@imf.org)
石原慎一 独立研究者 (E-mail: ishihara5683@gmail.com)
篠 潤之介 早稲田大学国際学術院 (E-mail: junnosuke.shino@waseda.jp)

(1) SHAP: 複雑な機械学習モデルを可視化する手法の近年の展開

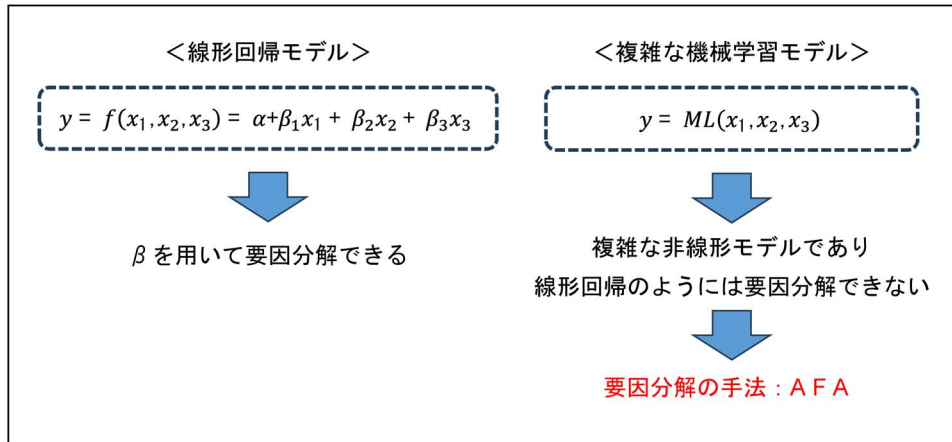
近年の人工知能（Artificial Intelligence, AI）や機械学習（Machine Learning, ML）分野における理論的・技術的手法の急速な発展を背景に、金融・ファイナンスや経済分野におけるデータ分析の手法にも、大きな変革がもたらされている。特に、高頻度・高粒度のデータやテキストデータといったいわゆるビッグ・データが分析対象として一般的になるもとの、機械学習モデルが従来のモデルを上回る予測力を示すことも少なくなく、資産価格の推定や信用リスク評価、マクロ経済指標の予測など幅広い分野で利用が進んでいる。複雑で非線形な関係を捉えられる点は機械学習モデルの大きな強みであり、ファイナンスの実務や経済分析において欠かせない手法となりつつある。

しかし、機械学習モデルは内部構造が複雑で理解が難しく、いわゆる「ブラックボックス」であるとの批判も根強い。経済・金融の分野では、モデルが出力する結果の根拠や判断過程を示すことが強く要請されるため、単に予測精度が高いだけでは十分とはいえない。モデルの予測や出力結果がどの要因によってどの程度もたらされているのかを明らかにすることは、実務上も倫理的にも不可欠である。このような課題を背景に近年急速に研究の進展がみられているのが、「説明可能な AI」（eXplainable AI, XAI）であり、特に本分析が焦点を当てる SHAP（SHapley Additive exPlanations）は、協力ゲーム理論における解概念であるシャープレイ値に基づいて、どんな機械学習モデルに対しても特徴量の寄与度を一貫して算出できる手法として評価され、幅広い分野で用いられている。

具体的には、XAI の枠組みに含まれる AFA（Additive Feature Attribution）とは、複雑な機械学習モデルの予測値を各特徴量の影響度に分解することで、個々の要因が予測に与える効果を定量化し可視化するアプローチである。たとえば、3 つの変数 $X \cdot Y \cdot Z$ を用いて資産価格を予測する場合、AFA は予測値のどの程度が X に由来するものなのか、どの程度が Y に由来するものなのか、どの程度が Z に由来するものなのかを分解・可視化する。線形回帰モデルであれば、こうした要因分解は推計されたパラメータを用いて実行することができる。ところが、ニューラルネットワークやアンサンブルツリーのように複雑な構造をもつモデルでは、そのような回帰係数による説明は実行不可能であるため、AFA の活用が重要となる（図 1）。

AFA の具体的な手法である SHAP は、協力ゲーム理論の解概念であるシャープレイ値（Shapley 1953）に基づく AFA であり、Lundberg and Lee（2017）（以下「LL 論

図 1 回帰分析における要因分解と機械学習モデルにおける AFA



文」と呼ぶ)によって定式化されて以降、近年、機械学習や AI の分野において、急速に分析・研究が進められている¹。例えば、計算コストの観点からは²、SHAP の計算速度を短縮化するための TreeSHAP (Lundberg *et al.* 2020) や FastSHAP (Jethani *et al.* 2022) などの手法が開発されている。実際のデータを用いた SHAP の適用としては、医療やヘルスケアの分野を中心として、様々な分野で分析が蓄積されてきた³。

さらに、ここ数年の間で、経済・ファイナンス・関連のデータに対して、SHAP をはじめとする XAI の手法を用いて機械学習モデルを解釈可能な形で適用・分析する研究が展開されつつある。Jabeur *et al.* (2024) は、商品価格（金価格）を 6 つの機械学習モデルを用いて予測した後、各予測値について SHAP を適用して比較分析を行い、XGBoost モデルとそれに対する SHAP の適用が分析上有効であることを主張した。Demirbaga and Xu (2024) は、株価リターンの将来予測モデルに機械学習モデルを構築したうえで、そこから得られた予測値に対して SHAP をはじめとする AFA の手法を適用し、予測値を各特徴量の貢献度に分解した。また、SHAP を用いた複数の可視化の手法 (SHAP summary plot, SHAP bar plot, SHAP force plot) を用いて、資産価格モデルがブラックボックス化することへの一連の対処方法を

.....
1 Lundberg and Lee (2017) (LL 論文) は、2017 年の NeurIPS (Conference on Neural Information Processing Systems、機械学習や人工知能 (AI) 分野でもっとも権威のある国際会議のひとつ) に掲載されたプロシーディングであるが、同論文の引用件数は 2025 年 9 月時点で 41,000 件を超えており、XAI におけるもっとも基礎的な文献の 1 つとなっている。

2 後述するように、SHAP は理論的に多くの優れた側面を有する一方で、大規模なデータに適用する際には、計算コストの大きさがしばしば問題となる。

3 医療分野における XAI の活用状況を体系的にレビューし、特に SHAP や LIME (後述) などの手法が診断支援や疾患予測モデルの解釈性向上に貢献していることを示した Loh *et al.* (2022) や、全身麻酔中の際の低酸素血症の予測に SHAP を適用した Lundberg *et al.* (2018) などが挙げられる。

示した。Misheva *et al.* (2021) は、信用リスクの要因分解に対し、また、Bussmann *et al.* (2021) は、融資を行うかどうかの意思決定を説明するためのリスク管理モデルに対し、XAI の手法を適用して特徴量と予測結果との関係性の可視化を行った。Ariza-Garzón *et al.* (2020) は、P2P レンディングにおけるスコアリング付与のための複数のモデルを評価する際に、SHAP の適用に基づく考察を行った。このように、SHAP をはじめとする XAI の手法は、複雑な機械学習モデルから透明性の高い解釈を提供することで、経済・ファイナンス分析において機械学習モデルを適用することの妥当性を大幅に高めたといえる。金融・ファイナンス分野における近年の XAI の研究動向については Wei *et al.* (2025) が詳しい⁴。

なお、XAI あるいは SHAP を用いた経済・ファイナンス関連データの実証分析は、中央銀行においても近年活発に進められている。英国中央銀行 (BOE) のワーキングペーパーとして公表され、その後 *International Journal of Central Banking* 誌に掲載された Buckmann and Joseph (2023) は、米国の失業率を対象に、機械学習モデルの予測精度の比較評価、SHAP による予測値の要因分解、変数間の非線形関係の可視化、SHAP の統計的な検証 (Shapley Regression) などの、SHAP をコアとする機械学習モデルを用いた分析フローを提唱し、政策当局の実体経済分析や情勢判断において、AFA ないし SHAP を活用することの有効性を示した。同じく BOE のワーキングペーパーとして公表された Bracke *et al.* (2019) は、XAI の 1 つである QII という手法を、機械学習モデルによる住宅ローンのデフォルト予測に適用し、LTV 比率や金利などが主要な決定要因であることを特定した。欧州中央銀行 (ECB) のワーキングペーパーとして公表され、その後 *Journal of International Economics* 誌に掲載された Bluwstein *et al.* (2023) は、SHAP を用いて金融危機の予測に有用な金融経済変数を特定し、可視化を行った。

わが国においても、SHAP を経済・ファイナンス関連のデータに適用する分析が進んでいる。金田ほか [2022] は、原油価格を題材に、機械学習モデルの推計とモデル説明のための SHAP を用いた分析ワークフローを示したうえで、その有用性と実務上留意すべき事項について考察している。鷲見 [2020] は、SHAP を用いて通貨オプション市場における投資家センチメントを分析し、その主要な変動要因が金融ストレス指数や米国イールドカーブであることを可視化した。森ほか [2021] も、125 か国の新規感染者数および 36 種類の特徴量からなるパネルデータに対して機械学習モデルとしてランダムフォレストモデルを適用した後、SHAP を用いて特徴量の重要度を計測した。

.....
4 XAI や SHAP を平易に解説した文献として、Molnar (2025) のほか、日本語では大坪ほか [2021] や森下 [2021] がある。和泉 [2025] は金融分野の因果 AI における XAI の位置づけについて論じている。

(2) 本分析の概要

本稿では、SHAP およびその代替的な手法を対象に、理論面および実証面の分析を行う。

まず、理論分析では、Hiraki, Ishihara and Shino (2024) (以下「HIS 論文」と呼ぶ) をベースに、それを発展させる形で、既存の AFA の代表的な手法である SHAP と、その代替的な手法について検討を行う。具体的には、まず、SHAP が協力ゲーム理論における解概念であるシャープレイ値に基づく AFA である点に着目し、協力ゲーム理論におけるシャープレイ値以外の解概念 (残余均等配分解やそれに関連する解概念) をベースにした AFA を提示し、それらの違いを考察する。次に、LL 論文をベースに、AFA の基本的な別の手法である、LIME (Local Interpretable Model-agnostic Explanations, Ribeiro *et al.* (2016) およびそのカーネル関数との関係性に着目して代替的な AFA を提示し、それらの違いを考察する。

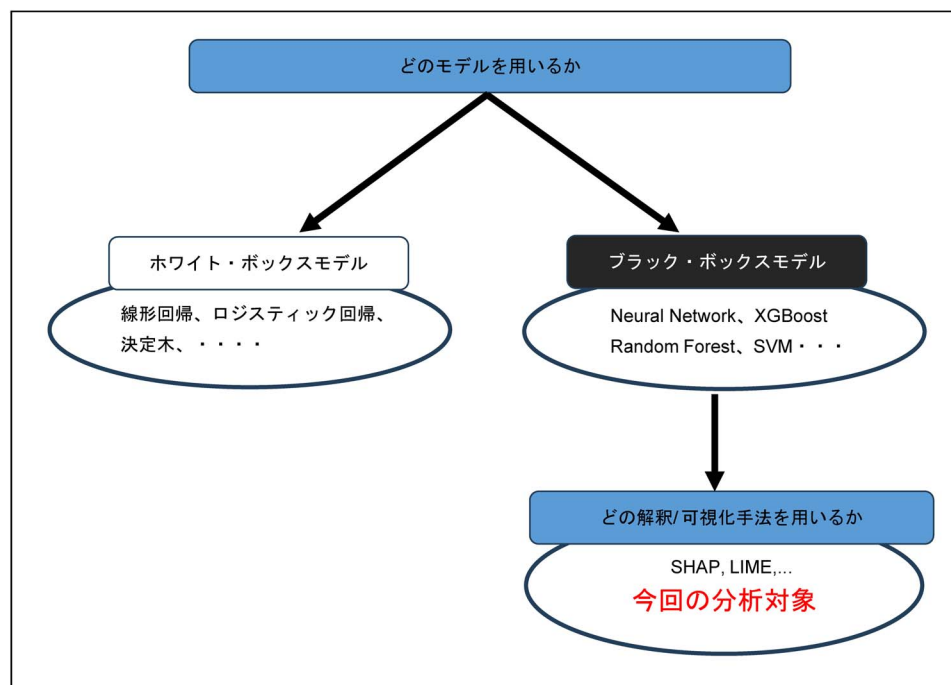
次に、実証分析では、理論分析で取り上げた SHAP およびその代替的な手法を、実際の経済・ファイナンスデータに適用し、比較分析を行う。具体的には、まず、わが国の長期金利 (10 年債利回り) の推移について、日本銀行 [2021] などの分析に基づき特徴量を選択した上で、機械学習モデルとして XGBoost を採用し、当該モデルを学習させる。そして、学習済の XGBoost モデルに対して理論分析パートで示した複数の AFA を適用し、分解パターンの違いや計算コストの大きさの観点から比較分析を行う。次に、米国失業率を対象にした Buckmann and Joseph (2023) を参考に、わが国失業率について、同様のフレームワークで分析を行う⁵。そして、複数のケースから得られた共通の知見を抽出し、SHAP および複数の代替的手法を暫定的に評価する。

実証分析の結果明らかになった点を簡潔に先取りすると、以下の通りである。1 点目に、協力ゲーム理論の解概念である残余均等配分解 (ES) に基づく AFA や、それに類似した解概念 (ENSC) に基づく AFA については、SHAP との間で分解パターンの違いが明確にみられた。2 点目に、それ以外の AFA の間では、視覚的に明確に確認できるほどの大きな差異はみられないものの、AFA 間の差異の大きさは、各 AFA に対応するカーネル関数の形状を反映したものとなっていることが確認された。3 点目に、ES に基づく AFA と、ENSC に基づく AFA を按分して定義される AFA は、計算コストが小さく、かつ SHAP との差異が極めて小さいなど、SHAP を近似計算する手法として優れた性質を持つことが明らかになった。

なお、以上の通り、本稿の実証分析パートは、(I) 特定のデータに対して機械学習モデルを構築し、(II) 構築した学習済のモデルに対して SHAP およびその代替的

5 さらに、補論 3 では、Jabeur *et al.* (2024) を参考に、金価格に対しても複数の AFA を用いた比較分析を行う。

図2 機械学習モデル分析における本稿の着目点



手法を適用する、という2のステップからなるが、本稿の主な分析対象は(II)である(図2)。すなわち、機械学習モデル自体の構築や異なる機械学習モデル間の予測精度の比較といった論点は(当然それ自体重要ではあるが)(I)に含まれ、本稿の分析対象ではない点に留意されたい(換言すると、本分析では機械学習モデルやハイパーパラメータは所与として扱う)。経済・ファイナンスデータを用いた機械学習モデルの構築とその予測精度については、AFAやSHAPの実データの適用についての分析に比べ、相対的に研究の蓄積が進んでいる。例えば、金融データについて、Gu *et al.* (2020)は、株価リターンに複数の機械学習モデルを適用し、これらのモデルによる予測精度が、従来の回帰ベースの予測を大きくアウトパフォームすることを示した。Chen *et al.* (2024)は、ニューラルネットワークを用いて個別銘柄の株価リターン予測を行い、全てのベンチマークモデルより優れた結果がもたらされることを示した。債券価格についてはBianchi *et al.* (2021)、オプション・リターンについてはBali *et al.* (2023)が、それぞれ機械学習モデルを適用し、従来のモデルより優れたパフォーマンスがもたらされることを示した。一方、実体経済関連への適用について、例えば、Coulombe *et al.* (2022)は、米国の5つのマクロ経済指標(鉱工業生産、失業率、消費者物価指数、長短金利差、住宅着工件数)を対象に機械学習モデルを用いた予測のパフォーマンスについて考察し、(I) 機械学習モデ

ルの最大の有意性はその非線形性にあること（例えば、マクロ経済的な不確実性が高い時期〈例：金融危機、住宅バブル崩壊〉に特に優位性が高まる）、(II) 時系列データにおいても K-fold クロスバリデーションがハイパーパラメータの選択法として有効であること、(III) 次元削減に主成分分析を用い、その後機械学習モデルで非線形な関係を学習するのが有効であること、といった点を明らかにしている。

本論文の次節以降の構成は以下の通りである。2 節は SHAP およびその代替的手法についての理論的な導出および詳細な比較分析である。3 節は SHAP およびその代替的手法の金融・経済データへの適用である。4 節はまとめと結論である。

2. 理論分析：SHAP とその代替的手法についての導出と考察

(1) LL 論文における SHAP と HIS 論文におけるその代替的手法に関する議論の全体感

SHAP とその代替的手法についての理論的考察を行ううえでは、まず、SHAP を提示した LL 論文の議論の全体感を把握することが極めて有用である。LL 論文では、SHAP を 2 つの観点から特徴づけている（図 3）。1 つめの観点は、協力ゲームの解概念であるシャープレイ値を、AFA の文脈に適用したものとして SHAP を特徴づけるものである（同論文の Theorem 1）。もう 1 つの観点は、AFA の別の基本的な手法の 1 つである、LIME の具体的な定式化として SHAP を特徴づけるものである。特に、LIME を定式化する際に必要となるカーネル関数 π を特定の形に限定することで（図 3 における $\pi^{SHAP}(S)$ ）。詳細は 2 節 (4) で解説）、それが SHAP と一致することを示している（同論文の Theorem 2）。

一方、HIS 論文では、この LL 論文の分析のフレームワークに依拠しつつ、それぞれの観点から SHAP の代替的手法を提示している（図 4）。まず、1 つめの「協力ゲームの解概念としての SHAP」について、HIS 論文では、協力ゲーム理論における解概念はシャープレイ値の他にも多くのものがあり、それらを同様に AFA として定式化し、それを SHAP と比較することの重要性を指摘した。そのうえで、協力ゲーム理論における解概念である残余均等配分解に基づく AFA と、LS（最小二乗）プレ仁に基づく AFA を定式化して、SHAP と代替的な手法として提示した（図 4 における Approach 1）。次に、2 つめの「LIME において特定のカーネル関数を仮定することで導出される SHAP」について、HIS 論文では、この SHAP に対応するカーネル関数が、後述する「カーネル関数が本来満たすべき性質」を満たしていない可能性を指摘した。そのうえで、この性質を満たすカーネル関数を定義して、そこか

図3 LL論文におけるSHAPの特徴づけ

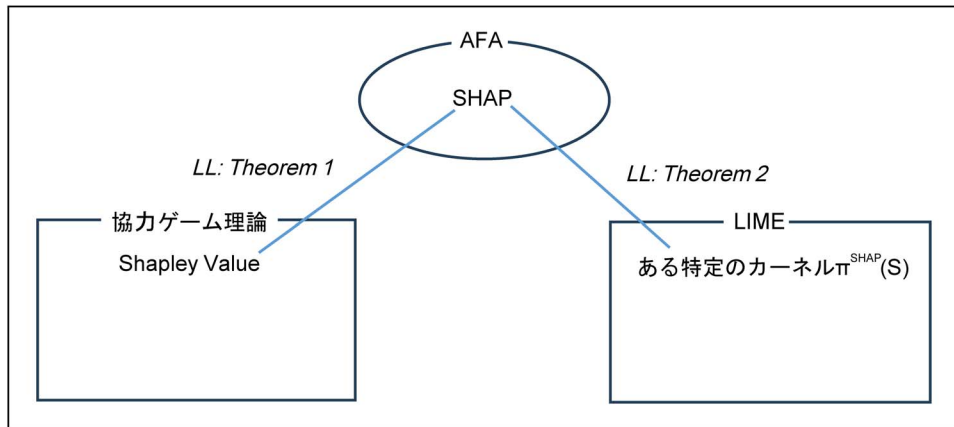
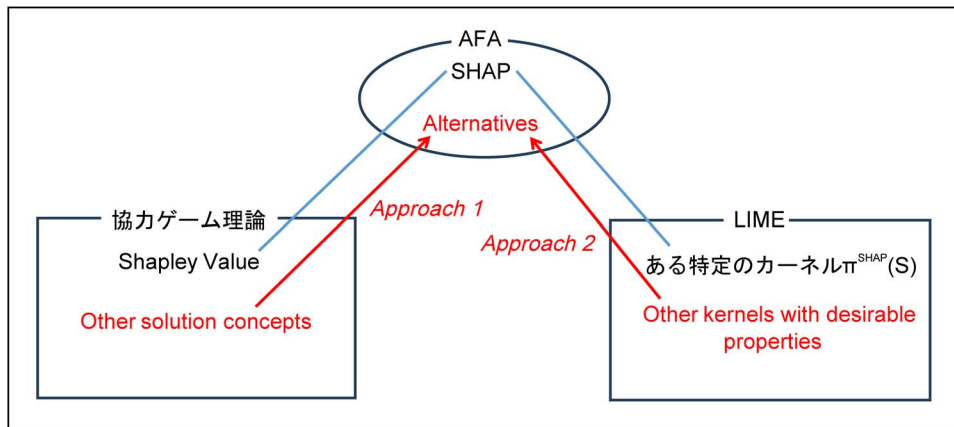


