IMES DISCUSSION PAPER SERIES

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（ひらきかずひろ）・石原慎一（いしはらしんいち）・篠 潤之介（しのじゅんのすけ）

Discussion Paper No. 2025-J-10

IMES

INSTITUTE FOR MONETARY AND ECONOMIC STUDIES

BANK OF JAPAN

日本銀行金融研究所

〒103-8660 東京都中央区日本橋本石町 2-1-1

IMES Discussion Paper Series 2025-J-10
2025 年 12 月

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***

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[16]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、（1）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、（2）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、（3）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

*　国際通貨基金（khiraki@imf.org）
**　独立研究者（ishihara5683@gmail.com）
***　早稲田大学国際学術院（junnosuke.shino@waseda.jp）

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）、また、渡辺真吾氏、池田大輔氏、崎山登志之氏、高橋耕史氏ほか日本銀行スタッフから有益なコメントを頂いた。なお、本稿の内容と意見は筆者ら個人に属するものであり、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。

1　はじめに

1.1　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で, 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.

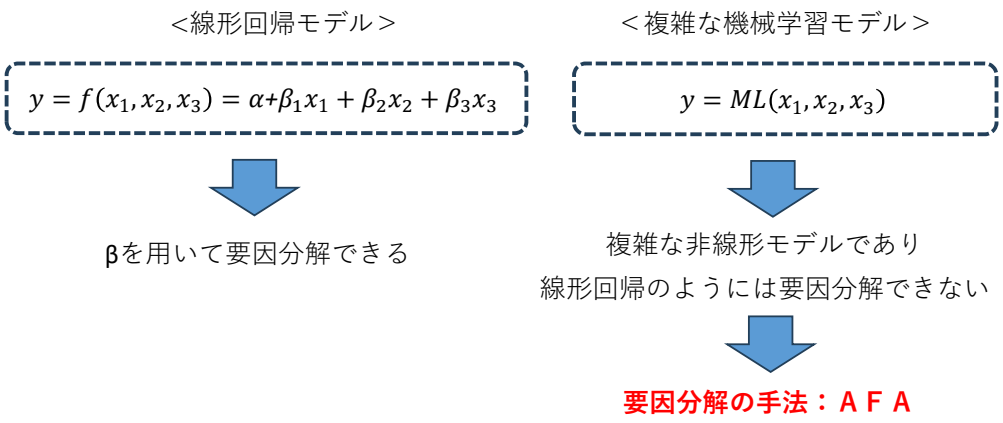
具体的には, XAI の枠組みに含まれる AFA（Additive Feature Attribution）とは, 複雑な機械学習モデルの予測値を各特徴量の影響度に分解することで, 個々の要因が予測に与える効果を定量化し可視化するアプローチである. たとえば, 3 つの変数 X・Y・Z を用いて資産価格を予測する場合, AFA は予測値のどの程度が X に由来するものなのか, どの程度が Y に由来するものなのか, どの程度が Z に由来するものなのかを分解・可視化する. 線形回帰モデルであれば, こうした要因分解は推計されたパラメータを用いて実行することができる. ところが, ニューラルネッ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 [33]) に基づく AFA であり, Lundberg and Lee [25] (以下「LL 論文」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]

[*1] Lundberg and Lee [25] (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.

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



SHAP の計算速度を短縮化するための TreeSHAP (Lundberg et al. [24]) や FastSHAP (Jethani et al. [21]) 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. [19] 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 [12] 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示した．Hadji et al. [27] は, 信用リスクの要因分解に対し, また, Bussmann et al. [7] 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. [1] 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. [34] 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

[*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. [23] 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.[26] などが挙げられる．

[*4] XAI や SHAP を平易に解説した文献として, Molnar [28] のほか, 日本語では大坪ほか [36] や森下 [40] がある．和泉 [35] 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．

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 [6] 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 (Shaprey 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. [5] 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. [4] 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 [41] 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 [42] 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 [39] 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.

1.2 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 [16] (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, [30]）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 [38] 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 [6] 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.

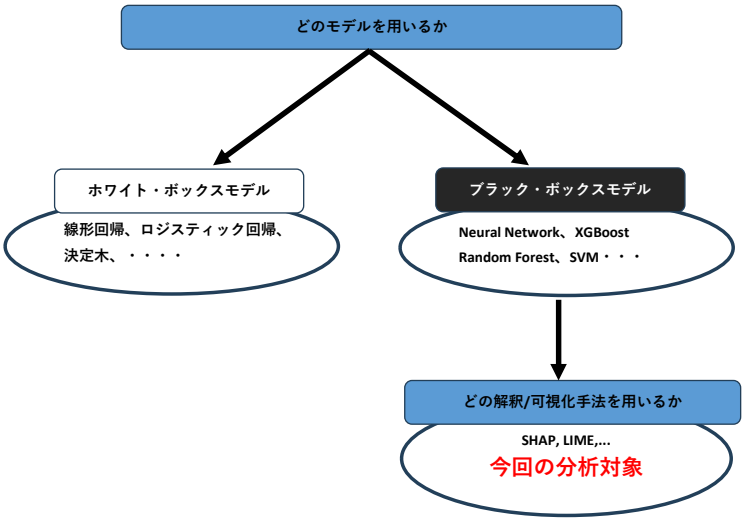
実証分析の結果明らかになった点を簡潔に先取りすると, 以下の通りである. 1 点目に, 協力ゲーム理論の解概念である残余均等配分解 (ES) に基づく AFA や, それに類似した解概念 (ENSC) に基づく AFA については, SHAP との間で分解パターンの違いが明確にみられた. 2 点目に, それ以外の AFA の間では, 視覚的に明確に確認できるほどの大きな差異はみられないものの, AFA 間の差異の大きさは, 各 AFA に対応するカーネル関数の形状を反映したものとなっていることが確認された. 3 点目に, ES に基づく AFA と, ENSC に基づく AFA を按分して定義される 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. [15] 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. [9] 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. [3], オプション・リターンについては Bali et al. [2] 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. [11] 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.