図4 HIS論文における代替的手法の提示



ら SHAP と代替的な AFA を提示した（図4における Approach 2）。

以上が LL 論文と HIS 論文の全体感である。2 節 (2)以降でこれらの AFA に関してのより詳細なレビューを行う前に、以下では、LL 論文における SHAP と、HIS 論文で提示された様々な代替的手法のうちもっともシンプルな、協力ゲーム理論の解概念である残余均等配分解を用いた AFA（以下では、残余均等配分（Equal Surplus）の頭文字をとって、SHAP に対して ES と呼ぶことにする）について、具体例を用いながらそのイメージを把握することにする。

SHAP と ES の違いを直感的に理解するために、ここでは簡単な例を用いる。学習済の機械学習モデルを f 、特徴量は A, B, C の 3 つであるとする。例えば、株価リターンを予測するとして、機械学習モデル f はランダムフォレストやニューラルネットワークといった複雑な「ブラックボックス」モデル、特徴量は計量分析における

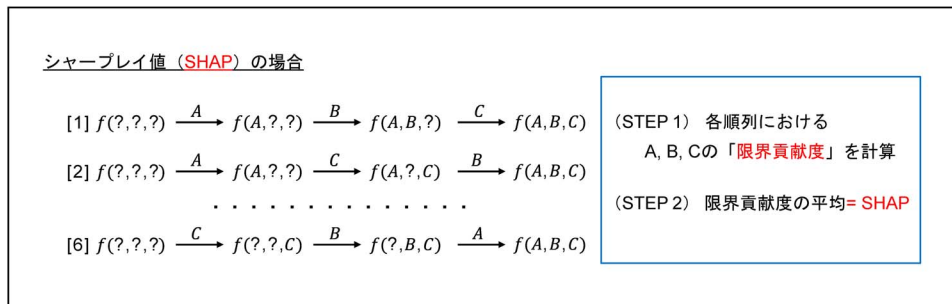
「独立変数」「説明変数」のことであり、企業収益、配当性向、為替レート、株価自身の過去のリターンやボラティリティといった変数が特徴量の候補となる。

AFA が解くべき問題は、ある観測値において、特徴量 A, B, C の値が全て既知である場合の予測値 $f(A, B, C)$ と、特徴量 A, B, C がすべて未知である場合の予測値 – これを $f(0)$ とする – の差、すなわち $f(A, B, C) - f(0)$ 、を、特徴量 A, B, C に配分する手法のことである。 $f(A, B, C) - f(0)$ は、すべての特徴量が既知となったときの「予測の改善度」とみなすことができる。これを A, B, C の「予測の貢献度」に応じて配分する手法が AFA である⁶。協力ゲーム理論においては $f(A, B, C) - f(0)$ はいわゆるプレイヤー間で配分される「パイの大きさ」にあたり、その配分方法として、シャープレイ値をはじめとする様々な解概念が提示されてきた。

それでは、まず SHAP について見てみよう (図 5)。SHAP では、まず、全ての特徴量が未知の状態である $f(0)$ (図 5 では $f(?, ?, ?)$ と表示) からスタートし、1 つずつ特徴量が加わっていく状況を想定する。図中で [1] とあるケースでは、まず特徴量 A が既知になる。ここで A の予測の「限界貢献度 (marginal contribution)」を $f(A) - f(0)$ とする ($f(A)$ は図中では $f(A, ?, ?)$ と表示)。次に、特徴量 B が既知となる。ここで B の限界貢献度は、 B が既知となる前の予測値 $f(A)$ と、 B が既知となったときの予測値 $f(A, B)$ の変化幅 $f(A, B) - f(A)$ とする。最後に、特徴量 C が既知となったときの C の限界貢献度を $f(A, B, C) - f(A, B)$ とする。

以上でケース [1] における特徴量 A, B, C の限界貢献度が算出できた。ケース [1] は順列 $A \rightarrow B \rightarrow C$ に対応するので、1 つずつ特徴量が加わっていくケースの総数は A, B, C の順列の場合の数、 $3! = 6$ 通りである (図 5 では [1], [2], ..., [6] と表示されている)。そして、それぞれのケースにおいて A, B, C それぞれの限界貢献度を計算し、それをケースの総数 6 で割った、いわば「限界貢献度の平均」が SHAP

図 5 SHAP (シャープレイ値を用いた AFA) の計算方法



6 データ全体に対する予測値と特徴量の平均的な関係ではなく、特定の観測値における予測値と特徴量の関係に注目することから、SHAP (および本稿で扱う AFA) を局所的 (ローカルな) 手法と呼ぶ。ただし、個々の観測値における SHAP を集計することで、グローバルな手法として用いることも可能である。グローバルな XAI の手法には、Permutation Importance や Partial Dependence Plot がある。

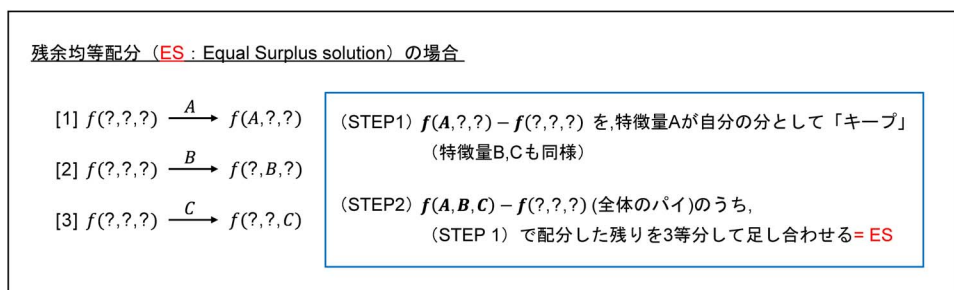
となる⁷。LL 論文ではこのシャープレイ値の考えに則って SHAP が定義された。

次に、残余均等配分解 (ES) についてみてみよう (図 6)。ES でも、まず、全ての特徴量が未知の状態である $f(\emptyset)$ からスタートする。しかし、3 つの特徴量全てが加わる順列を考えるのではなく、考慮するのは $f(\emptyset)$ から 1 つめの特徴量が加わる状態のみである。図 6 で [1] とあるケースにおいては、すべての特徴量が未知である場合の予測値 $f(\emptyset)$ から、特徴量 A が既知になったときの予測値 $f(A)$ の変化幅 $f(A) - f(\emptyset)$ を、SHAP 同様に A の予測の「限界貢献度 (marginal contribution)」とする。ES では、これを特徴量 A が自分の取り分としてまず「キープ」すると考える。同様に、 B は $f(B) - f(\emptyset)$ 、 C は $f(C) - f(\emptyset)$ を自分の取り分として「キープ」する。この「各特徴量が自分の取り分をキープする」という段階が 1 番目のステップにあたる。そして、2 番目のステップにおいて、「全体のパイの大きさ」である $f(A, B, C) - f(\emptyset)$ から、1 番目のステップにおいて各特徴量がキープした分を差し引いた「残余」を、3 つ特徴量で均等配分し、1 番目のステップでキープした分に加える。これが ES に基づく AFA である。

SHAP と ES を比べると、当然、SHAP の方がより多くの情報を用いて計算されている。すなわち、上記の例において、SHAP は 2 つの特徴量が既知の場合における予測値である $f(A, B)$ 、 $f(A, C)$ および $f(B, C)$ の情報を考慮して算出されるが、ES はこの情報を考慮していない。したがって、「様々なケースにおける各特徴量の予測の貢献度に応じて配分する」AFA として、SHAP は ES に比べると「フェア」な方法であると言える。また、特徴量の数が大きくなるほど、「SHAP では考慮しているが ES では考慮していない情報量」は大きくなることから、両者の違いは大きくなっていくと考えられる。

一方で、ES にはメリットもある。情報量についての議論といわばコインの裏表の関係であるが、SHAP の計算においては、 n 個の特徴量があるとき、計算すべき予測値の数は 1 観測値あたり 2^n 個になることから、特徴量の数が増えると SHAP

図 6 ES (残余均等配分解を用いた AFA) の計算方法



.....
7 この計算によって得られた 3 つの特徴量の SHAP の合計値が、ちょうど「パイの大きさ」である $f(A, B, C) - f(\emptyset)$ と一致することは容易に確かめられる。

の計算コストは指数関数的に増大していく。一方で、ES においては、計算すべき予測値の数は 1 観測値あたり $n + 2$ 個であることから、特徴量の数が増えた場合の SHAP の計算コストの増加ペースは線形なものにとどまる。したがって、両者の計算コストの差は、特徴量の数が大きくなるほど拡大していく。機械学習においては、非常に多くの特徴量を扱うケースが一般的であることから、計算コストの小ささは ES の大きなメリットであると言える。1 節でも言及したが、SHAP は AFA として様々なメリットを持つ一方で、計算コストの大きさが最大の問題とみなされてきた。このため、相対的に少ない計算コストで近似的に SHAP を計算する、TreeSHAP や FastSHAP などの手法が提案されてきたが、これらの手法も適用可能な学習モデルが制限されていたり、なお計算コストが比較的大きいといった点が指摘されている (Molnar 2025、森下 [2021] などを参照)。

SHAP と ES についての直感的な議論は以上である。HIS 論文において提示された他の代替的な手法や、さらなる新たな手法について考察するために、2 節 (2) 以降では、まず数学的な準備を行ってから、SHAP とその代替的手法についての考察をより掘り下げていく。

(2) 準備

観測値を t 個、特徴量の数を n 個とし ($N = \{1, \dots, n\}$ および $T = \{1, \dots, t\}$)、特徴量のベクトルを $t \times n$ 次元ベクトル $X = (X_1, \dots, X_j, \dots, X_n)$ とする。 j 番目の特徴量のベクトルを $X_j = (x_{1,j}, \dots, x_{t,j})'$ 、また、 τ 番目の観測値の特徴量のベクトルを $x_\tau = (x_{\tau,1}, \dots, x_{\tau,j}, \dots, x_{\tau,n})$ とする。 f を学習済モデルとし、 $Y = (y_1, \dots, y_t)'$ を f による予測値とする ($Y = f(X)$)。

N のべき集合の要素 $S \in 2^N$ (協力ゲーム理論では提携と呼ぶ) に対し、 $x_{\tau,S} = \{x_{\tau,j} | j \in S\}$ とする。すなわち、 $x_{\tau,S}$ は、 τ 番目の観測値における、 S に含まれる特徴量からなるベクトルである。また、 $X_S = \{X_j | j \in S\}$ とする。 S の要素数を $|S|$ とする。

協力ゲーム理論において、特性関数形ゲームは (N, v) で表現される。 $N = \{1, \dots, n\}$ をプレイヤーの集合と呼ぶ。 v はべき集合 2^N 上の実数値関数である。本来の協力ゲーム理論が主に扱うのは、「全体提携 N が形成されるという前提のもとで、(1 人提携を含む) 各提携が自力で獲得できる値 $v(S)$ の情報を踏まえながら、 $v(N)$ をどのようにプレイヤー間に配分するか」という問題である。

今、 τ 番目の観測値において、特徴量の集合 N をプレイヤー集合とする特性関数形ゲームを作ることを考える。 τ 番目の観測値において、提携 S に対して実数値関数 $v_\tau : 2^N \rightarrow R$ を以下の (1) 式で定義すると、 τ についての特性関数形ゲームが 1

つ定まる：

$$v_{\tau}(S) = E[f(x_{\tau,S}, X_{N \setminus S})]. \quad (1)$$

(1) 式において、 $v_{\tau}(S)$ は、「 τ 番目の観測値において、 S に含まれる特徴量 $x_{\tau,j}(j \in S)$ は既知だが、 S に含まれない特徴量 $x_{\tau,k}(k \in N \setminus S)$ は未知であるときの、学習済モデル f が予想する予測値」である。ここで、 $E[f(x_{\tau,S}, X_{N \setminus S})] = (1/t) \sum_{t' \in T} f(x_{\tau,S}, x_{t', N \setminus S})$ である（以下の例 1 も参照）。すなわち、 $v_{\tau}(S)$ を求める際には、 S に含まれない特徴量は、各観測値におけるそのような特徴量の組み合わせが等確率で生じると仮定して、すべての観測値についての期待値をとる。一方、 S に含まれる特徴量は既知と仮定しているので一定とする。

$v_{\tau}(N) = E[f(x_{\tau,1}, \dots, x_{\tau,n})] = f(x_{\tau,1}, \dots, x_{\tau,n})$ かつ $v_{\tau}(\emptyset) = E[f(X_1, \dots, X_n)] = E[f(X)]$ であり、前者は τ 番目の観測値において特徴量が全て既知である場合の理論値、後者は τ 番目の観測値において特徴量がすべて未知である場合の理論値である。協力ゲーム理論の分析においては、（一般性を失わないという前提のもとで） $v(\emptyset) = 0$ を仮定することが多い。一方、上記の定式化（すなわち AFA の文脈）においては $v_{\tau}(\emptyset)$ は一般的に非ゼロであるため、注意が必要である。

例 1 観測値 4 個 ($t = 4$)、特徴量 3 個 ($n = 3$) の場合を考える。

$$X = \begin{pmatrix} x_{11}, x_{12}, x_{13} \\ x_{21}, x_{22}, x_{23} \\ x_{31}, x_{32}, x_{33} \\ x_{41}, x_{42}, x_{43} \end{pmatrix}. \quad (2)$$

4 番目の観測値 $\tau = 4$ に対応する特性関数形ゲーム v_4 は、以下の通り定まる：

$$\begin{aligned} v_4(\emptyset) &= \frac{1}{4} \sum_{i=1}^4 f(x_i) \quad \text{ただし } x_i = (x_{i1}, x_{i2}, x_{i3}) \\ v_4(1) &= \frac{1}{4} \{f(x_{41}, x_{12}, x_{13}) + f(x_{41}, x_{22}, x_{23}) + f(x_{41}, x_{32}, x_{33}) + f(x_{41}, x_{42}, x_{43})\} \\ &\dots \dots \dots v_4(2), v_4(3) \text{ も同様} \dots \dots \dots \\ v_4(12) &= \frac{1}{4} \{f(x_{41}, x_{42}, x_{13}) + f(x_{41}, x_{42}, x_{23}) + f(x_{41}, x_{42}, x_{33}) + f(x_{41}, x_{42}, x_{43})\} \\ &\dots \dots \dots v_4(13), v_4(23) \text{ も同様} \dots \dots \dots \\ v_4(123) &= f(x_{41}, x_{42}, x_{43}) \end{aligned}$$

$v_4(\emptyset)$ は、「4 番目の観測値において、すべての特徴量が未知である場合の理論値」なので、 x_1, x_2, x_3, x_4 が等確率で発生すると仮定し、期待値を計算する。 $v_4(1)$ は、「4 番目の観測値において、1 番目の特徴量のみが既知である場合の理論値」なの

で、 x_{41} は固定し、 (x_{12}, x_{13}) , (x_{22}, x_{23}) , (x_{32}, x_{33}) , (x_{42}, x_{43}) , が等確率で発生すると仮定し、理論値の期待値を計算する。 $v_4(12)$ も同様である。最後に、 $v_4(123)$ は、「4 番目の観測値においてすべて特徴量が既知である場合の理論値」なので、学習済モデル f に $x_4 = (x_{41}, x_{42}, x_{43})$ を代入するだけである⁸。

機械学習における AFA とは、 τ 番目の観測値に着目し、「すべての特徴量が既知である場合の予測値と、すべての特徴量が未知である場合の予測値の差」である $v_\tau(N) - v_\tau(\emptyset)$ を、各特徴量の貢献度（寄与度）に応じて配分する手法である。例えば、目的変数 y を株価リターン、3 つの特徴量を鉱工業生産、為替レート、インフレ率とする。ある特定の観測値（例えば $\tau = 2024$ ）において、鉱工業生産と為替レートが株価リターンを説明する主要因である一方、インフレ率が株価リターンに与える影響は小さかったとする。このとき、 $v_\tau(N) - v_\tau(\emptyset)$ の要因分解（あるいは寄与度分解）において、鉱工業生産と為替レートに大きな値を与え、インフレ率には小さな値を与える、というのが AFA の基本的な考え方である⁹。学習モデル f が線形回帰などの単純なモデルであれば、学習済（すなわち推計された）回帰パラメータを用いて要因分解を行うことが可能である。一方で、 f が複雑な機械学習モデルである場合、パラメータを用いた要因分解は不可能であることから、特定の手法を用いた貢献度の可視化が必要となる。この具体的な手法の 1 つが AFA である。

具体的には、 τ 番目の観測値に関する特性関数形ゲーム (N, v_τ) とそこでの特徴量 $j \in N$ に対し、実数値関数 $\Psi_\tau(j) : N \rightarrow \mathbb{R}$ を考える（以後、既存研究の表記法に沿って、 $\Psi_\tau(j)$ を $\Psi_{\tau,j}$ と表記する）。また、 $\Psi_\tau = (\Psi_{\tau,1}, \dots, \Psi_{\tau,n})$ とする。 Ψ_τ が以下の (3) 式を満たすとき、 Ψ_τ を AFA と呼ぶ：

.....
8 正確には、以上の計算においては、Feature Independence (FI) の仮定が用いられている (Buckmann and Joseph 2023, Lundberg and Lee 2017)。ある特徴量 X の実現値 x を所与としたときの、特徴量全体の条件付き分布は x に依存しない、すなわちこの例においては $E[f(x_1, x_2, x_3) | x_1] = E[f(x_1, x_2, x_3)]$ というのがその内容である。つまり FI は特徴量の間の相関を仮定しない。極端な例として、 $v_4(1)$ の計算において、 X_1 が (X_2, X_3) と完全相関しているとする。このとき、 (x_{41}, x_{12}, x_{13}) , (x_{41}, x_{22}, x_{23}) , あるいは (x_{41}, x_{32}, x_{33}) の組み合わせが生じることはありえないので、この相関関係が事前に分かっているのであれば、 $v_4(1)$ は、本文中で示した形ではなく、 $v_4(1) = f(x_{41}, x_{42}, x_{43})$ とするのが妥当であると言える。

もっとも、FI の仮定は SHAP の既存研究全般で一般的に用いられている仮定の 1 つであり、本分析では FI の妥当性についてこれ以上掘り下げることはしない。FI を支持する立場の論文としては Janzing *et al.* (2020) がある。Chen *et al.* (2020) は、FI を仮定すべきかどうかは、分析の目的（データの自体の構造や振る舞いを知りたいのか、その背後にある因果関係を知りたいのか）に拠るとしている。一方、FI の仮定を外すと、Missingness という AFA が満たすべき性質（協力ゲーム理論におけるいわゆる null player property に該当する性質）を SHAP が満たさないことが知られている (Molnar 2025)。

9 ここでは便宜上時系列データを例に挙げたが、もちろん、AFA はより広いクラスのデータに適用可能な手法である。

$$\sum_{j \in N} \Psi_{\tau,j} = v_{\tau}(N) - v_{\tau}(\emptyset). \quad (3)$$

AFA とは、「観測値における特徴量がすべて既知である場合の予測値」と「観測値における特徴量がすべて未知である場合の予測値」の差を、各特徴量に過不足なく配分したものである。以降、 Ψ_{τ} が AFA であるとき、 Ψ_{τ}^{AFA} とも表記する。

(3) 協力ゲーム理論の解概念に基づいた AFA: SHAP およびその代替的手法

以上の準備のもと、2 節 (1) で取り上げた SHAP と ES は、それぞれ以下の (4) 式および (5) 式によって定義される：

$$\Psi_{\tau,j}^{SHAP} = \sum_{S \subseteq N \setminus j} \frac{|S|!(n - |S| - 1)!}{n!} (v_{\tau}(S \cup \{j\}) - v_{\tau}(S)) \quad (4)$$

$$\Psi_{\tau,j}^{ES} = v_{\tau}(\{j\}) - v_{\tau}(\emptyset) + \frac{(v_{\tau}(N) - v_{\tau}(\emptyset)) - \sum_{i \in N} \{v_{\tau}(\{i\}) - v_{\tau}(\emptyset)\}}{n}. \quad (5)$$

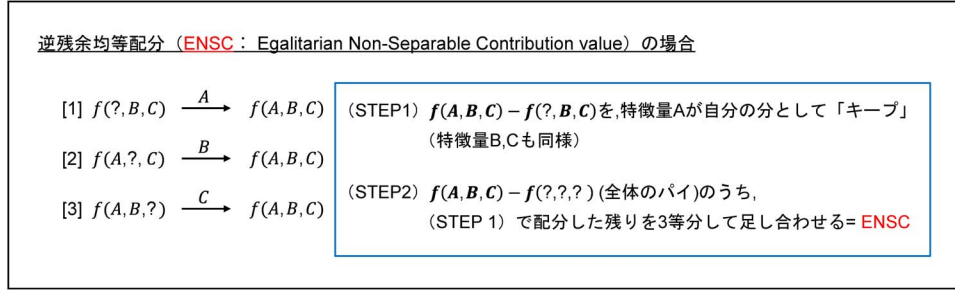
$\Psi_{\tau,j}^{SHAP}$ は、2 節 (2) で定義した特性関数形ゲーム (N, v_{τ}) における、協力ゲーム理論における代表的な解概念であるシャープレイ値 (Shapley 1953) そのものである。また、 $\Psi_{\tau,j}^{ES}$ も、協力ゲーム理論における別の代表的解概念である残余均等配分解を (N, v_{τ}) において定めたものである。(5) 式の $v_{\tau}(\{j\}) - v_{\tau}(\emptyset)$ の部分が図 6 における STEP 1、そのあとの n を分母とした分数部分が STEP 2 に対応している。

$\Psi_{\tau,j}^{ES}$ は HIS 論文および Condevaux *et al.* (2022) において考察の対象とされていたものである。本稿ではあと 2 つ、協力ゲーム理論における解概念をベースとした AFA を定式化する。

このうち 1 つめの AFA は、残余均等配分解に対して、「逆残余均等配分解」とも呼ぶべきものである。協力ゲーム理論においては、Egalitarian Nonseparable Contribution (ENSC) という名で、1990 年代に定式化・公理化が行われている (Dragan *et al.* 1996 および Drissen and Funaki 1991 参照)。

2 (1) における、特徴量が 3 つの場合の残余均等配分解 (ES) についての議論を思い出されたい。ES では、1 番目のステップにおいて、全ての特徴量が未知の状態である $f(\emptyset)$ からスタートし、そこから特徴量が 1 つだけ既知の状態における予測値を考え、その差を限界貢献度として、その特徴量の「取り分」としたのであった。一方、ENSC は、逆に全ての特徴量が既知の状態である $f(A, B, C)$ からスタートする。そして、そこから後ろ向きに特徴量が 1 つだけ未知の状態を考え、その差

図 7 NSC (逆残余均等配分解を用いた AFA) の計算方法



を限界貢献度として当該特徴量の「取り分」とするのである。そして、2 番目のステップは ES 同様、「全体のパイの大きさ」である $f(A, B, C) - f(\emptyset)$ から、1 番目のステップにおいて各特徴量がキープした分を差し引いた「残余」を、3 つの特徴量で均等配分し、キープした分に加える。これが ENSC に基づく AFA である (図 7)。
正確には、AFA としての ENSC $\Psi_{\tau,j}^{ENSC}$ は、以下の (6) 式によって定義される。

$$\Psi_{\tau,j}^{ENSC} = v_{\tau}(N) - v_{\tau}(N \setminus \{j\}) + \frac{v_{\tau}(N) - v_{\tau}(\emptyset) - \sum_{k \in N} \{v_{\tau}(N) - v(N \setminus \{k\})\}}{n}. \quad (6)$$

(6) 式の $v_{\tau}(N) - v_{\tau}(N \setminus \{j\})$ が、図 7 における STEP 1, そのあとの分数部分が STEP 2 に対応している。

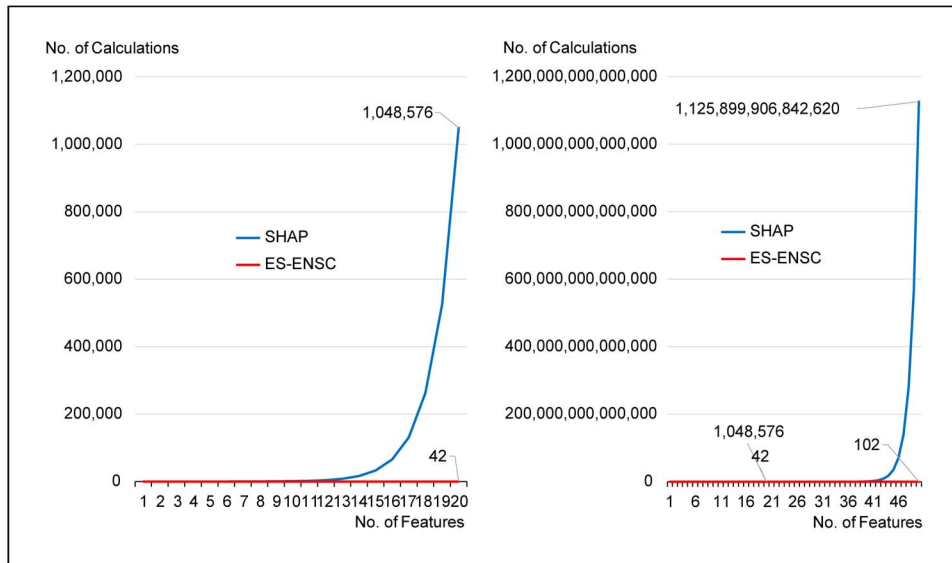
本項で新たに提示する 2 つめの AFA は、ES と ENSC を按分した $\Psi_{\tau,j}^{ES-ENSC}$ であり、以下の (7) 式によって定義される。

$$\Psi_{\tau,j}^{ES-ENSC} = \frac{1}{2} (\Psi_{\tau,j}^{ES} + \Psi_{\tau,j}^{ENSC}). \quad (7)$$

改めて既存の Ψ_{τ}^{SHAP} と、その代替的手法である $\Psi_{\tau,j}^{ES}$, $\Psi_{\tau,j}^{ENSC}$ および $\Psi_{\tau,j}^{ES-ENSC}$ を比較すると、 Ψ_{τ}^{SHAP} はありうる全ての順列における、全ての限界貢献値を用いて計算される。一方で、 $\Psi_{\tau,j}^{ES}$ は $v_{\tau}(\{j\}) - v_{\tau}(\emptyset)$ 、すなわち、空集合に対する j の貢献度のみを考慮し、 $\Psi_{\tau,j}^{ENSC}$ も $v_{\tau}(N) - v_{\tau}(N \setminus \{j\})$ のみを考慮する。したがって、用いる情報量は SHAP がもっとも大きく、この点においてフェアな方法であると言える。

一方、計算コストの観点からこれらの AFA を比較すると、どれだけの $v(S)$ を計算する必要があるかがポイントとなる。 Ψ_{τ}^{SHAP} は全ての S について $v(S)$ を計算する必要があるため、特徴量が n 個のときの計算コストは 2^n である。これに対し、 Ψ_{τ}^{ES} および Ψ_{τ}^{ENSC} は (i) n 個の $v(\{j\})$ (ES の場合) または $v(N \setminus \{j\})$ (ENSC の場合)、(ii) $v(\emptyset)$ と $v(n)$ 、の合計 $n+2$ である。 $\Psi_{\tau}^{ES-ENSC}$ は Ψ_{τ}^{ES} と Ψ_{τ}^{ENSC} の計算コストの合計から重複分を差し引いた $2n+2$ である。 Ψ_{τ}^{SHAP} と $\Psi_{\tau}^{ES-ENSC}$ の計算コストのギャップを図 8 に示した。ここから分かるように、特徴量の数が大きくなるほど計算コストのギャップは指数関数的に拡大していく。

図8 SHAP と ES-ENSC の計算コスト比較



以上が協力ゲーム理論の解概念に基づく代替的 AFA の理論的導出および考察である。次に、図 4 における Approach 2 で示した、SHAP の代替的手法を定式化する際のもう 1 つのアプローチである、カーネル関数を起点とした議論を概観する。

(4) LIME とカーネル

HIS 論文では、LIME (Ribeiro *et al.* 2016、以下 LIME 論文と呼ぶ) におけるカーネル関数の観点から、SHAP におけるカーネル関数は、「分析対象の観測値に近い摂動サンプルほど大きなウェイトが付与されるべき」という、LIME 論文において明記されていた条件を満たしていない点を指摘した。そのうえで、任意のカーネル関数を用いた AFA の一般的表現を導出し、上記条件を満たす複数の AFA を、SHAP の代替的な手法として提示した。2 節 (4) と 2 節 (5) ではこの点をレビューしつつ、AFA とカーネル関数の関係についてさらに掘り下げた議論を行う。

まず、LL 論文および LIME 論文の表記に従い、 x を分析対象である観測値、 z を x から生成された摂動サンプル (すべての特徴量が既知である x をベースに、一部の特徴量を未知としたときのデータ) とする。LL 論文および LIME 論文では、二値ベクトル z' および $z = h_x(z')$ を満たす写像 h_x を用いて z を z' に置き換えた上で分析しているが¹⁰、ここでは単純化のために $x = x'$ および $z = z'$ とする。すなわ

10 例えば、分析対象の観測値を $x = (20 \text{ 代、大卒、東京居住})$ とすると、 $x' = (1, 1, 1)$ であり ((20 代、

ち、元の観測値とその摂動サンプルははじめから二値ベクトルに変換されているものとする。

LIME 論文では、「複雑な機械学習モデル f を、分析対象である x の近傍において、「説明可能な複雑ではない」モデル g で局所的に近似する手法」として、以下の最小化問題を提示した：

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g).$$

ここで、 $g(z) = \phi_0 + \sum_{i=1}^n \phi_i z_i$ (ただし $\phi_i \in \mathbb{R}$)、すなわち、説明可能なモデルは特徴量に関して線形である。 G をすべての g の集合とし、 $\phi = (\phi_1, \dots, \phi_n) \in \mathbb{R}^n$ とする。 π_x はカーネル関数であり (本分析においてキーとなる重要な概念である)、 $\pi_x(z)$ で、分析対象である観測値 x と摂動サンプル z の近接度が測られる。 L は f および π_x のもとで、 g が f をどの程度近似しているかどうかを測る損失関数であり、近似度が高いほど損失が小さくなる。 $\Omega(g)$ は説明可能なモデル g の複雑さを測るペナルティ項である。

以上の定式化のもと、LIME 論文では、カーネル関数 π_x が満たすべき条件として、以下の点を挙げている：

- ・ x と z との近接度が高い (距離が小さい) ほど、 z に付与されるカーネル (重み) は大きくなるべきである。

説明可能なモデル g は、分析対象となる観測値 x の近くにおいて、 f をより正確に近似すべきであるため、そのような g を求めるうえで、 x に近い摂動サンプルを重視するように損失関数を定式化することを要請する上記の条件は、カーネル関数が当然満たすべき条件であるといえる¹¹。そして、LIME 論文では、具体的にカーネル関数を $\pi_x(z) = \exp(-D(x, z)^2)/\sigma^2$ (ただし D は距離関数、 σ は散らばりの程度) と定式化している。さらに、損失関数については、 $L(f, g, \pi_x) = \sum_{z \in Z} [f(z) - g(z)]^2 \pi_x(z)$ という局所的加重二乗損失関数 (locally weighted square loss function) を想定している。すなわち、摂動サンプル z において、説明可能なモデルが与える予測値 $g(z)$ がそもそもの学習モデルの予測値 $f(z)$ から乖離するほど損失は大きくなる。そして、そのような損失は、 z が x に近いほど (すなわち $\pi_x(z)$ が大きいほど)、より重視される。

LIME 論文では、以上の最小化問題の解を直接解析的に求めることはしていない。

大卒、東京居住) = $h_x((1, 1, 1))$ 、 $z = (40 \text{ 代、高卒、東京居住})$ であれば $z' = (0, 0, 1)$ ((40 代、高卒、東京居住) = $h_x((0, 0, 1))$)、 $z = (20 \text{ 代、大卒、大阪居住})$ であれば $z' = (1, 1, 0)$ である ((20 代、大卒、大阪居住) = $h_x((1, 1, 0))$)。

11 LIME 論文の 3 節を参照。例えば、Figure 3 において、 x に近い摂動サンプルは相対的に大きく表示されているが、これは当該サンプルにより大きな重みを付与していることを表している。

一方、LIME と SHAP の関係を考察した LL 論文では、この最小化問題に、追加的に $\Omega(g) = 0$ の仮定を置いている。これにより、LIME の最小化問題は以下で表される：

$$\arg \min_{g \in G} \sum_{z \in Z} [f(z) - g(z)]^2 \pi_x(z) = \arg \min_{\phi \in R^n} \sum_{z \in Z} \left[\sum_{i=1}^n \phi_i z_i - \{f(z) - \phi_0\} \right]^2 \pi_x(z). \quad (8)$$

ここで、以下の分析のために、(8) 式の最小化問題を、2 節 (2) で導入したノーテーションで書き換える。まず、(8) 式では $z \in Z$ についての和をとっているが、 z はすべての特徴量が既知である場合に対応する x を基に、一部の特徴量を未知として生成された摂動サンプルなので、これは $S \in 2^N$ についての和をとることを意味する。例えば、特徴量が 3 つのケースにおいて、 $z = (1, 1, 0)$ は、本稿のノーテーションでは $S = \{1, 2\}$ である。2 番目に、特定の z のもとでの $\sum_{i=1}^n \phi_i z_i$ は、 $z_i = 1$ 、すなわち $i \in S$ である i について ϕ_i の和をとることであるから、 $\sum_{i \in S} \phi_i$ となる。3 番目に、特定の z のもとでの $f(z)$ を考える。例えば $z = (1, 1, 0)$ のとき、 $f(1, 1, 0)$ は、「1 番目と 2 番目の特徴量が既知であるときの、学習モデルの予測値」であるので、 $v_\tau(\{1, 2\})$ となる。したがって、 $f(z)$ は $v_\tau(S)$ と置き換えられる。

以上より、(8) 式の最小化問題は、以下の通りに書き換えられる：

$$\arg \min_{\phi \in R^n} \sum_{S \in 2^N} \left[\sum_{i \in S} \phi_i - \{v_\tau(S) - \phi_0\} \right]^2 \pi_{x_\tau}(S). \quad (9)$$

さらに、LL 論文では、説明可能なモデル $g(z) = \phi_0 + \sum_{i=1}^n \phi_i z_i$ に関して、いくつかの制約を課している。1 点目は、 $z = (0, \dots, 0)$ のとき、 $\phi_0 = f(0, \dots, 0)$ 、すなわち、 ϕ_0 は、全ての特徴量が未知のときの学習モデルの予測値と一致していなければならない。これは、本稿のノーテーションでは $\phi_0 = v_\tau(\emptyset)$ である。したがって、(9) 式は以下に置き換えられる。

$$\arg \min_{\phi \in R^n} \sum_{S \in 2^N} \left[\sum_{i \in S} \phi_i - \{v_\tau(S) - v_\tau(\emptyset)\} \right]^2 \pi_{x_\tau}(S). \quad (10)$$

2 点目は、 $f(x) = g(x)$ 、すなわち、分析対象の観測値 x においては、 $g(x)$ は $f(x)$ と一致しなければならない。この条件を局所的正確性条件 (local accuracy) または効率性条件 (efficiency) と呼び、この条件を課すことで最小化問題の解は常に AFA となる。以上を整理すると、LL 論文の制約条件付最小化問題は、 Ψ_τ^{AFA} を解とすると、以下で表される。

$$\Psi_\tau^{AFA} = \arg \min_{\substack{\phi \in R^n \text{ with } \sum_{i \in N} \phi_i = v_\tau(N) - v_\tau(\emptyset)}} \sum_{S \in 2^N} \left[\sum_{i \in S} \phi_i - \{v_\tau(S) - v_\tau(\emptyset)\} \right]^2 \pi_{x_\tau}(S). \quad (11)$$

なお、制約条件なし最小化問題 (10) および制約条件付き最小化問題 (11) のいずれにおいても、カーネルをすべての S について等倍しても、最適化解は不変であることに注意されたい。

1 節で述べたように、LL 論文は (11) 式の解が SHAP と一致するようなカーネル関数 $\pi_{x_r}(S)$ が存在することを示し、さらにそのカーネル関数の具体的表現を導出した (LL 論文の Theorem 2)。一方、以下では、HIS 論文に基づき、カーネル関数についての対称性条件を課した上で、(10) 式および (11) 式の最小化問題に対する解の一般的表現を導く手順を紹介する。

■**カーネルについての条件** (10) 式および (11) 式におけるカーネル関数 $\pi_{x_r}(S)$ について、本項では以下の対称性条件を課す：

$$\pi_{x_r}(S) = \pi_{x_r}(T) \quad \left(\forall S, T \in 2^N \text{ with } |S| = |T| \right) \quad (12)$$

(12) 式は、特徴量の数に関して、 S と T が N と等距離であれば、カーネル関数は両者に等しい重みを与える、ということ意味する。これはある種の対称性条件であり、妥当な条件であろう。この条件と、前述した「カーネルをすべての S について等倍しても、最適化解は不変」という性質を利用することで、以下でみるように、最小化問題 (10) 式および (11) 式の解を解析的に求めることができる。

■**制約条件がない場合の最小化問題 (10) 式の最適解** まず、効率性条件を制約条件として課さない最小化問題 (10) 式の最適解を導出する。 ϕ_j についての最適化の 1 階の条件は、以下の通りである：

$$\sum_{S \in 2^N: j \in S} 2 \left(\sum_{i \in S} \phi_i - \{v_r(S) - v_r(\emptyset)\} \right) \pi_{x_r}(S) = 0. \quad (13)$$

したがって、 $i \neq j$ を満たす任意の $i, j \in N$ について、以下が成り立つ：

$$\sum_{S \in 2^N: i \in S} \left(\sum_{k \in S} \phi_k - \{v_r(S) - v_r(\emptyset)\} \right) \cdot \pi_{x_r}(S) = \sum_{S \in 2^N: j \in S} \left(\sum_{k \in S} \phi_k - \{v_r(S) - v_r(\emptyset)\} \right) \cdot \pi_{x_r}(S) \quad (14)$$

$$\begin{aligned} &\iff \sum_{S \subseteq N \setminus \{i, j\}} \left(\pi_{x_r}(S \cup \{i\}) \cdot \phi_i - \pi_{x_r}(S \cup \{j\}) \cdot \phi_j \right) \\ &= \sum_{S \subseteq N \setminus \{i, j\}} \left(\pi_{x_r}(S \cup \{i\}) \cdot v_r(S \cup \{i\}) - \pi_{x_r}(S \cup \{j\}) \cdot v_r(S \cup \{j\}) \right) \end{aligned} \quad (15)$$

$$\iff \phi_i - \phi_j = \sum_{S \subseteq N \setminus \{i, j\}} \left(\pi_{x_r}(S \cup \{i\}) \cdot v_r(S \cup \{i\}) - \pi_{x_r}(S \cup \{j\}) \cdot v_r(S \cup \{j\}) \right). \quad (16)$$

(16) 式の等号がなぜ成り立つかについてのロジックは、本節のその後の分析にも関連するのでここでやや詳しく述べる。(15) 式と (16) 式の右辺は同じであるので、左辺同士が等しいことを示せばよい。(15) 式の左辺第 1 項 $\phi_i \sum_{S \subseteq N \setminus \{i, j\}} \pi_{x_r}(S \cup \{i\})$ 、第 2 項 $\phi_j \sum_{S \subseteq N \setminus \{i, j\}} \pi_{x_r}(S \cup \{j\})$ において、 $\sum_{S \subseteq N \setminus \{i, j\}} \pi_{x_r}(S \cup \{i\}) = \sum_{S \subseteq N \setminus \{i, j\}} \pi_{x_r}(S \cup \{j\}) = 1$ とする。すなわち、空集合 $\emptyset$ と全体集合 N 以外の全ての提携についてのカーネルの和が 1 になるという条件を課す。すなわち、この条件を満たすように、既述の通り「カーネルを全ての S について等倍する」のである。この結果、(15) 式の左辺 $= \phi_i - \phi_j$ となる。

2 節 (5) 以降では、AFA に対するカーネル関数を求めたり、あるいは逆にカーネル関数を定めてから AFA を導出するが、その際には、この条件（カーネルの和が 1 となるという条件）を満たすようにカーネル関数を定式化していく。

最適解の導出に戻ると、(16) 式より、以下が成り立つ：

$$\phi_1 - \sum_{S: 1 \in S, S \neq N} \pi_{x_r}(S) \cdot v_r(S) = \cdots = \phi_n - \sum_{S: n \in S, S \neq N} \pi_{x_r}(S) \cdot v_r(S). \quad (17)$$

さらに、(13) 式より、以下が成り立つ：

$$\left(\sum_{S \in 2^N: j \in S} \pi_{x_r}(S) \right) \phi_j + \sum_{i \in N: i \neq j} \left(\sum_{S \in 2^N: i, j \in S} \pi_{x_r}(S) \right) \phi_i = \sum_{S \in 2^N: j \in S} \pi_{x_r}(S) \cdot (v_r(S) - v_r(\emptyset)). \quad (18)$$

したがって、両辺をそれぞれすべての $j \in N$ について足し合わせると、以下が成り立つ：

$$\left(\sum_{S \in 2^N: j \in S} \pi_{x_r}(S) + (n-1) \cdot \sum_{S \in 2^N: i, j \in S} \pi_{x_r}(S) \right) \sum_{j \in N} \phi_j = n \cdot \sum_{S \in 2^N: j \in S} \pi_{x_r}(S) \cdot (v_r(S) - v_r(\emptyset)). \quad (19)$$

(19) 式より、 $\phi = (\phi_1, \dots, \phi_n)$ は以下で表される：

$$\phi_j = \sum_{S: j \in S \neq N} \pi_{x_r}(S) \cdot v_r(S) + \frac{T - \sum_{i \in N} \{ \sum_{S: i \in S \neq N} \pi_{x_r}(S) \cdot v_r(S) \}}{n} \quad (20)$$

ここで、 T は以下の通りである：

$$T = \frac{n \cdot \sum_{S \in 2^N: j \in S} \pi_{x_r}(S) \cdot (v_r(S) - v_r(\emptyset))}{\sum_{S \in 2^N: j \in S} \pi_{x_r}(S) + (n-1) \cdot \sum_{S \in 2^N: i, j \in S} \pi_{x_r}(S)}. \quad (21)$$

(20) および (21) 式が、制約なし最小化問題 (10) の最適解である。前述した通り、LIME 論文では最適化問題にペナルティ項 $\Omega(z)$ が含まれており、最適化問題の解を解析的に導くことはせず、その代わりに解を近似的に導くアルゴリズムを提案し

ている。一方、本稿では、LL 論文と同様にペナルティ項をゼロと単純化したうえで、追加的に対称性の条件を課すことで解を解析的に導出している。また、この結果は、次にみる制約条件がある場合の最適解の一般化となっている。

■制約条件がある場合の最小化問題 (11) 式の最適解 次に、最適化問題 (11) の解 Ψ_{τ}^{AFA} を導出する¹²。(11) のラグランジアンは以下の通りである：

$$\mathcal{L}(\phi_1, \dots, \phi_n, \lambda) = \sum_{S \in 2^N} \left[\sum_{i \in S} \phi_i - \{v_{\tau}(S) - v_{\tau}(\emptyset)\} \right]^2 \cdot \pi_{x_{\tau}}(S) - \lambda \left[\sum_{i \in N} \phi_i - v_{\tau}(N) + v_{\tau}(\emptyset) \right].$$

ϕ_j についての最適化の 1 階の条件は、

$$\sum_{S \in 2^N: j \in S} 2 \left(\sum_{i \in S} \phi_i - \{v_{\tau}(S) - v_{\tau}(\emptyset)\} \right) \cdot \pi_{x_{\tau}}(S) - \lambda = 0$$

であり、これは制約条件がない場合同様、(17) 式が満たされることを意味する。したがって、(17) 式および $\sum_{j \in N} \phi_j = v_{\tau}(N) - v_{\tau}(\emptyset)$ を満たす $\phi = (\phi_1, \dots, \phi_j, \dots, \phi_n)$ は、以下で表される：

$$\Psi_{\tau,j}^{AFA} = \phi_j = \sum_{S: j \in S} \pi_{x_{\tau}}(S) \cdot v_{\tau}(S) + \frac{v_{\tau}(N) - v_{\tau}(\emptyset) - \sum_{i \in N} \{ \sum_{S: i \in S} \pi_{x_{\tau}}(S) \cdot v_{\tau}(S) \}}{n}. \quad (22)$$

制約条件がある場合の最小化問題 (11) 式において効率性条件 $\sum_{i \in N} \phi_i = v_{\tau}(N) - v_{\tau}(\emptyset)$ を課すことは、制約条件がない場合の最小化問題 (10) 式において、 $S = N$ のときのカーネルを無限に大きくすることに等しい。実際、(20) 式および (21) 式において $\pi_{x_{\tau}}(N) = \infty$ とすると、(22) 式が得られる¹³。すなわち、(20) 式および (21) 式は (22) 式の一般化である。

(22) 式は、AFA (Ψ_{τ}^{AFA}) を、カーネル関数 $\pi_{x_{\tau}}(S)$ の関数として表現している。このことにより、任意のカーネル関数に対して AFA を求めることができ、特定のカーネル関数に基づいた AFA を定式化するうえで極めて有用である。図 4 に戻ると、右下の四角形内でカーネル関数を 1 つ定めると、それに応じて上部の楕円形内で SHAP の代替的手法が 1 つ求まるという、右側の赤い矢印で示された Approach 2 を、(22) 式を用いて実行することができる。また、逆に、特定の AFA を所与として、それに対応するカーネル関数を求めることもできる（この点についての正確な議論は補論 1 を参照）。

以上に基づき、次の 2 節 (5) では、HIS 論文をさらに発展させるかたちで、(I) 前節で取り上げた協力ゲーム理論の解概念に基づく AFA を、それに対応するカーネ

.....
12 なお、Ruiz *et al.* (1998) は、協力ゲーム理論の枠内で、既にこの制約条件付き最小化問題と類似の問題を分析し、同様の結論を導いている。

13 具体的には、 $\pi_{x_{\tau}}(N) = \infty$ とすることは、 N 以外の S について $\pi_{x_{\tau}}(S) = 0$ とすることと等しい。これを (21) 式に代入すると $T = v_{\tau}(N) - v_{\tau}(\emptyset)$ となり、(22) 式と一致する。

ル関数の観点から評価する。さらに、(II) 先述した「 x と z との近接度が高い（距離が小さい）ほど、 z に大きな重みを与える」という条件を満たすカーネル関数を定式化し、それに対応する AFA を導出する。

(5) カーネル関数に基づく AFA の理論的比較

ここでは、2 節 (4) で得られた AFA とカーネル関数の関係を基に、まずは、2 節 (3) で取り上げた、協力ゲーム理論の解概念に基づく AFA に対応するカーネル関数を導出する (AFA → カーネル関数、2 節 (5) イ.)。次に、「既知の特徴量の数に関して増加関数となる」という、LIME 論文においてカーネル関数を持つべきとされた性質を有するカーネル関数を定式化したうえで、そのカーネル関数に対応する AFA を導出する (カーネル関数 → AFA、2 節 (5) ロ.)。

イ. 協力ゲーム理論の解概念に基づく AFA に対応するカーネル関数の導出

■ SHAP まず、 Ψ_{τ}^{SHAP} に対応するカーネル関数は、以下の (24) 式である (補論 1 参照)。すなわち、(24) 式を (22) 式に代入して得られる AFA は、SHAP Ψ_{τ}^{SHAP} となる¹⁴：

$$\pi_{x_{\tau}}^{SHAP}(S) = \frac{n}{{}_nC_{|S|} \cdot |S| \cdot (n - |S|)}. \quad (24)$$

上式のカーネル関数を用いて最適化問題 (11) を解いて得られた SHAP を、KernelSHAP と呼ぶことがある。SHAP と KernelSHAP は概念的には同じものであるが、SHAP の計算上は区別される。すなわち、実際のデータおよび学習済みモデルを所与として SHAP を導出する際、SHAP の定義式に基づいて計算しようとすると、前述のように、1 つの観測値 τ に対して 2^n 個の $v(S)$ を計算する必要がある。このため、特徴量の数が大きい場合、厳密な SHAP を導出するには多大な計算コストがかかる (しばしば計算不能となる)。そこで、定義から SHAP を計算するのではなく、最小化問題 (11) に着目して、一定基準以上のカーネル (ウェイト) $\pi_{x_{\tau}}(S)$ をカバーするようにサンプリングして最小化問題を解く手法が提示されている¹⁵。そして、これによって解かれた SHAP を特に KernelSHAP と呼ぶ。ただし、KernelSHAP もなお計算コストが比較的大きいとされている。

14 なお、LL 論文では、SHAP に対応するカーネル関数は表現は以下の通りである：

$$\pi_{x_{\tau}}(S) = \frac{n-1}{{}_nC_{|S|} \cdot |S| \cdot (n - |S|)}. \quad (23)$$

(24) の右辺は (23) の右辺を定数倍 ($n/(n-1)$ 倍) したものであるので、どちらのカーネルを用いても最適化問題 (11) の解は不変である。LL 論文においてなぜ分子が $n-1$ とされたかは不明である。

15 Python の SHAP パッケージでは、(I) ウェイトの大きい S から順に計算対象に含めていくという作

(24) 式において、 $|S| = 0$ または $|S| = n$ のとき $\pi_{x_r}^{SHAP}(S) = \infty$ であるほか、それ以外の $|S|$ の領域においてもカーネル関数は $|S|$ について U 字型（凹型）となっている。これは、LIME 論文においてカーネル関数が持つべき性質とされた、「 z が x に近いほど、より大きな重みが付与される」すなわち π_{x_r} が $|S|$ に関する増加関数であるという条件を満たしていない。

■ ES 次に、 Ψ_{τ}^{ES} に対応するカーネル関数は、以下の (25) 式である。

$$\pi_{x_r}^{ES}(S) = \begin{cases} 1 & \text{if } |S| = 1 \\ 0 & \text{if } 2 \leq |S| \leq n-1. \end{cases} \quad (25)$$

$\pi_{x_r}^{ES}$ も、 $|S|$ に関する増加関数という条件を明らかに満たしていない。すなわち、2 節 (3) でも言及したように、ES は、計算コストが相対的に小さいという大きなメリットがある一方で、カーネル関数が $|S|$ に関する増加関数とはなっていないという点においては、(SHAP と同様に) 必ずしも望ましい性質を有してはいないといえる。

■ ENSC Ψ_{τ}^{ENSC} に対応するカーネル関数は、以下の (26) 式である。

$$\pi_{x_r}^{ENSC}(S) = \begin{cases} 0 & \text{if } 1 \leq |S| \leq n-2 \\ 1 & \text{if } |S| = n-1. \end{cases} \quad (26)$$

$\pi_{x_r}^{ENSC}$ は、(極端な形ではあるが) $|S|$ に関する増加関数になっている。また、ENSC は、協力ゲーム理論において既に公理化が完了している (Drissen and Funaki 1991 参照)。 $v_{\tau}(N)$ と $v_{\tau}(N \setminus \{i\})$ 以外の提携値についての情報は用いていないが、その分、SHAP などと異なり、厳密に計算したとしても計算上の負荷は小さいというメリットがある。

■ ES-ENSC 最後に、 $\Psi_{\tau}^{ES-ENSC}$ に対応するカーネル関数は、以下の (27) 式である。

$$\pi_{x_r}^{ES-ENSC}(S) = \begin{cases} \frac{1}{2} & \text{if } |S| = 1 \text{ or } n-1 \\ 0 & \text{if } 2 \leq |S| \leq n-2. \end{cases} \quad (27)$$

$\pi_{x_r}^{ES-ENSC}$ は、 $|S|$ に関して U 字型となっており、増加関数という条件は満たしていない。ただし、SHAP とカーネル関数の形状が類似しており、かつ SHAP に比べ

業を、計算対象に含める S のウェイトが一定基準に到達するまで続ける、(II) (I) で計算対象に含まれなかった S をランダムに選んで計算対象に加える、という手順で最小化問題を解き、KernelSHAP を算出する。

ると計算コストが大幅に小さいことから（図 8 参照）、SHAP を近似的に計算する手法としては好ましい性質を有していると言える。また、協力ゲーム理論では、この解を含むクラスの解について、既に公理化が完了している（Kongo 2024 参照）。

ロ. 特定のカーネル関数に対応する AFA の導出

前項では、特定の AFA から出発し、それに対応するカーネル関数を導出した。本稿では、逆に、LIME 論文で示されたカーネル関数を持つべき条件である「既知の特微量の数に関する増加関数」一すなわち、 $|S|$ についての増加関数という条件一を満たす特定のカーネル関数を定式化し、そのカーネル関数に対応する AFA を導出する。

■線形に増加するカーネル関数に基づく AFA まず、以下のカーネル関数 $\pi_{x_\tau}^{LnK}$ は $|S|$ に関して線形に増加しており、LIME 論文で示された条件を満たしている¹⁶：

$$\pi_{x_\tau}^{LnK}(S) = \frac{|S|}{n \cdot 2^{n-3}}. \quad (28)$$

(28) を (22) に代入して、以下の AFA を得る：

$$\Psi_{\tau,j}^{LnK} = \phi_j = \sum_{S:j \in S} \frac{|S|}{n \cdot 2^{n-3}} \cdot v_\tau(S) + \frac{v_\tau(N) - v_\tau(\emptyset) - \sum_{i \in N} \left\{ \sum_{S:i \in S} \frac{|S|}{n \cdot 2^{n-3}} \cdot v_\tau(S) \right\}}{n}, \quad (29)$$

$\Psi_{\tau,j}^{LnK}$ は、カーネル関数として望ましい性質を有する、SHAP と代替的な 1 つめの AFA である。

■指数関数的に増加するカーネル関数に基づく AFA 2 節 (4) で触れたように、LIME 論文（2016）において、カーネル関数は以下で定義されている：

$$\pi_{x_\tau}(z) = \exp\left(\frac{-D(x,z)^2}{\sigma^2}\right)$$

ただし D は距離関数、 σ は散らばりの程度 (width) である。ここで、 x は分析対象となる観測値であり z は x からの摂動サンプル、さらにこれらは二値関数であったことに留意されたい。したがって、本稿のノーテーションに従うと、カーネル関数 π_{x_τ} は以下のように表現できる：

$$\pi_{x_\tau}(S) = \exp\left(\frac{-\left(\sqrt{\sum_{i \in S} 0^2 + \sum_{i \notin S} 1^2}\right)^2}{\sigma^2}\right) = \exp\left(\frac{-(n - |S|)}{\sigma^2}\right) = \frac{\left(e^{\frac{1}{\sigma^2}}\right)^{|S|}}{\left(e^{\frac{1}{\sigma^2}}\right)^n}. \quad (30)$$

.....
16 $n \cdot 2^{n-3}$ で割るのは、カーネルの合計を 1 にするためである。2 節 (4) の (16) 式のすぐ後の議論を参照。

さらに、カーネルの和が 1 になるように定数倍すると、以下が得られる：

$$\pi_{x_\tau}(S) = \frac{\left(e^{\frac{1}{\sigma^2}}\right)^{|S|-1}}{\left(e^{\frac{1}{\sigma^2}} + 1\right)^{n-2}}.$$

さらに、 $\sigma = \sqrt{1/\log 2}$ と仮定することで、以下の簡潔なカーネル関数 $\pi_{x_\tau}^{ExK}$ を得る：

$$\pi_{x_\tau}^{ExK}(S) = \frac{2^{|S|-1}}{3^{n-2}}. \quad (31)$$

$\pi_{x_\tau}^{ExK}$ は $|S|$ に関して指数関数的に増加しており、それに基づく AFA である $\Psi_{\tau,j}^{ExK}$ は (32) 式の通りとなる。

$$\Psi_{\tau,j}^{ExK} = \phi_j = \sum_{S:j \in S} \frac{2^{|S|-1}}{3^{n-2}} \cdot v_\tau(S) + \frac{v_\tau(N) - v_\tau(\emptyset) - \sum_{i \in N} \left\{ \sum_{S:i \in S} \frac{2^{|S|-1}}{3^{n-2}} \cdot v_\tau(S) \right\}}{n} \quad (32)$$

$\Psi_{\tau,j}^{ExK}$ が、カーネル関数上の望ましい性質を有する 2 つめの代替的 AFA である。上添字の ExK は指数関数的 (exponentially) に増加するカーネル関数を意味する。

■対数関数的に増加するカーネル関数に基づく AFA (28) 式で定義されるカーネル関数は、統計学における三角カーネル関数 (triangular kernel) に対応している。また、(31) 式のカーネル関数は、凸型カーネル関数に対応している。一方、以下で定義される凹型カーネル関数は、エパネチニコフ・カーネル (Epanechnikov kernel) あるいはコサイン・カーネル (cosine kernel) に対応するものである¹⁷。

$$\pi_x(S)^{CvK} = \frac{|S|(2n - |S|)}{(3n^2 - n + 2) \cdot 2^{n-4}}. \quad (33)$$

$\pi_x(S)^{CvK}$ に基づく AFA である $\Psi_{\tau,j}^{CvK}$ は、以下の (34) 式の通りである：

$$\begin{aligned} \Psi_{\tau,j}^{CvK} = & \sum_{S:j \in S} \frac{|S|(2n - |S|)}{(3n^2 - n + 2) \cdot 2^{n-4}} \cdot v_\tau(S) \\ & + \frac{v_\tau(N) - v_\tau(\emptyset) - \sum_{i \in N} \left\{ \sum_{S:i \in S} \frac{|S|(2n - |S|)}{(3n^2 - n + 2) \cdot 2^{n-4}} \cdot v_\tau(S) \right\}}{n}. \end{aligned} \quad (34)$$

カーネル関数が (31) 式のように指数関数的に増加しているのであれば、摂動サンプル z が分析対象となる観測値 x に近づくにしたがい、 z に付与される重みは急激に増加していく。これは、「説明可能なモデル g が x 近傍で学習モデル f を近似

17 ここでもカーネルの和が 1 になるように調整されているほか、分子に $2n$ を加えているのは、定義域において常に傾きがプラスになるようにするためである。

しているかどうか」という点をより重視して最適な g を探索することを意味する。一方、カーネル関数が (33) のように対数関数的に増加しているのであれば、 x から離れた摂動サンプルも比較的重視して g を探索することを意味する。

■一定のカーネル関数に基づく AFA: LS プレ仁との一致 最後に、以下のカーネル関数を考える¹⁸：

$$\pi_{x_\tau}^{PNucl}(S) = \frac{1}{2^{n-2}}. \quad (35)$$

これは定数、すなわち $|S|$ に関して独立なカーネル関数である。(35) 式を (22) 式に代入することで、以下の AFA を得る：

$$\Psi_{\tau,j}^{PNucl} = \phi_j = 2 \left(\frac{1}{2^{n-1}} \sum_{S:j \in S} v_\tau(S) \right) + \frac{v_\tau(N) - v_\tau(\emptyset) - \sum_{i \in N} \left\{ 2 \left(\frac{1}{2^{n-1}} \sum_{S:i \in S} v_\tau(S) \right) \right\}}{n} \quad (36)$$

$\Psi_{\tau,j}^{PNucl}$ は、協力ゲームにおける解概念である LS（最小二乗）プレ仁（Ruiz *et al.* 1996, 1998）と一致することが証明できる¹⁹。カーネル関数は特微量の数から独立（すなわち定数）であるが、単純なカーネル関数に基づいて導出された AFA が、LS プレ仁という解概念で既に 1990 年代には協力ゲーム理論の枠内で定式化されていたのは、興味深い事実であると言える。

(6) SHAP とその代替的手法のまとめ

表 1 は、本節で定式化・分析対象とした 8 つの AFA をまとめたものである。最初の 4 つが協力ゲーム理論の解概念に基づいた AFA であり、それぞれシャーププレイ値、残余均等配分解（ES）、逆残余均等配分解（ENSC）および ES と ENSC の按分解（ES-ENSC）に対応している。5 番目から 7 番目は既知の特微量の数 $|S|$ （すなわち、LIME およびカーネル関数の観点からは、分析対象となっている観測値と摂動サンプルの近さ）に関して増加するカーネル関数から導出された AFA であり、それぞれ $|S|$ に関して線形に増加（ $\Psi_{\tau,j}^{LnK}$ ）、指数関数的に増加（ $\Psi_{\tau,j}^{ExK}$ ）、対数関数的に増加（ $\Psi_{\tau,j}^{CvK}$ ）するカーネル関数に基づいている。8 番目は一定のカーネル関数から導出される AFA であり、かつ協力ゲーム理論の解概念としては LS プレ仁に一致する。

なお、これらの AFA は、学習モデルが（複雑な「ブラックボックスモデル」ではなく）線形である場合は、すべて一致し、またその分解パターンは線形回帰モデルの回帰パラメータを用いた要因分解と一致する。このことは、(I) 本稿で示した

18 (35) 式で定義されるカーネル関数は、一様カーネル関数（uniform kernel）に対応する。

19 証明については著者に問い合わせされたい。

表 1 本分析における AFA 一覧

記号	AFA 式番号	カーネル 関数式番号	文献	協力ゲーム理論の 解概念としての特徴	カーネルの 特徴
$\Psi_{\tau,j}^{SHAP}$	(4)	(24)	LL 論文 (2017)	シャープレイ値 (1953)	U 字型
$\Psi_{\tau,j}^{ES}$	(5)	(25)	Condevaux <i>et al.</i> (2022)	ES (残余均等配分)	減少関数
$\Psi_{\tau,j}^{ENSC}$	(6)	(26)	本分析	ENSC (逆残余均等 配分) (1996, 1991)	増加関数
$\Psi_{\tau,j}^{ES-ENSC}$	(7)	(27)	本分析	ES と ENSC の按分 (2024) 等	U 字型
$\Psi_{\tau,j}^{LnK}$	(29)	(28)	HIS 論文 (2024)	—	線形増加
$\Psi_{\tau,j}^{ExK}$	(32)	(31)	HIS 論文 (2024)	—	指数関数的増加
$\Psi_{\tau,j}^{CvK}$	(34)	(33)	HIS 論文 (2024)	—	対数関数的増加
$\Psi_{\tau,j}^{PNucl}$	(36)	(35)	HIS 論文 (2024)	LS プレ仁 (1998)	一定

(注) 「文献」 および 「ゲーム理論解としての特徴」 に付されている番号は末尾の参考文献の番号に対応している。

AFA はすべて線形回帰モデルの要因分解の一般化であること、(II) 非線形なモデルにこれらの AFA を適用することによって、またそのことによってのみ、AFA 間で分解パターンに違いが生じること、を意味する。補論 2 では、この点についての数学的な議論を行っている。

3. SHAP とその代替的手法の金融・経済データへの適用

3 節では、表 1 で示した 8 つの AFA を実際の金融・経済データに適用し、各特徴量の予測貢献度に関する分解パターンが、AFA 間でどの程度異なるのかを比較分析していく²⁰。具体的には、(1)でわが国の 10 年国債利回り、(2)でわが国の失業率を対象にする²¹。

比較分析の際に用いる手法は、(A) 時系列グラフによる視覚的な比較、(B) 特定

20 公表データおよび Bloomberg から取得したデータ等を用いた。

21 既述のとおり、本稿で取り上げた AFA は、時系列データだけではなく、様々な数値データ、画像や音声データといった全てのタイプのデータに適用可能である。本稿では、著者らのドメイン知識も活用して AFA 評価を行う観点から、わが国の経済・ファイナンスに関する時系列データを対象とした。

の観測値のある特徴量において、異なる AFA によって与えられた値の差の絶対値をベースにした比較、の 2 つである。

なお、本分析の目標は、SHAP およびその代替的手法の分解パターンの違いを見ることであり、機械学習モデルそのものの予測精度や汎化性能を評価することではない。したがって、以下では、機械学習モデルは XGBoost で固定し、かつ、全てのデータを学習モデルとした In-sample の分析を行う²²。また、補論 3 では、平木ほか（2025）で扱った金価格への AFA の適用を ENSC や ES-ENSC にまで拡張し、本節で得られた洞察や傾向が、金価格のケースでも明確に観察されることを示している。

（1）わが国 10 年債利回りの AFA 分解

8 つの AFA を適用する最初のデータは、わが国の長期金利である（図 9）。具体的には、1998 年 1 月から 2024 年 12 月までの、10 年国債利回り（月次）の前月差を被説明変数とする。特徴量は日本銀行〔2021〕で示された長期金利推計などを参考に、以下の 4 つとした²³。

- ・ US10y: 米国長期金利（10 年物、%、1 期ラグ）
- ・ IIP: 鉱工業生産指数（OR、指数、1 期ラグ）
- ・ Inflation: 消費者物価指数（前年比、%、1 期ラグ）
- ・ BOJ_JGB: 日本銀行国債保有比率（資金循環統計ベース）

機械学習の標準的な手法に倣い、各特徴量は平均 0、分散 1 に標準化したものを学習データとして用いる。学習モデルは前述の通り XGBoost である。

表 1 で示した 8 つの AFA にもとづく分解パターンの違いは、図 10、図 11 および図 12 に示されている。それぞれ青い実線が実際の長期金利の推移、緑の実線が学習モデルによる予測値、そして棒グラフが各特徴量に割り当てられた AFA の値で

.....
22 したがって、以下で示すグラフから分かるように、いわゆるオーバーフィッティングがみられるが、これも、「学習済みモデルを所与として、異なる AFA 間の分解パターンの違いを比較する」という目的と照らし合わせると、ここでの論点とはならない。なお、XGBoost とは、多数の決定木を順に構築し、前のモデルの誤差を修正しながら予測精度を高めていく勾配ブースティング法をベースとした機械学習モデルである。

23 日本銀行〔2021〕では、10 年物国債金利を、(I) 有効求人倍率、消費者物価（除く生鮮食品、前年比、%）、米国債金利（10 年物、%）と、(II) 米国債金利（10 年物、%）、実質 GDP 成長率予想（%）、日本銀行の国債保有割合（%）、という 2 つのパターンで推計している。ここでは、AFA の比較を視覚的に行う観点から、比較が容易となるように特徴量の数を押さえつつ、これらの特徴量選定を参考にした。

図 9 長期金利（10 年国債利回り）の推移



ある（いずれについても、視覚的な比較を容易にする観点から、2012 年初からの累積変化幅ベースで示している）。したがって、ある特定の月において、各特徴量に対応する棒グラフを積み上げた高さは、その月の緑実線の高さと一致する。

図 10 の左パネルは、参考までに、学習モデルを XGBoost ではなく線形回帰 (OLS) モデルとして AFA 分解を行ったものである。学習モデルが線形回帰モデルの場合には、本分析で示したどの AFA を用いても、分解パターンは同一になる（この点についての厳密な証明は補論 2 を参照）。線形モデルと XGBoost の予測精度や特徴量の貢献度のパターンの違いは本分析の主目的ではないのでこれ以上の詳述は行わないが、XGBoost と比べると、回帰モデルによる予測は、青実線で示されている実績値と緑実線で示されている予測値の乖離が相対的に大きいこと、また、線形性を仮定していることから、各特徴量の寄与度の推移は、その特徴量自体の推移と比較的類似したパターンとなっていることがわかる。

本分析の主な対象である 8 つの AFA による分解パターンの違いは、図 10 の中央パネル以降で示されている。図 10 の中央パネルが SHAP（表 1 における $\Psi_{\tau,j}^{SHAP}$ ）、右パネルが LS プレ仁に対応する AFA（表 1 における $\Psi_{\tau,j}^{PNucl}$ ）、図 11 は左パネルから残余均等配分解 (ES) に対応する AFA ($\Psi_{\tau,j}^{ES}$)、逆残余均等配分解 (ENSC) に対応する AFA ($\Psi_{\tau,j}^{ENSC}$)、そしてそれらを按分する解 (ES-ENSC) に対応する AFA ($\Psi_{\tau,j}^{ES-ENSC}$) が示されている。ここまでは協力ゲーム理論における既存の解概念を

図 10 長期金利の AFA 分解 (1)

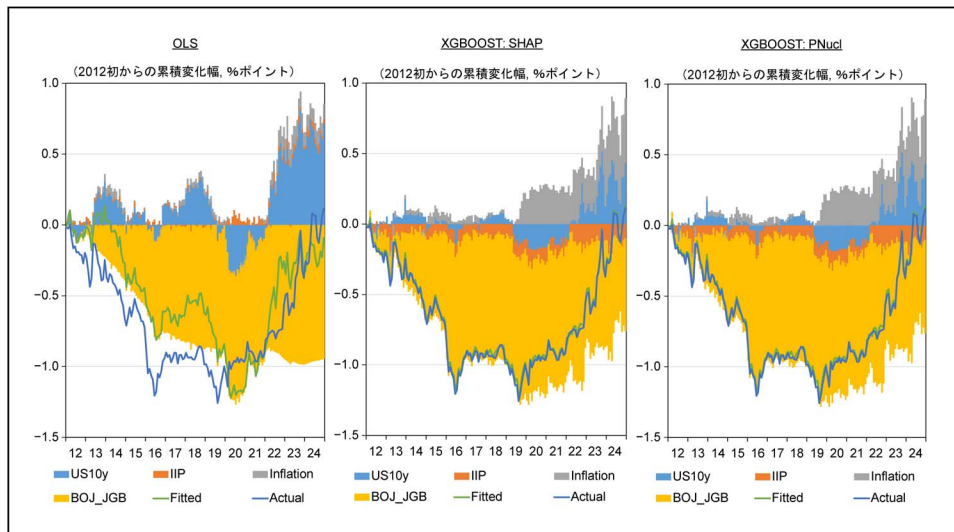
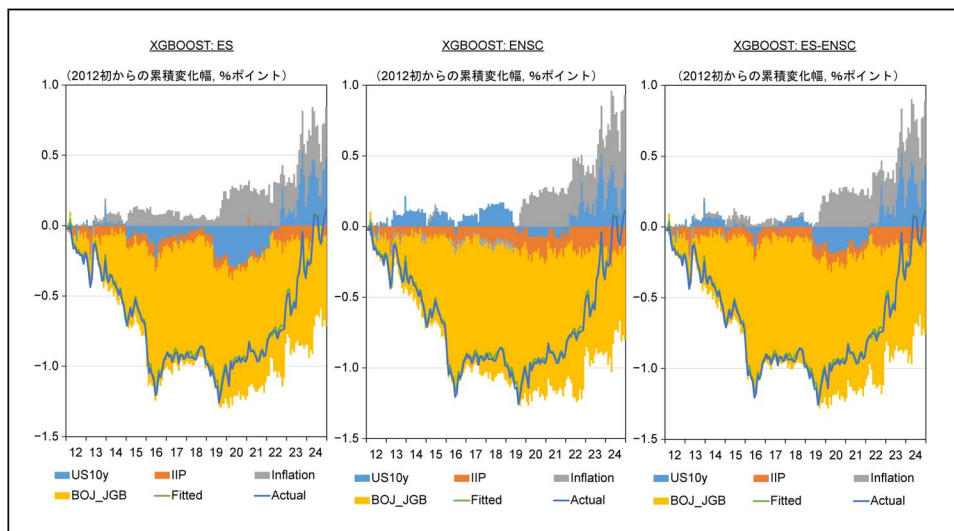
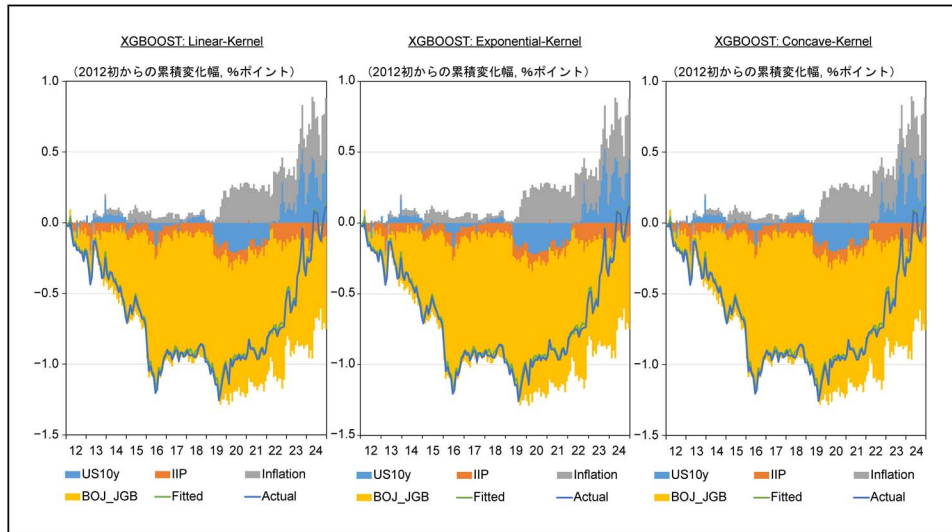


図 11 長期金利の AFA 分解 (2)



ベースとした AFA の分解パターンを示したものである。次の図 12 では、既知の特徴量の数について増加関数となっているカーネル関数に基づく 3 つの AFA を用いた分解パターンを示している。左パネルが線形に増加するカーネル関数に基づく AFA (表 1 における $\Psi_{\tau,j}^{LnK}$)、中央パネルが指数関数的に増加するカーネル関数に基づく AFA ($\Psi_{\tau,j}^{ExK}$)、右パネルが対数関数的に増加するカーネル関数に基づく AFA

図 12 長期金利の AFA 分解 (3)



($\Psi_{\tau,j}^{CvK}$) である。

これらのグラフにおける分解パターンの違いを視覚的にみると、いくつかの傾向が確認できる。

1 点目は、全体としてみると、AFA の間の分解パターンは概ね類似しているということである。例えば、どの AFA による分解パターンをみても、黄色で示された日本銀行国債保有比率が、2012 年以降の長期金利低下の最大の押し下げ要因となっている。また、2020 年以降、インフレ率（消費者物価前年比）が長期金利の押し上げ要因となっていること、米国長期金利が期中を通じて本邦長期金利に影響を及ぼしていたこと、IIP が継続的に長期金利の下押し圧力となっていたこと、などの点も、全ての分解パターンに共通してみられ、こうした全体感について AFA の間に大きな違いはない。

2 点目は、SHAP と ES、および SHAP と ENSC との比較に関するものである。SHAP と ES、および SHAP と ENSC の違いは、SHAP とそれ以外の AFA との違いと比べて相対的に大きい。例えば、図 11 の左パネルで示された ES をみると、SHAP に比べて、期中前半（特に 2015 年から 2018 年頃にかけて）の米国長期金利のマイナス寄与が大きい。また、インフレ率の寄与の大きさが期中を通じて SHAP よりも大きくなっている。これらの違いは、視覚的に確認できる程度に明確なものである。一方、図 11 の中央パネルで示された ENSC をみると、SHAP に比べて、インフレ率の寄与が特に期中前半においては小さく（あるいは、マイナスになっている）。また、米国長期金利については、ES とは逆に、期中前半にはプラス寄与となっているほか、2019 年から 2021 年頃にかけてのマイナス寄与が SHAP

対比小さい。このように、SHAP との分解パターンの違いは、特に ES と ENSC において、視覚的に容易に確認できる程度に明確になっている。ES および ENSC が SHAP と大きく異なる点は、既知の特微量の数が中程度（あるいは、協力ゲームの言葉で言えば、提携のサイズが中程度）である場合の予測値の情報を考慮していないということである。分解パターンの違いは、この点に由来するものであると考えられる。

3 点目は、上記 2 点目とも関連するが、ES と ENSC を按分した AFA である ES-ENSC の分解パターンについてである。図 11 の右パネルを見ると、ES-ENSC の分解パターンは、ES と ENSC の分解パターンを按分したイメージとなっている。これは、ES-ENSC の定義から自然に予想されることである。そして、ES-ENSC と SHAP の分解パターンをみると、視覚的には違いが見出しがたいほど両者が類似していることが分かる。これは、両者のカーネル関数の形状が似ていることから（表 1 参照）、理論的に想定できる結果である。(3) 節および図 8 で言及したように、ES-ENSC は、SHAP に対して計算コストが小さく、その差は特微量の数が増えるほど指数関数的に拡大していく。したがって、ES-ENSC は、SHAP を近似する AFA として優れた性質を持っていることを、この例は強く示唆しているといえる。

4 点目は、図 12 で示した、既知の特微量の数について増加関数となっているカーネル関数に基づく AFA についてである。図 12 をみると、カーネル関数の増加パターンの違い（線形か（左パネル）、指数関数的か（中央パネル）、対数関数的か（右パネル））による分解パターンの違いは、視覚的にはほとんど確認できない。さらに、これらの分解パターンは SHAP と類似したものになっている。この間、一定のカーネル関数を持ち、かつ協力ゲーム理論の解概念である LS プレ仁に基づく AFA の分解パターンも、SHAP と類似している（図 10 の右パネル）。以上より、この長期金利データを用いた例においては、カーネル関数の違いが、視覚的に見たときの分解パターンに大きな違いをもたらすといった傾向は（前述した ES および ENSC を除いては）観察されなかった。これは、SHAP およびその代替的な手法を用いた AFA による可視化分析の頑健性を示した結果であるともいえる。

次に、異なる AFA を適用したときの貢献度の差を定量化した比較を行う。表 2 で示された行列は、左下の領域と右上の領域に分かれている。まず、左下の赤字で示された領域に着目すると、 i 行 j 列のセルの数値は、ある特定の観測値におけるある特微量について、 i 行と j 列に対応した AFA を差の絶対値を計算し、それを全ての特微量と全ての観測値について平均した値を示している。したがって、セル内の値が大きいほど、対応する 2 つの AFA が与える寄与度が乖離していること意味する。左下の領域は、学習モデルを XGBoost にした場合の差、右上の領域は、学習モデルを回帰モデルにした場合の差である。既述の通り、学習モデルが回帰モデルであれば、本稿で取り上げた AFA は全て同じ値を与えることから、右上の領域は

表 2 AFA 間の相違度（長期金利データ）

	左下：XGBoost				右上：Linear Model			
	SHAP	PNucl	ES	ENSC	ES-ENSC	Linear	Exponential	Concave
SHAP	—	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
PNucl	0.000011	—	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
ES	0.004159	0.004156	—	0.00000	0.00000	0.00000	0.00000	0.00000
ENSC	0.004169	0.004172	0.008328	—	0.00000	0.00000	0.00000	0.00000
ES-ENSC	0.000021	0.000032	0.004164	0.002769	—	0.00000	0.00000	0.00000
Linear	0.001044	0.001041	0.003115	0.005213	0.001054	—	0.00000	0.00000
Exponential	0.001390	0.001387	0.002769	0.005559	0.001398	0.000346	—	0.00000
Concave	0.000728	0.000725	0.003432	0.004896	0.000739	0.000316	0.000663	—

（注）8 種類の AFA のそれぞれの組み合わせについて、平均絶対差（ある観測値におけるある特徴量について、2 つの AFA の差の絶対値を計算し、それを全ての特徴量およびすべての観測値について平均したもの）を表示。右上の領域が学習モデルを線形回帰モデル、左下の領域が学習モデルを XGBoost にしたときの平均絶対差。

常にゼロとなる。

学習モデルを XGBoost にした表 2 の左下の領域に着目すると、以下の点が分かる。

1 点目は、グラフによる分析からも明らかになった通り、SHAP との乖離は、ES および ENSC で相対的に大きくなっている。例えば、SHAP と ES を比べると、ある特徴量の両者の違いの平均は 0.004159%、すなわち 0.4bps 程度である。一方、分析期間中の長期金利前月差（絶対値）の平均は、約 0.07%（7bps）であった。したがって、ある特定の月におけるある特徴量について、2 つの AFA が与える値は平均的に 6% 程度（0.4bps/7bps）乖離するといえる。

2 点目は、EN-ENSC および LS プレ仁をベースにした AFA は、SHAP と乖離が極めて小さく（0.001～0.002bps 程度）、上記の基準で比較すると、両者と SHAP との乖離は 0.01% 程度にとどまる。

3 点目は、AFA 間の差異の大きさは、各 AFA に対応するカーネル関数の形状を反映したものとなっている。例えば、特徴量の数とは独立の、一定のカーネル関数を持つ LS プレ仁（PNucl）をベースにした AFA を基準に、それ以外の AFA との差をみると、一定の範囲内で PNucl のカーネル関数と類似した形状となる、U 字型のカーネル関数を持つ SHAP との差がもっとも小さくなっている。さらに、増加関数型のカーネル関数に基づく AFA と PNucl との違いをみると、傾きが急激に高まる指数関数型のカーネル関数に基づく AFA（Exponential）がもっとも大きな差を示している一方、傾きが徐々に緩やかになっていく Concave 型の AFA は PNucl との差が相対的に小さい。また、線形に増加するカーネル関数を持つ Linear の乖離はこれらの中間となっている。このように、視覚的な観点からは必ずしも明らかではな

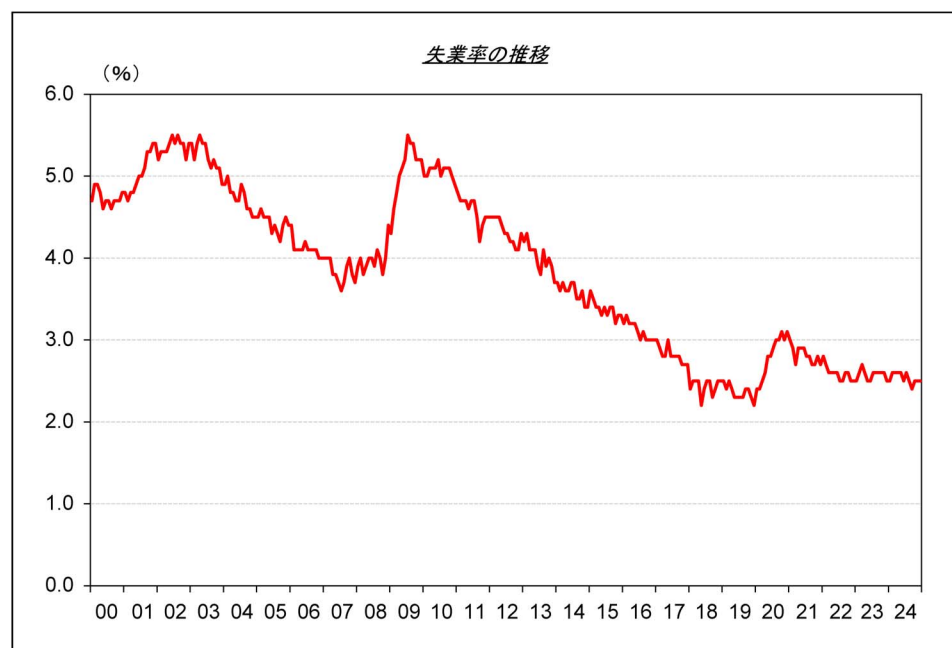
かったものの、表2の分析結果は、カーネル関数の形状の違いを背景として、異なるAFA間の中で分解パターンの違いが生じうることを示唆している。また、これら増加関数型のカーネル関数を持つAFAとSHAPとの平均的な乖離度は1~2%程度と、EN-ENSCおよびLSプレ仁をベースとしたAFAに比べるとはっきりと大きくなっている。

(2) わが失業率のAFA分解

Buckmann and Joseph (2023) は、米国失業率を対象に、勾配ブースティングやニューラルネットワークなどに基づく予測モデルを構築したうえで、SHAPを用いてこうした「ブラックボックスモデル」の可視化を行った(1節参照)。ここでは、わが国の失業率に対して、前節同様に8つのAFAを適用して、特徴量の貢献度についての分解パターンの違いが手法間でどの程度異なるのかを比較分析していく(図13)。ここでの着眼点は、前節の長期金利のケースと同様の傾向がここでも見出せるか否かである。

具体的には、被説明変数は、わが国失業率の2001年1月から2024年12月までの前年差である。Buckmann and Joseph (2023) を参考に、特徴量は以下の9つとし

図13 失業率の推移



た。前節同様、学習モデルは XGBoost であり、学習データには各特徴量を平均 0、分散 1 に標準化したものを用いている：

- ・ Lag_Unrate: 失業率（1 期ラグ）
- ・ 3MTB: 3 か月金利（1 期ラグ）
- ・ RealPerIncome: 実質個人所得（対数値、1 期ラグ）
- ・ IIP: 鉱工業生産（対数値、1 期ラグ）
- ・ Consumption: 個人消費（対数値、1 期ラグ）
- ・ NKY: 株価（対数値、1 期ラグ）
- ・ Loan: 企業向貸出前年比：1 期ラグ
- ・ CPI: 消費者物価指数前年比：1 期ラグ
- ・ M2: マネーサプライ（M2 前年比、1 期ラグ）

表 1 で示した 8 つの AFA にもとづく分解パターンの違いは、図 14、図 15 および図 16 に示されている（見やすさの観点から 2007 年以降の推移を示している）。赤実線が実際の失業率（前年差）の推移、紺色の実線が XGBoost モデルによる予測値、そして棒グラフが各特徴量に割り当てられた AFA の値である。図 14 から図 16 まで、合計 9 個のグラフが表示されているが、その構成は前節の長期金利のケースと同様である。すなわち、図 14 には線形回帰モデルを学習モデルとした時の AFA 分解（左パネル）と SHAP $\Psi_{\tau,j}^{SHAP}$ （中央パネル）、LS プレ仁に対応する AFA $\Psi_{\tau,j}^{PNucl}$ 、図 15 には左から残余均等配分解（ES）に対応する AFA $\Psi_{\tau,j}^{ES}$ 、逆残余均等配分解（ENSC）に対応する AFA $\Psi_{\tau,j}^{ENSC}$ 、そしてそれらを按分する解（ES-ENSC）に対応する AFA $\Psi_{\tau,j}^{ES-ENSC}$ が示されている。図 16 は、左パネルが既知の特徴量の数に関して線形に増加するカーネル関数に基づく AFA $\Psi_{\tau,j}^{LnK}$ 、中央パネルが指数関数的に増加するカーネル関数に基づく AFA $\Psi_{\tau,j}^{ExK}$ 、右パネルが対数関数的に増加するカーネル関数に基づく AFA $\Psi_{\tau,j}^{CrK}$ の分解パターンを示している。

これらの失業率に関する各 AFA の分解パターンの違いを視覚的にみると、前節の長期金利の分解パターンと同様の傾向がみられることが分かる。

すなわち、1 点目として、全体としてみると、AFA の間の分解パターンは類似している。長期金利のケースと比べると特徴量の数が多いため、視覚的なパターンの把握は必ずしも容易ではないが、例えば、2008 年頃の金融危機における失業率の上昇パターンとその後の低下局面を見ると、どの AFA においても、消費や株価（の低迷）が失業率を押し上げ、その後短期金利の低下が失業率を押し下げる姿となっている。また、2020 年のパンデミック時の失業率の上昇は、いずれの AFA においても、金融危機時に比べると相対的に IIP の寄与が大きくなっている。このように、分解パターンの全体感について、AFA の間に大きな違いはないといえる。

図 14 失業率の AFA 分解 (1)

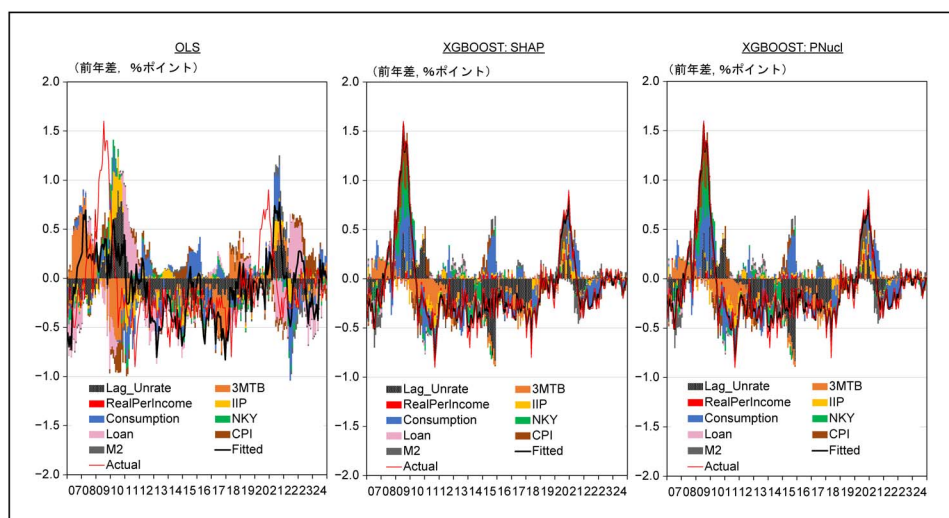
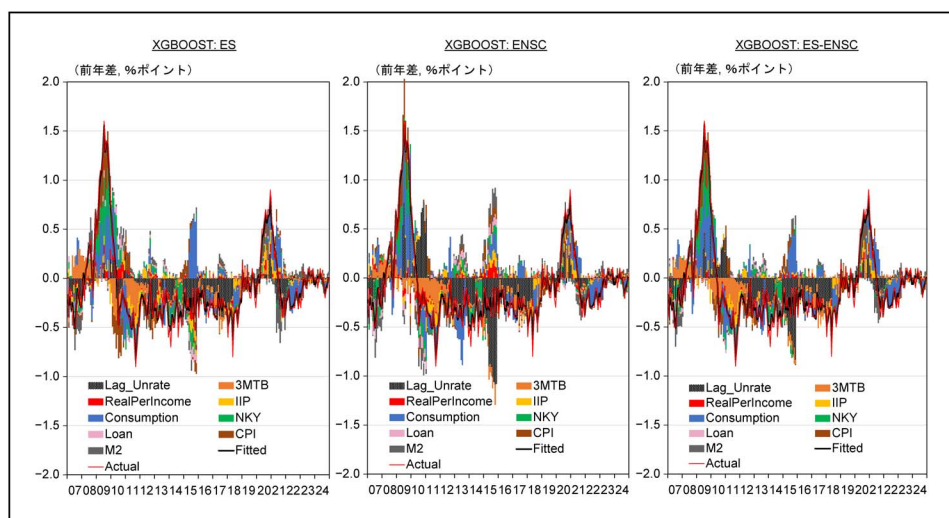
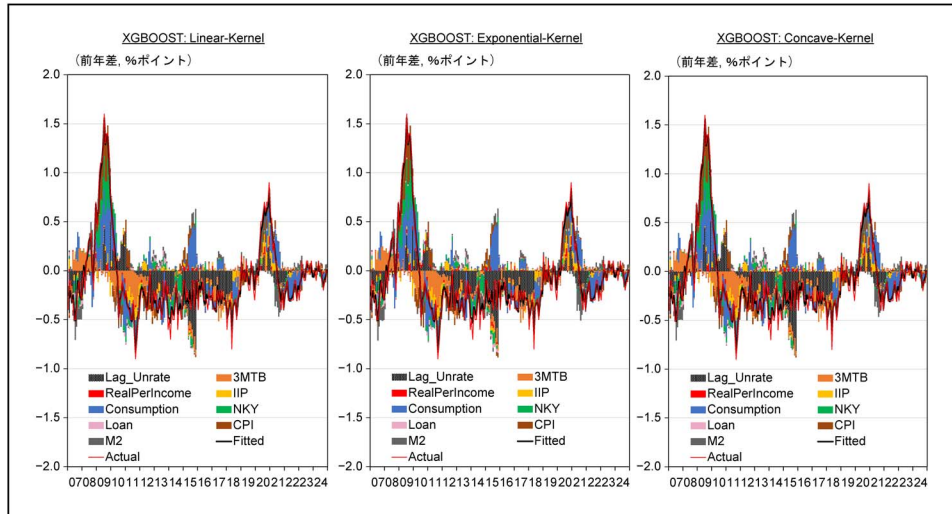


図 15 失業率の AFA 分解 (2)



2 点目に、長期金利のケース同様、SHAP と ES、および SHAP と ENSC の違いは、SHAP とそれ以外の AFA との違いと比べて相対的に大きい。例えば、失業率の自己ラグ (Lag_Unrate) に着目すると、SHAP と比べて、図 15 の左パネルで示された ES は自己ラグの寄与が全体的に小さく、一方で中央パネルで示された ENSC は自己ラグの寄与が SHAP より期間を通じて大きい。また、SHAP の分解パターンを見ると、2014 年から 2015 年にかけて、個人消費 (の弱さ) が失業率の押し上げ

図 16 失業率 AFA 分解 (3)



要因となっていたが、同時期の ES ではこの寄与がより大きく、一方で ENSC ではより小さくなっている。SHAP との分解パターンの違いは、特に ES と ENSC において視覚的にも捉えられるほど大きいという点は、長期金利のケースと同様の傾向である。

さらに、長期金利のケースで 3 点目に指摘した点も、共通の傾向として確認できる。すなわち、ES と ENSC を按分した AFA である ES-ENSC の分解パターンは、ES と ENSC の分解パターンを按分したイメージとなっており、かつ SHAP の分解パターンとの類似性が高い。ここでも、ES-ENSC が、SHAP を近似する AFA として優れた性質を持っていることが示唆されている。

4 点目の、既知の特徴量の数について増加関数となっているカーネル関数に基づく AFA や一定のカーネル関数を持つ LS プレ仁型 AFA についても、長期金利のケースと同様の傾向を指摘できる。図 16 をみると、カーネル関数の増加パターンの違いに由来する分解パターンの違いは、視覚的にはほとんど確認できず、かつ SHAP との類似性も高い（ただし、例えば、金融危機時における個人消費の押し上げ効果などについては、子細にみると違いが確認できる）。

次に、異なる AFA 間の違いを定量的に確認する。表 3 は表 2 と同じ形式のものである。すなわち、左下の赤字で示された領域のセルは、機械学習モデルを XGBoost としたときの、行と列に対応する AFA 間の差の絶対値を全ての特徴量と全ての観測値について平均した値であり、値が大きいほど 2 つの AFA が与える寄与度が乖離していることを意味する。右上の領域は学習モデルを回帰モデルにした場合であり、長期金利のケース同様、理論的に常にゼロとなる。以下では左下の領域に着目

表 3 AFA 間の相違度（失業率データ）

	左下：XGBoost					右上：Linear Model		
	SHAP	PNucl	ES	ENSC	ES-ENSC	Linear	Exponential	Concave
SHAP	—	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
PNucl	0.000509	—	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
ES	0.022005	0.022159	—	0.00000	0.00000	0.00000	0.00000	0.00000
ENSC	0.021563	0.021493	0.043513	—	0.00000	0.00000	0.00000	0.00000
ES-ENSC	0.001018	0.001527	0.021756	0.021756	—	0.00000	0.00000	0.00000
Linear	0.002386	0.002417	0.019754	0.023903	0.002778	—	0.00000	0.00000
Exponential	0.007188	0.007291	0.014890	0.028751	0.007146	0.004875	—	0.00000
Concave	0.001672	0.001657	0.020512	0.023145	0.002236	0.000761	0.005635	—

（注）8 種類の AFA のそれぞれの組み合わせについて、平均絶対差（ある観測値におけるある特徴量について、2 つの AFA の差の絶対値を計算し、それを全ての特徴量およびすべての観測値について平均したもの）を表示。右上の領域が学習モデルを線形回帰モデル、左下の領域が学習モデルを XGBoost にしたときの平均絶対差。

するが、前節でみた長期金利のケース同様、以下の傾向が確認できる。

1 点目は、SHAP と ES、および SHAP と ENSC との乖離幅である。定量的にみると、SHAP と ES、および SHAP と ENSC の特徴量ごとの平均的な乖離幅は 0.02% 程度である。一方、期間中の失業率前年差（絶対値）の平均は 0.32% であった。したがって、ある特定の月におけるある特徴量について、両者は平均的に 6% 程度 (0.02/0.32) 乖離するといえる。この乖離幅は、前節の長期金利のケースにおける乖離幅と概ね同じである。2 点目は、ES-ENSC および LS プレ仁をベースにした AFA は、SHAP と乖離が極めて小さく (0.001% 弱)、両者と SHAP との乖離幅はわずかに 0.0001% 未満である。3 点目は、カーネル関数の形状と AFA 間の乖離幅の関係であり、一定のカーネル関数を持つ LS プレ仁対比でみると、SHAP、Concave、Linear、Exponential の順に乖離幅が大きくなっていく。この点についても長期金利のケースで観察されたパターンと同じであり、カーネル関数の形状を反映した一般的な傾向であることが示唆されている。

4. まとめと結論

本稿では、SHAP およびその代替的な手法を対象に、理論面および実証面の分析を行った。まず、理論分析では、Hiraki, Ishihara and Shino (2024) (HIS 論文) をベースに、それを発展させる形で、既存の AFA の代表的な手法である SHAP と、その代替的な手法について検討を行った。特に、協力ゲーム理論におけるシャープ

レイ値以外の解概念に基づく AFA（特に残余均等配分解や逆残余均等配分解、あるいはそれらを按分した解に基づく AFA）を新たに提示し、それらの違いを考察した。次に、実証分析では、理論分析で取り上げた SHAP およびその代替的な手法を、わが国の長期金利および失業率に適用し、これら複数のケースから得られた共通の傾向を探ることで、SHAP および複数の代替的手法を、分解パターンの違いや計算コストの観点から暫定的に評価した。

分析の結果、まず、協力ゲーム理論の解概念である残余均等配分解に基づく AFA や逆残余均等配分解に基づく AFA については、SHAP との間で分解パターンの違いが明確に存在することが分かった。次に、それ以外の AFA の間では、視覚的に明確に確認できるほどの大きな差異はみられないものの、AFA 間の差異の大きさは、各 AFA に対応するカーネル関数の形状を反映したものとなっていることが明らかになった。さらに、残余均等配分解と逆残余均等配分解の按分として定義される AFA は、計算コストが小さく、かつ SHAP との差異が極めて小さいなど、SHAP を近似的かつ短期間に計算する手法として優れた性質を持つことが確認された。

SHAP を用いた意思決定は、経済・ファイナンスだけではなく、医療やマーケティングなど、社会経済活動の非常に重要な局面で急速に広がっていることから、その分解パターンのばらつき（不安定性）に対する理解を深めることや、代替的な手法の開発に取り組むことは、社会的にも非常に価値の高いテーマであるといえる。最後に、本分析で得られたインプリケーションを改めて挙げつつ、それらの点との関連で今後更なる分析が有益であると考えられるポイントを挙げて、本稿を結ぶこととする。

■既知の特徴量の数について増加するカーネル関数を持つ AFA の実データへの適用 本稿でも述べたように、カーネル関数を用いて機械学習モデルの解釈可能性を高める手法として、LIME（Local Interpretable Model-agnostic Explanations）がある。LIME を提示した Ribeiro *et al.*（2016）においては、カーネル関数は、「実際の分析対象となる観測値に近い摂動サンプルほど、より大きな重みを与えて評価する」という考えに沿ったものであるべきとされている。本稿で示した AFA のうち、この考え方をもっとも純粋に踏襲したものは指数関数的に増加するカーネル関数を持つ AFA（表 1 における $\Psi_{\tau,j}^{ExK}$ ）である（2 節 (5) 口. の議論も参照）。今回の実証分析では、SHAP とこれらの指数関数的に増加するカーネル関数を持つ AFA との違いは、視覚的に確認できるほどははっきりとしたものではなかった。一方で、これらの違いが、カーネル関数の形状を反映したものであるという実証的な結果は、今回示した理論分析の妥当性を裏付けるものであった。今後は、様々なデータにこれらの AFA を適用することで、SHAP とこれらの AFA の間で分解パターンに明確な違いが生じることがあるのか、また、そうであればそれは特にどのようなデータ特性において生じる傾向があるのか、などについての分析を進める必要がある。

■**計算コストへの対処** 本稿でも述べたように、SHAP は、計算コストが大きく、かつ特徴量の数が増えるほど指数関数的に増加することが知られている。Python などの機械学習用のソフトウェアにおいては、近似計算用のパッケージが提供されているが、それらを用いたとしてもかなりの計算時間を要するケースが頻繁に生じうる。本稿で示した AFA のうち、残余均等配分解と逆残余均等配分解を按分した AFA（表 1 における $\Psi_{\tau,j}^{ES-ENSC}$ ）は、計算コストが相対的に小さく、かつ SHAP との差異が極めて小さいなど、SHAP の近似計算アルゴリズムとして優れた性質を持つことが明らかになった。今回示した按分型の AFA や、同様の性質を持つ AFA を理論的に正当化し、かつ実装することができれば、学術的にも実務的にも大きな貢献となる²⁴。

■**様々なタイプのデータへの活用** 本稿では、実際の金融・経済への適用として、時系列データを取り上げた。もっとも、本稿で分析対象とした AFA は、時系列データだけではなく、あらゆるタイプのデータに適用できる。時系列データ以外への適用として興味深いものの 1 つとして、株式のクロス・セクショナル・リターンを挙げておく。Gu *et al.* (2020) は、米国の約 30,000 銘柄の株式データを対象に、決定木やニューラルネットワークといった機械学習モデルを用いて、時系列だけでなく、クロスセクショナルなリスクプレミアムのモデル学習と予測を行った。彼らは、回帰ベースの従来手法と比較し、機械学習モデルの予測精度が大幅に高まること、また、実際の投資という観点からも、シャープレシオの改善などの便益がもたらされることを示した。こうしたクロスセクショナルなデータを非線形の機械学習モデルで学習させ、それに基づく予測に対して本稿で用いた様々な AFA を適用し、説明可能性を高めることによって、CAPM やマルチファクターモデルといった従来の資産価格モデルでは捉えることのできない関係性を把握する事ができる可能性がある。

■**区間推定への拡張** Napolitano *et al.* (2024) は、SHAP を区間予測の機械学習モデルにまで拡張するために、協力ゲーム理論の 1 分野である、特性関数形ゲームに区間不確実性が存在すると仮定する「区間ゲーム」の枠組みを利用した。区間ゲームにおけるシャープレイ値（Interval Shapley value や Shapley mapping として定式化されている）については、Ishihara and Shino (2023, 2025) などによって既に研究が蓄積されており、これらを活かしつつ、本分析で示した AFA を区間予測の機械学習モデルに適用することは興味深いトピックである。

.....
 24 例えば、 $\Psi_{\tau,j}^{ES-ENSC}$ は、「予測の改善に全く貢献していない特徴量にも非ゼロの値を割り当ててしまう可能性がある」という点で、いわゆるシャープレイ値の公理系における null player property を満たさない。このため、(I) 今回の分析で得られた知見を踏まえつつ、理論的に null player property を満たすような計算コストの小さい AFA を理論的に示すことや、(II) null player property を満たさなくても、 $\Psi_{\tau,j}^{ES-ENSC}$ がそうした特徴量に与える非ゼロの値は無視できるほど小さいことを実証的に確かめていく、といったことが興味深いテーマとなりうる。

参考文献

- Ariza-Garzón, Miller, Janny, Javier Arroyo, Antonio Caparrini, and Maria-Jeses Segovia-Vargas. “Explainability of a Machine Learning Granting Scoring Model in Peer-to-peer Lending.” *IEEE Access* 8 (2020): 64873–90.
- Bali, Turan, G, Heiner Beckmeyer, Mathis Mörke, and Florian Weigert. “Option Return Predictability with Machine Learning and Big Data.” *Review of Financial Studies* 36, no. 9 (2023): 3548–602.
- Bianchi, Daniele, Matthias Buchner, and Andrea Tamoni. “Bond Risk Premiums with Machine Learning.” *Review of Financial Studies* 34, no. 2 (2021): 1046–89.
- Bluwstein, Kristina, Marcus Buckmann, Andreas Joseph, Sujit Kapadia, and Özgür Şimşek. “Credit Growth, the Yield Curve and Financial Crisis Prediction: Evidence from a Machine Learning Approach.” *Journal of International Economics* 145 (2023).
- Bracke, Phillipe, Anupam Datta, Carsten Jung, and Shayak Sen. “Machine Learning Explainability in Finance: An Application to Default Risk Analysis.” Bank of England Staff Working Paper 816 (2019).
- Buckmann, Marcus and Andreas Joseph. “An Interpretable Machine Learning Workflow with an Application to Economic Forecasting.” *International Journal of Central Banking* 19, no. 4 (2023): 449–522.
- Bussmann, Niklas, Paolo Giudici, Dimitori Marinelli, and Jochen Papenbrock. “Explainable Machine Learning in Credit Risk Management.” *Computational Economics* 57, no. 1 (2021): 203–16.
- Chen, Hugh, Joseph. D. Janizek, Scott M Lundberg, and Su-In Lee. “True to the Model or True to the Data?” *arXiv* (2020). <https://arxiv.org/abs/2006.16234>
- Chen, Luyang, Markus Pelger, and Jason Zhu. “Deep Learning in Asset Pricing.” *Management Science* 70, no. 2 (2024): 714–50.
- Condevaux, Charles, Sébastien Harispe, and Stéphane Mussard. “Fair and Efficient Alternatives to Shapley-based Attribution Methods.” In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2022.
- Coulombe, Phillipe, Goulet, Maxime Leroux, Dailbor Stevanovic, and Stéphane Surprenant. “How is Machine Learning Useful for Macroeconomic Forecasting?” *Journal of Applied Econometrics* 37, no. 5 (2022): 920–64.
- Demirbaga, Umit, and Yue Xu. “Empirical Asset Pricing Using Explainable Artificial Intelligence.” SSRN Working Paper, 2024.
- Dragan, Irinel, Theo Driessen, and Yukihiro Funaki. “Collinearity between the Shapley Value and the Egalitarian Division Rules for Cooperative Games.” *OR SPEKTRUM* 18 (1996): 97–105.

- Driessen, Theo and Yukihiro Funaki. "Coincidence of and Collinearity between Game Theoretic Solutions." *OR SPEKTRUM* 13 (1991): 15–30.
- Gu, Shihao, Bryan Kelly, and Dacheng Xiu. "Empirical Asset Pricing via Machine Learning." *Review of Financial Studies* 33, no. 5 (2020): 2223–73.
- Hiraki, Kazuhiro, Shinichi Ishihara, and Junnosuke Shino. "Alternative Methods to Shap Derived from Properties of Kernels: A Note on Theoretical Analysis." In *Proceedings of the International Conference on Big Data*, 2024.
- Ishihara, Shinichi and Junnosuke Shino. "Some Properties of Interval Shapley Values: An Axiomatic Analysis." *Games* 14, no. 3 (2023): 1–10.
- Ishihara, Shinichi and Junnosuke Shino. "An Axiomatization of the Shapley Mapping Using Strong Monotonicity in Interval Games." *Annals of Operational Research* 345, no. 1 (2025): 147–68.
- Jabeur, Sami, Ben, Salma Mefteh-Wali, and Jean-Laurent Viviani. "Forecasting Gold Price with the Xgboost Algorithm and Shap Interaction Values." *Annals of Operational Research* 334, no. 1 (2024): 679–99.
- Janzing, Dominik, Lenon Minorics, and Patrick Bloebaum. "Feature Relevance Quantification in Explainable AI: A Causal Problem." *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR 108 (2020): 2907–16.
- Jethani, Neil, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. "Fastshap: Real-time Shapley Value Estimation." *International Conference on Learning Representations*, 2022.
- Kongo, Takumi. "Equal Support from Others for Unproductive Players: Efficient and Linear Values That Satisfy the Equal Treatment and Weak Null Player Out Properties for Cooperative Games." *Annals of Operations Research* 338, no. 2 (2024): 973–89.
- Loh, Hui Wen, Chui Ping Ooi, Silvia Seoni, Prabal Datta Barua, Fillipo Molinari, and Rajendra Acharya. "Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011–2022)." *Computer Methods and Programs in Biomedicine* 226 (2022): 107161.
- Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Roint Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (2020): 56–67.
- Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017).
- Lundberg, Scott M, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim,

- and Su-In Lee. “Explainable Machine-learning Predictions for the Prevention of Hypoxaemia During Surgery.” *Nature Biomedical Engineering* 2 (2018): 749–60.
- Misheva, Branka, Hadji, Joerg Osterrieder, Ali Hirs, Onkar Kulkarni, and Stephan Fung Lin. “Explainable Ai in Credit Risk Management.” arXiv preprint arXiv.2103.00949. <https://arxiv.org/abs/2103.00949>.
- Molnar, Christoph. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. 3rd ed., Lulu, 2025.
- Napolitano, Davide, Lorenzo Vaiani, and Luca Cagliero. “Efficient Neural Network-based Estimation of Interval Shapley Values.” *IEEE Transactions on Knowledge and Data Engineering* 36, no. 12 (2024): 8108–19.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?: Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44, New York, August 13–17, 2016.
- Ruiz, Luis M., Federico Valenciano, and Jose M. Zarzuelo. “The Least Square Prenucleolus and the Least Square Nucleolus. Two Values for TU Games Based on the Excess Vector.” *International Journal of Game Theory* 25, no. 1 (1996): 113–34.
- Ruiz, Luis M., Federico Valenciano, and Jose M. Zarzuelo. “The Family of Least Square Values for Transferable Utility Games.” *Games and Economic Behavior* 24, no. 1–2 (1998): 109–30.
- Shapley, Lloyd S. “A Value for n-Person Games.” *Annals of Mathematics Studies* 28 (1953): 307–18.
- Yeo, Wei Jie, Wihan Van Der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. “A comprehensive Review on Financial Explainable AI.” *Artificial Intelligence Review* (2025): article 189.
- 和泉 潔、「金融分野における因果推論の展開—統計的手法・因果 AI・自然言語処理の三潮流とその展望—」、SBI Research Review 8, SBI 金融経済研究所、2025 年
- 大坪直樹・中江俊博・深沢裕太・豊岡 祥・坂元哲平・佐藤 誠・五十嵐健太・市原大暉・堀内新吾、『XAI（説明可能な AI）—そのとき人工知能はどう考えたか—』、リックテレコム、2021 年
- 平木一浩・石原慎一・篠 潤之介、「カーネル関数に基づく SHAP とその代替的な手法の比較分析」、『人工知能学会全国大会論文集』、一般社団法人人工知能学会、2025 年
- 日本銀行、「より効果的で持続的な金融緩和を実施していくための点検」、日本銀行、2021 年 (<https://www.boj.or.jp/mopo/mpmdeci/transparency>)
- 森 いづみ・中村俊文・乗政喜彦、「グローバルにみた感染症の家計等の行動への

影響：機械学習によるアプローチ」、日銀レビュー 2021-J-5、日本銀行、2021 年
森下光之助、『機械学習を解釈する技術: 予測力と説明力を両立する実践テクニック』、リクテレコム、2021 年
金田規靖・木全友則・平木一浩・松栄共紘、「SHAP を用いた機械学習モデルの解釈
—原油価格の変動要因分析を例に—」、『日本銀行金融研究所ファイナンス・ワー
クショップの模様』、日本銀行金融研究所、2022 年 ([https://www.imes.boj.or.jp/jp/
conference/finance/2022_Fworkshop.html](https://www.imes.boj.or.jp/jp/conference/finance/2022_Fworkshop.html))
鷺見和昭、「通貨オプション市場における投資家センチメントの要因分析：機械学
習アプローチ」、日本銀行ワーキングペーパー、No. 20-J-8、日本銀行、2020 年

補論 1. 所与の AFA に対応するカーネル関数の導出方法

2 節 (4) および 2 節 (5) では、所与のカーネル関数に対して、(22) 式を用いることで AFA を導くという手法を用いて、複数の AFA を提案した（カーネル関数 → AFA）。一方で、ある特定の条件を持つ²⁵ 所与の AFA（またはそれに対応した協力ゲームの解概念）に対して、それに対応するカーネル関数を導出することも可能である（AFA → カーネル関数）。以下では、この点について説明する。

以下では、AFA に対応する協力ゲーム解の観点から、また、単純化のために $(\emptyset) = 0$ を仮定して議論を進める。特性関数形ゲーム (N, v) における、全体合理性を満たす解 $\theta(v) = (\theta_1(v), \dots, \theta_n(v))$ が²⁶、重み $w_i(S)$ を用いて以下のように表せるとする。

$$\theta_i(v) = \sum_{S \in 2^N \setminus \emptyset} w_i(S) v(S). \quad (37)$$

ただし、 $w_i(S)$ は、

$$w_i(S) = \begin{cases} M(|S|) & \text{if } i \in S \\ m(|S|) & \text{if } i \notin S \end{cases} \quad (38)$$

であり、 $M(|S|) \geq 0$ 、 $m(|S|) \leq 0$ 、 $\sum_{|S|=1}^{n-1} (M(|S|) - m(|S|)) \cdot (n-2)C_{|S|-1} = 1$ を満たすものとする。このとき、解 θ に対応するカーネル関数 $\pi(S)$ は以下で表現できる²⁷。

$$\pi(S) = M(|S|) - m(|S|). \quad (40)$$

.....
25 以下の (37) 式で表現できる解概念。協力ゲーム理論においては、代表的な解概念はこの和の形で表現できる。

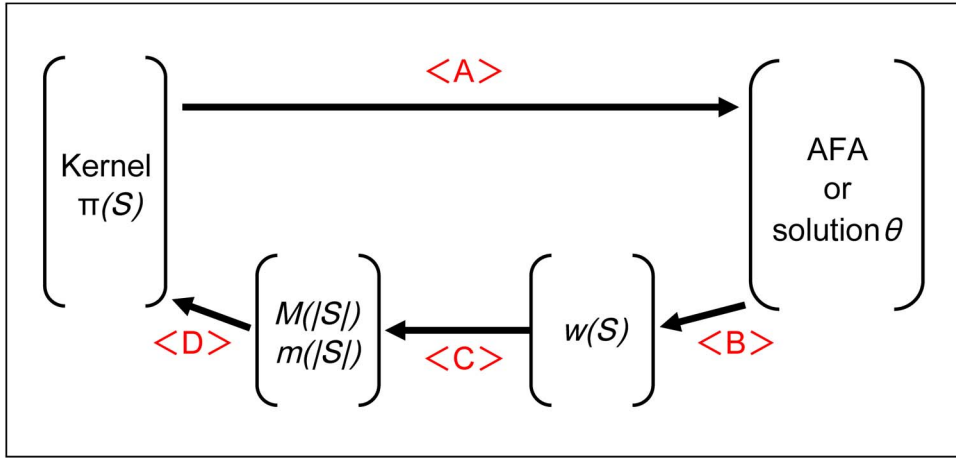
26 $\sum_{i \in N} \theta_i(v) = v(N)$ を満たすとき、解 $\theta(v)$ は全体合理性を満たすと言う。

27 証明の概略は以下の通り：

$$\begin{aligned} \theta_i(v) - \theta_j(v) &= \sum_{S \in 2^N \setminus \emptyset} w_i(S) v(S) - \sum_{S \in 2^N \setminus \emptyset} w_j(S) v(S) = \sum_{S \in 2^N \setminus \emptyset} [w_i(S) - w_j(S)] v(S) \\ &= \sum_{\substack{S \in 2^N \setminus \emptyset \\ \text{with } i, j \in S}} [w_i(S) - w_j(S)] v(S) + \sum_{\substack{S \in 2^N \setminus \emptyset \\ \text{with } i \in S \text{ and } j \notin S}} [w_i(S) - w_j(S)] v(S) \\ &\quad + \sum_{\substack{S \in 2^N \setminus \emptyset \\ \text{with } i \notin S \text{ and } j \in S}} [w_i(S) - w_j(S)] v(S) + \sum_{\substack{S \in 2^N \setminus \emptyset \\ \text{with } i, j \notin S}} [w_i(S) - w_j(S)] v(S) \\ &= \sum_{\substack{S \in 2^N \setminus \emptyset \\ \text{with } i \in S \text{ and } j \notin S}} [M(|S|) - m(|S|)] v(S) - \sum_{\substack{S \in 2^N \setminus \emptyset \\ \text{with } i \notin S \text{ and } j \in S}} [M(|S|) - m(|S|)] v(S) \\ &= \sum_{S \subset N \setminus \{i, j\}} \left([M(|S|) - m(|S|)] v(S \cup \{i\}) - [M(|S|) - m(|S|)] v(S \cup \{j\}) \right) \end{aligned} \quad (39)$$

となり、これは (16) 式と同じ形式である。ゆえに、 $M(|S|) - m(|S|)$ がカーネル関数となる。

図 17 本分析で示したカーネル関数と AFA の関係



すなわち、(37) 式で表すことのできる解 (=AFA) $\theta(v)$ に対応するカーネル関数は、(40) 式によって求めることができる。

ここで AFA とカーネル関数の関係を、図 17 でまとめておこう。所与のカーネル関数に対して AFA を導出するという、図中の<A>で示されている関係は、(22) 式によって表現された。一方で、所与の AFA または協力ゲーム理論解に対しては、その解を重み $w(S)$ を用いて表し（図中および (37) 式）、 $w(S)$ を $M(|S|)$ および $m(|S|)$ で表し（図中<C>および (38) 式）、それを用いて $\pi(S)$ が導出される（図中<D>および (40) 式）。

以下では、既存の協力ゲーム理論解であるシャープレイ値および ENSC に対応するカーネル関数を導出する過程を示しておく。

■シャープレイ値に対応するカーネル関数の導出 シャープレイ値 $\phi(v)$ は、

$$\begin{aligned}\phi_i(v) &= \sum_{S \subseteq N \setminus i} \frac{|S|!(n-|S|-1)!}{n!} (v(S \cup \{i\}) - v(S)) \\ &= \sum_{S \in 2^N \setminus \emptyset: i \in S} \frac{(|S|-1)!(n-|S|)!}{n!} v(S) + \sum_{S \in 2^N \setminus \emptyset: i \notin S} -\frac{|S|!(n-|S|-1)!}{n!} v(S)\end{aligned}$$

と表すことができる。したがって、

$$M(|S|) = \frac{(|S|-1)!(n-|S|)!}{n!}, \quad m(|S|) = -\frac{|S|!(n-|S|-1)!}{n!}$$

とすれば、シャープレイ値は (37) の形式で表現することができる。このとき、

$$M(|S|) - m(|S|) = \frac{(|S|-1)!(n-|S|)!}{n!} + \frac{|S|!(n-|S|-1)!}{n!} = \frac{(|S|-1)!(n-|S|-1)!}{(n-1)!}$$

であり、

$$\begin{aligned} \sum_{s=1}^{n-1} (M(|S|) - m(|S|)) \cdot \binom{n-2}{|S|-1} &= \sum_{|S|=1}^{n-1} \frac{(|S|-1)!(n-|S|-1)!}{(n-1)!} \cdot \frac{(n-2)!}{(|S|-1)!(n-|S|-1)!} \\ &= \sum_{|S|=1}^{n-1} \frac{1}{n-1} = 1 \end{aligned}$$

を満たす。よって、シャープレイ値に対応するカーネル関数 $\pi(S)$ は

$$\begin{aligned} \pi(S) &= M(|S|) - m(|S|) = \frac{(|S|-1)!(n-|S|-1)!}{(n-1)!} = \frac{n \cdot |S| \cdot (n-|S|) \cdot (|S|-1)! \cdot (n-|S|-1)!}{n \cdot |S| \cdot (n-|S|) \cdot (n-1)!} \\ &= \frac{|S|! \cdot (n-|S|)! \cdot n}{n! \cdot |S| \cdot (n-|S|)} = \frac{n}{{}_nC_{|S|} \cdot |S| \cdot (n-|S|)} \end{aligned}$$

となり、(24) 式に一致する（最後の等号は ${}_nC_{|S|} = n!/[|S|! \cdot (n-|S|)!]$ による）。

■ ENSC に対応するカーネル関数の導出 ENSC $e(v)$ は、

$$\begin{aligned} e_i(v) &= v(N) - v(N \setminus i) + \frac{\sum_{j \in N} (v(N) - v(N \setminus j))}{n} \\ &= \sum_{S \in 2^N \setminus \emptyset: i \in S, |S|=n-1} \frac{1}{n} v(S) + \left(-\frac{n-1}{n} v(N \setminus i) \right) + \frac{1}{n} v(N) \end{aligned}$$

となるので、 $M(s), m(s)$ を以下のように定義すると、(37) の形式で表現することができる：

$$M(s) = \begin{cases} 0 & \text{if } s \leq n-2 \\ \frac{1}{n} & \text{otherwise} \end{cases}$$

$$m(s) = \begin{cases} 0 & \text{if } s \leq n-2 \\ -\frac{n-1}{n} & \text{otherwise} \end{cases}$$

このとき、

$$M(s) - m(s) = \begin{cases} 0 & \text{if } s \leq n-2 \\ 1 & \text{otherwise} \end{cases}$$

であり、

$$\sum_{s=1}^{n-1} (M(s) - m(s)) \cdot \binom{n-2}{s-1} = \sum_{s=1}^{n-2} 0 \cdot \frac{(n-2)!}{(s-1)!(n-s-1)!} + 1 = 1$$

を満たす。よって、ENSCに対応するカーネル関数 $\pi(S)$ は以下ようになる。

$$\pi(S) = M(s) - m(s) = \begin{cases} 0 & \text{if } s \leq n-2 \\ 1 & \text{otherwise} \end{cases}$$

補論 2. 本分析で取り上げた AFA の一般的性質

ここでは、表 1 で示した 8 つの AFA、あるいはその一般的表現である (22) 式で表される AFA について、特に学習モデル f との関係を中心に、いくつかの特徴を示す。具体的には、まず、学習モデルが特徴量に関して加法的であれば、(22) 式で定義された AFA はカーネル関数の形状に関わらず常に同一になることを示す。このことは、本稿で扱った AFA が異なる分解パターンを示すのは、学習モデルが少なくとも非加法的である必要があることを示している。次に、加法的な学習モデルの特別なケースとして線形回帰を取り上げ、(22) 式で定義された AFA は、カーネル関数の形状に関わらず、常に線形回帰モデルにおける回帰パラメータから得られる AFA と一致することを示す。これは、本稿で示した AFA はすべて線形回帰モデルの要因分解の一般化であることを意味し、表 1 の各 AFA に一定の妥当性を与える結果であるといえる。

■学習モデルが加法的である場合の AFA の一致性

特徴 4.1 学習モデル f が、以下の通り、特徴量について加法的であるとする：

$$Y = f(X) = \sum_{j=1}^n f_j(X_j). \quad (41)$$

このとき、(22) 式で与えられる Ψ_τ^{AFA} はカーネル関数の形状に関わらず同一であり、 $\Psi_{\tau,j}^{AFA} = v_\tau(N) - v_\tau(N \setminus \{j\})$ となる。

特徴 4.1 を証明するため、以下の定義 4.1 および補題 4.1 を挙げておく。

定義 4.1 特性関数形ゲーム (N, v) が以下の条件を満たすとき、 (N, v) を加法的（特性関数形ゲーム）と呼ぶ。

$S \cap T = \emptyset$ を満たす任意の $S, T \in 2^N$ について、

$$v(S \cup T) - v(\emptyset) = \{v(S) - v(\emptyset)\} + \{v(T) - v(\emptyset)\}. \quad (42)$$

特性関数形ゲーム (N, v) が加法的であるとき、単に v が加法的であるとも言う。

なお、(42) は以下の条件と同値である：

$$\exists a = (a_1, \dots, a_n) \in R^n, \forall S \in 2^N, v(S) - v(\emptyset) = \sum_{j \in S} a_j. \quad (43)$$

補題 4.1 学習モデル f が特徴量 X_j について加法的であれば、(1) 式に基づいて作られる特性関数形ゲーム (N, v_τ) は加法的である。

補題 4.1 の証明 学習モデル f が特徴量 X_j について加法的であれば、(41) 式および (1) 式より、 $v_\tau(S) = E[f(x_{\tau,S}, X_{N \setminus S})] = \sum_{k:k \in S} f_k(x_{\tau,k}) + \sum_{l:l \notin S} E[f_l(X_l)]$ が成り立つ。 $a_j = f_j(x_{\tau,j}) - E[f_j(X_j)]$ とすると、任意の $S \in 2^N$ について、以下が成り立つ：

$$\begin{aligned} v_\tau(S) - v_\tau(\emptyset) &= \left(\sum_{k:k \in S} f_k(x_{\tau,k}) + \sum_{l:l \notin S} E[f_l(X_l)] \right) - \sum_{j=1}^n E[f_j(X_j)] \\ &= \sum_{j:j \in S} (f_j(x_{\tau,j}) - E[f_j(X_j)]) = \sum_{j \in S} a_j. \blacksquare \end{aligned}$$

特徴 4.1 の証明 学習モデル f が特徴量 X_j について加法的であるとする。補題 4.1 より、特性関数形ゲーム (N, v_τ) は加法的である。したがって、 $a_j = f_j(x_{\tau,j}) - E[f_j(X_j)]$ とおくと、(22) 式より以下が成り立つ：

$$\begin{aligned} \Psi_{\tau,i}^{AFA} - \Psi_{\tau,j}^{AFA} &= \sum_{S \subseteq N \setminus \{i,j\}} \left(\pi_{x_\tau}(S \cup \{i\}) \cdot v_\tau(S \cup \{i\}) - \pi_{x_\tau}(S \cup \{j\}) \cdot v_\tau(S \cup \{j\}) \right) \\ &= \sum_{S \subseteq N \setminus \{i,j\}} \left(\pi_{x_\tau}(S \cup \{i\}) \cdot \sum_{k \in S \cup \{i\}} a_k - \pi_{x_\tau}(S \cup \{j\}) \cdot \sum_{k \in S \cup \{j\}} a_k \right) = a_i - a_j. \end{aligned} \quad (44)$$

(N, v_τ) は加法的なので、(43) 式より $v(N) - v(\emptyset) = \sum_{j \in N} a_j$ 、すなわち $\sum_{j \in N} \Psi_{\tau,j}^{AFA} = \sum_{j \in N} a_j$ であり、かつ (44) 式より、 $\Psi_{\tau,j}^{AFA} = a_j$ 。一方、以下が成り立つことと；

$$v_\tau(N) - v_\tau(N \setminus \{j\}) = \sum_{k=1}^n f_k(x_{\tau,k}) - \left(\sum_{k:k \neq j} f_k(x_{\tau,k}) + E[f_j(X_j)] \right) = f_j(x_{\tau,j}) - E[f_j(X_j)] = a_j,$$

$$\Psi_{\tau,j}^{AFA} = a_j \text{ より、} \Psi_{\tau,j}^{AFA} = v_\tau(N) - v_\tau(N \setminus \{j\}). \blacksquare$$

■**学習モデルが線形回帰モデルの場合** 次に、学習モデル f が、以下の通り、線形回帰モデルであるとする：

$$Y = f(X) = \beta_0 + \sum_{j=1}^n \beta_j X_j. \quad (45)$$

学習モデルが線形回帰モデルであれば、回帰パラメータを用いた AFA が可能である。すなわち、(45) 式より、 $v_\tau(S) = E[f(x_{\tau,S}, X_{N \setminus S})] = \beta_0 + \sum_{k:k \in S} \beta_k x_{\tau,k} + E[\sum_{l:l \notin S} \beta_l X_l]$ 。

したがって、 $v_{\tau}(N) - v_{\tau}(\emptyset) = \sum_{j=1}^n \beta_j (x_{\tau,j} - E[X_j])$ 。ここで $\Psi_{\tau}^{LM} = (\Psi_{\tau,1}, \dots, \Psi_{\tau,n})$ を；

$$\Psi_{\tau,j}^{LM} = \beta_j (x_{\tau,j} - E[X_j]), \quad (46)$$

と定義すると、 Ψ_{τ}^{LM} は AFA となっている ((3) 式参照)。

特徴 4.2 学習モデル f が線形回帰モデルのとき、 $\Psi_{\tau}^{LM} = \Psi_{\tau}^{AFA}$ (ただし Ψ_{τ}^{AFA} は (22) 式で定義したもの)。

SHAP については、 f が線形回帰であれば Ψ_{τ}^{LM} と一致することが知られている ($\Psi_{\tau}^{LM} = \Psi_{\tau}^{SHAP}$ 。例えば森下 (2021) を参照)。特徴 4.2 は、これと同じ性質が (22) 式で定義されているすべての AFA で成り立つことを意味している。これは、われわれが本稿で提示した AFA を実際のデータに適用して分析することの正当性を少なくとも部分的に示すものであるといえる。

特徴 4.2 の証明 f が線形回帰モデルであれば、 f は加法的である。特徴 4.1 の証明で示した通り、 f が加法的であれば $\Psi_{\tau,j}^{AFA} = a_j$ 。一方、線形回帰モデルであれば、 $a_j = f_j(x_{\tau,j}) - E[f_j(X_j)] = \beta_j (x_{\tau,j} - E[X_j])$ 。(46) 式より、 $\Psi_{\tau,j}^{LM} = \Psi_{\tau,j}^{AFA}$ 。■

補論 3. 金価格の AFA 分解

Jabeur *et al.* (2024) は、金価格を 6 つの機械学習モデル (Linear regression, Neural networks, Random forest, Light gradient boosting machine, CatBoost algorithm, XGBoost algorithm) を用いて予測した後、これに SHAP を適用して比較分析を行い、XGBoost とそれに対する SHAP の適用が分析上有効であることを主張した。さらに、平木ほか [2025] は、Jabeur *et al.* (2024) をベースにしつつ、適用する AFA を表 1 における *ES*、*LnK*、*ExK*、*CvK* にまで拡張し、分解パターンの違いを考察した。この補論では、同じデータを用いつつ、適用する AFA を *ES* および *ES-ENSC* にまでさらに拡張し、3 節で得られた洞察や傾向が、金価格のケースでも明確に観察されることを示す²⁸。

具体的には、分析対象は 1998 年 1 月から 2023 年 12 月までの、ドル建て金価格 (1 オンスあたり、月次) データである (図 18)。学習モデルは XGBoost、特徴量は Jabeur *et al.* (2024) に沿って以下の 6 つとする。各特徴量を平均 0、分散 1 に標準化して学習データとする点も 3 節と同様である。

- ・ Silverprice: 銀価格 (ドル／オンス、1 期ラグ)
- ・ Oilprice: 原油価格 (ドル／バレル、1 期ラグ)
- ・ USD_EUR: ユーロ対ドルレート (1 期ラグ)

.....
28 3 節同様公表データおよび Bloomberg から取得したデータ等を用いた。

図 18 金価格の推移



- ・ USD_CNY: 人民元対ドルレート
- ・ CPI: 米国消費者物価指数（指数、レベル、1 期ラグ）
- ・ SP500: 米国株価（SP500、ドル、1 期ラグ）

図 19、図 20 および図 21 に示されている 9 つのパネルの構成は、3 節で示したものと同様であり、また、そこから 3 節と同じ傾向を見出すことができる。すなわち、1 点目に、全体として各 AFA が示す分解パターンは概ね似通っている。2 点目に、SHAP と ES、SHAP と ENSC は、視覚的にも確認できる程度の分解パターンの違いがある。3 点目は、ES と ENSC を按分した AFA である ES-ENSC の分解パターンは SHAP のそれと極めて似たものとなっている。4 点目は、特徴量の数について増加関数となっているカーネル関数に基づく AFA 間の違いは、視覚的に確認できるほど大きくはない。

次に、3 節同様、表 4 の左下の領域で、学習モデルを XGBoost としたときの AFA 間の違いを定量的に確認する。まず、SHAP と ES、および SHAP と ENSC の平均的な乖離幅は 1.8USD/OZ 程度である。一方、期間中の金価格前月差の平均は 28.8USD/OZ。すなわち、両者は平均的に 6% 程度乖離することを意味する。この乖離幅は、3 節で示した長期金利および失業率のケースと概ね等しい。6% という値が理論的に求まるわけではないが、金融・経済データに今回提示した AFA を適

図 19 金価格の AFA 分解 (1)

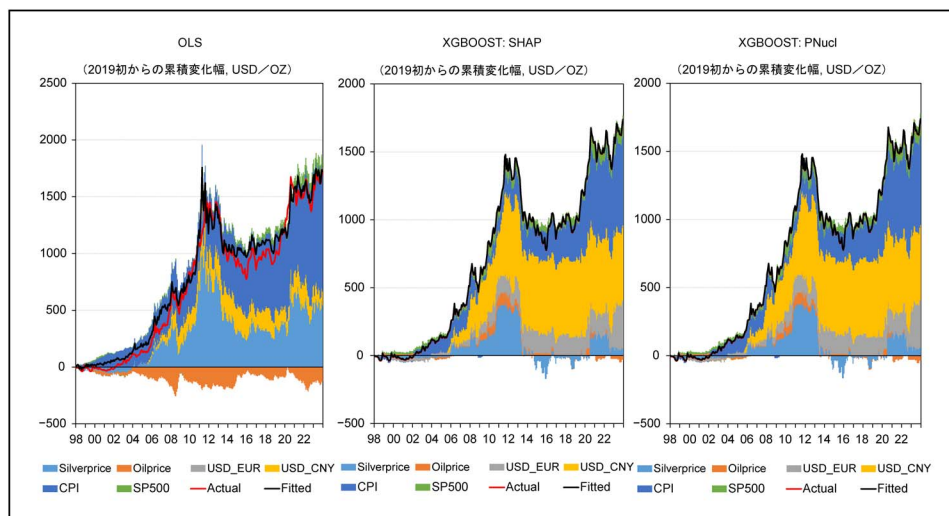
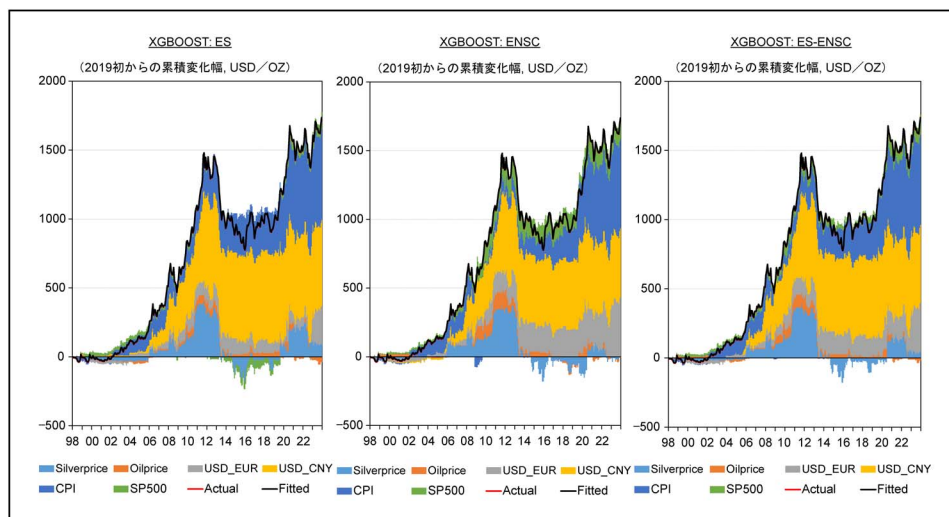


図 20 金価格の AFA 分解 (2)



用したときの乖離幅の最大値のイメージとして、1つの目安となるかもしれない。

また、EN-ENSC および LS プレ仁をベースとした AFA と SHAP の乖離幅が相対的に小さい点、また、特徴量の数について増加関数となっているカーネル関数に基づく AFA のうち、Exponential が LS プレ仁との乖離がもっとも大きいといった点も、3 節と同様であり、本稿で提示した AFA を適用した際にははっきりとした傾向として捉えることができる。

図 21 金価格の AFA 分解 (3)

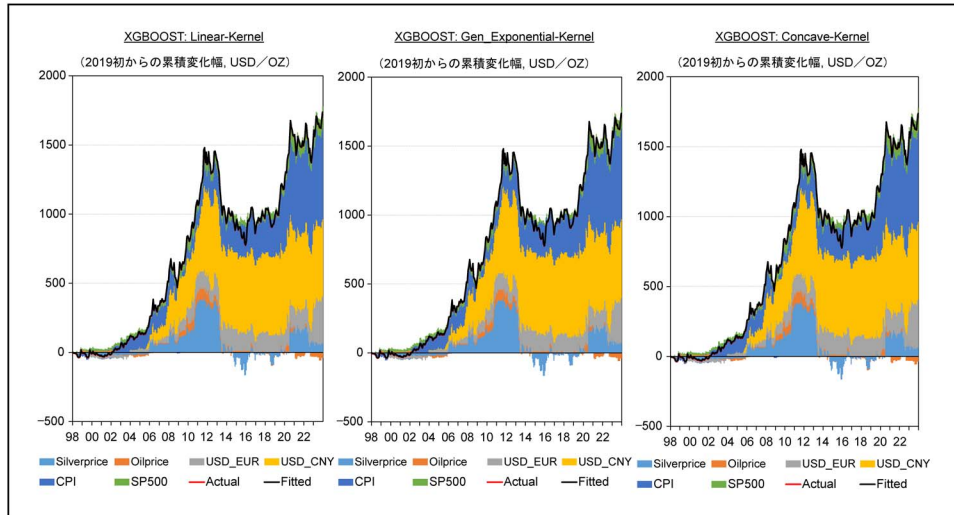


表 4 AFA 間の相違度 (金価格データ)

	左下 : XGBoost					右上 : Linear Model		
	SHAP	PNucl	ES	ENSC	ES-ENSC	Linear	Exponential	Concave
SHAP	—	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
PNucl	0.042518	—	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
ES	1.799410	1.808279	—	0.00000	0.00000	0.00000	0.00000	0.00000
ENSC	1.780546	1.778156	3.577296	—	0.00000	0.00000	0.00000	0.00000
ES-ENSC	0.085043	0.127561	1.788648	1.788648	—	0.00000	0.00000	0.00000
Linear	0.297137	0.298271	1.512723	2.075696	0.340193	—	0.00000	0.00000
Exponential	0.593852	0.598276	1.213298	2.374398	0.600511	0.300027	—	0.00000
Concave	0.209885	0.206226	1.604352	1.983996	0.267145	0.092083	0.392110	—

(注) 8 種類の AFA のそれぞれの組み合わせについて、平均絶対差 (ある観測値におけるある特徴量について、2 つの AFA の差の絶対値を計算し、それを全ての特徴量およびすべての観測値について平均したもの) を表示。右上の領域が学習モデルを線形回帰モデル、左下の領域が学習モデルを XGBoost にしたときの平均絶対差。

金融研究所の概要

金融研究所（Institute for Monetary and Economic Studies, IMES）は1982年10月に日本銀行創立百周年を記念して、日本銀行の内部組織の1つとして設立されました。その主な目的は、（1）金融経済の理論、制度、歴史に関する基礎的な研究の充実を図り、日本銀行の政策の適切な運営に役立てること、（2）学界等との交流を促進すること、（3）外部の研究活動の便宜に資する各種情報、資料等を公に提供することです。

主な活動

研究活動

- ・金融経済の基本的問題に関する理論・実証的研究。
- ・金融関連の制度基盤（法律・会計・中央銀行制度等）に関する研究。
- ・金融関連の情報技術に関する研究。
- ・金融経済の歴史に関する研究。
- ・金融経済に関する歴史的資料の収集・保存・公開（アーカイブ・貨幣博物館）。

顧問および客員研究員の招へい

- ・国内外の有力学者若干名に研究所顧問（非常勤）を委嘱し、研究所の運営、研究活動の進め方、内外学界との交流等に関する助言を受けています。
- ・国内外より学者数名を客員研究員として招へいし、研究活動の強化を図っています。

会議の開催

- ・海外中央銀行、国際機関の研究者や内外の著名学者を交えた会議を開催。
- ・外部の研究者、専門家を交えた研究会、セミナー等を適宜開催。

刊行物の発行

金融研究、Monetary and Economic Studies、IMES Discussion Paper Series、日本銀行の機能と業務、国際コンファランス議事録等を公刊（刊行物の一覧は後掲）。

貨幣博物館

国内を中心とした貴重な貨幣や貨幣にかかわる資料を貨幣博物館に展示するとともに、貨幣に関する歴史資料をホームページ等で公開（博物館の案内は後掲）。

歴史的公文（歴史的文書）の公開

日本銀行に関連する歴史的資料のうち、歴史的価値を有するものについて、日本銀行金融研究所アーカイブで保管・整理し、一般に公開（アーカイブの案内は後掲）。

研究の委託

人員面、資料面の制約等から所内での研究が困難なテーマについて外部の学者等に研究を委託。

主な刊行物

金融研究

邦文機関誌、年4回程度発行。金融研究所の研究論文や各種ワークショップの様様、研究会報告等を公表。

Monetary and Economic Studies

英文機関誌。『金融研究』と同様、金融研究所の研究論文等を公表（『金融研究』掲載論文の英訳のほか、英文オリジナル論文を含む）。

IMES Discussion Paper Series (E-Series：英語版、J-Series：日本語版)

金融研究所スタッフおよび外部研究者による研究成果の未定稿を随時公表。学界、金融機関、関係者等から、広くコメントを求めることを目的としている。

国際コンファランス議事録

- ・ *Growth, Integration, and Monetary Policy in East Asia* (*Monetary and Economic Studies*, Vol. 25, No. S-1, 2007年)
- ・ *Financial Markets and the Real Economy in a Low Interest Rate Environment* (*Monetary and Economic Studies*, Vol. 24, No. S-1, 2006年)
- ・ *Incentive Mechanisms for Economic Policymakers* (*Monetary and Economic Studies*, Vol. 23, No. S-1, 2005年)
- ・ *Challenges for Sustained Economic Growth under Changing Economic, Social, and International Environments* (*Monetary and Economic Studies*, Vol. 22, No. S-1, 2004年)
- ・ *Exchange Rate Regimes in the 21st Century* (*Monetary and Economic Studies*, Vol. 20, No. S-1, 2002年)
- ・ *The Role of Monetary Policy under Low Inflation: Deflationary Shocks and Policy Responses* (*Monetary and Economic Studies*, Vol. 19, No. S-1, 2001年)
- ・ *Monetary Policy in a World of Knowledge-Based Growth, Quality Change and Uncertain Measurement* (Palgrave, 2001年)
- ・ *Towards More Effective Monetary Policy* (Macmillan Press, 1997年)
- ・ *Financial Stability in a Changing Environment* (Macmillan Press, 1995年)
- ・ *Price Stabilization in the 1990s: Domestic and International Policy Requirements* (Macmillan Press, 1993年)
- ・ *The Evolution of the International Monetary System: How Can Efficiency and Stability Be Attained?* (University of Tokyo Press, 1990年)
- ・ *Toward a World of Economic Stability: Optimal Monetary Framework and Policy* (University of Tokyo Press, 1988年)
- ・ *Financial Innovation and Monetary Policy: Asia and the West* (University of Tokyo Press, 1986年)
- ・ *Monetary Policy in Our Times* (MIT Press, 1985年)

なお、2008年以降に開催された国際コンファランスについては、議事要旨等を *Monetary and Economic Studies* に所収。

日本銀行の機能と業務（有斐閣、2011 年発行）

日本銀行がどのような役割・機能を担い、また、これを果たすため、どのような業務を行っているのかについて、網羅的かつ具体的に解説。

貨幣博物館 常設展示図録（ときわ総合サービス、2017 年発行）

貨幣博物館の展示資料を紹介した図録です。

日本銀行金融研究所（IMES）

〒103-8660 東京都中央区日本橋本石町 2-1-1

TEL：03-3279-1111

E-mail：imes.journals-info@boj.or.jp

ホームページ：https://www.imes.boj.or.jp

資料公開サービス

以下の資料公開サービスを行っておりますので、ご利用下さい。

(1) 貨幣博物館の公開

2015年11月、貨幣博物館はリニューアルオープンしました。日本の貨幣を中心に、和同開珎や大判・小判の実物、貨幣に関する絵画など、貴重な資料を多数展示するとともに、所蔵資料を貨幣博物館ホームページ等で一般に公開しています。20名以上の団体で見学を希望される場合は予め電話でご連絡下さい。

開館時間：9時30分～16時30分（入館は16時まで）

休館日：月曜日（ただし祝休日は開館）

年末年始（12月29日～1月4日）

※最新の情報は下記の当館ホームページで

お知らせいたしますので、ご確認ください。

貨幣博物館ホームページ：<https://www.imes.boj.or.jp/cm/>

〈連絡先〉金融研究所 貨幣博物館

TEL：03-3277-3037（直通）

(2) 日本銀行が保有する歴史的公文（歴史的文書）の公開

金融研究所では、日本銀行に関連する歴史的資料のうち、歴史的価値を有するものについて保管・整理し、一般に公開しています。閲覧をご希望の方は、当館ホームページ掲載の目録で検索のうえ、予めメール・電話にてご連絡下さい。

歴史的公文の目録：<https://www.imes.boj.or.jp/archives/hozon.html>

〈連絡先〉金融研究所 アーカイブ

E-mail：imes.archives@boj.or.jp

TEL：03-3277-2151（直通）

金 融 研 究 (第 45 卷 第 3 号)

ISSN 2759-5099

2026 年 7 月 21 日 発行

日本銀行金融研究所長

編集兼発行者 渡 辺 真 吾

発 行 所 日本銀行 金融研究所

郵便番号 103-8660

東京都中央区日本橋本石町 2-1-1

電 話 : (03) 3279-1111 (大代表)

国際文献社

本誌に関する照会は、日本銀行金融研究所までお寄せ下さい。