[*5] さらに, 補論 3 では, Jabeur et al. [19] 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



2　理論分析：SHAP とその代替的手法についての導出と考察

2.1　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 [25] 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 2.4 節で解説), それが SHAP と一致することを示している (同論文の Theorem2).

一方, HIS 論文 [16] 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, そこから SHAP と代替的な AFA を提示した (図 4 における Approach 2).

図 3: LL 論文における SHAP の特徴づけ

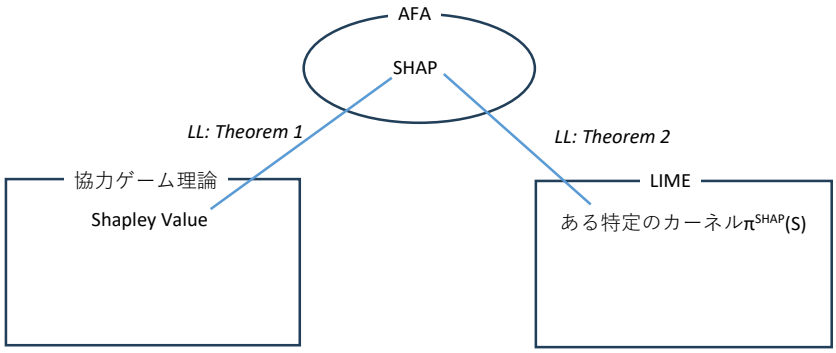
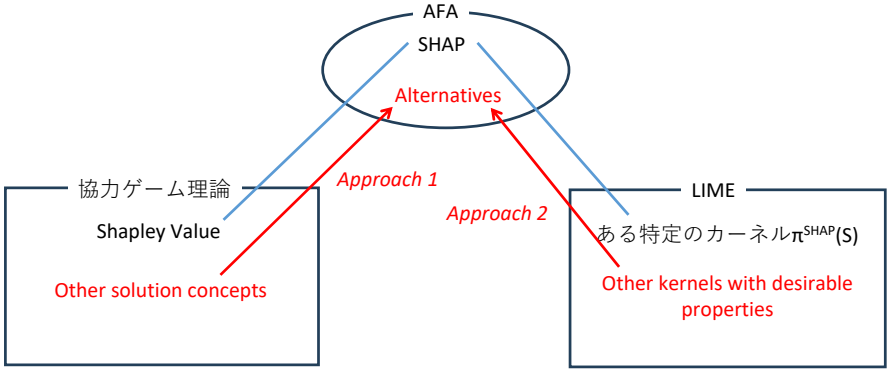


図 4: HIS 論文における代替的手法の提示



以上が LL 論文と HIS 論文の全体感である. 2.2 節以降でこれらの 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.

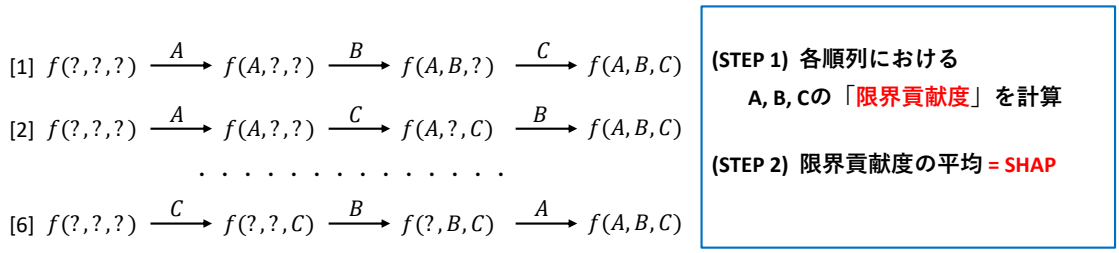
SHAP と ES の違いを直感的に理解するために, ここでは簡単な例を用いる. 学習済の機械学習モデルを f, 特徴量は A, B, C の 3 つであるとする. 例えば, 株価リターンを予測するとして, 機械学習モデル f はランダムフォレストやニューラルネットといった複雑な「ブラックボックス」モデル, 特徴量は計量分析における「独立変数」「説明変数」のことであり, 企業収益, 配当性向, 為替レート, 株価自身の過去のリターンやボラティリティといった変数が特徴量の候補となる.

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 $f(\emptyset)$ とする – の差, すなわち $f(A, B, C) - f(\emptyset)$,

を, 特徴量 A, B, C に配分する手法のことである. $f(A,B,C) - f(\emptyset)$ 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 $f(A,B,C) - f(\emptyset)$ 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 $f(\emptyset)$ (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 $f(A) - f(\emptyset)$ 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.

図 5: SHAP（シャープレイ値を用いた AFA）の計算方法



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 $A \to B \to C$ 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値を計算し, 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 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.

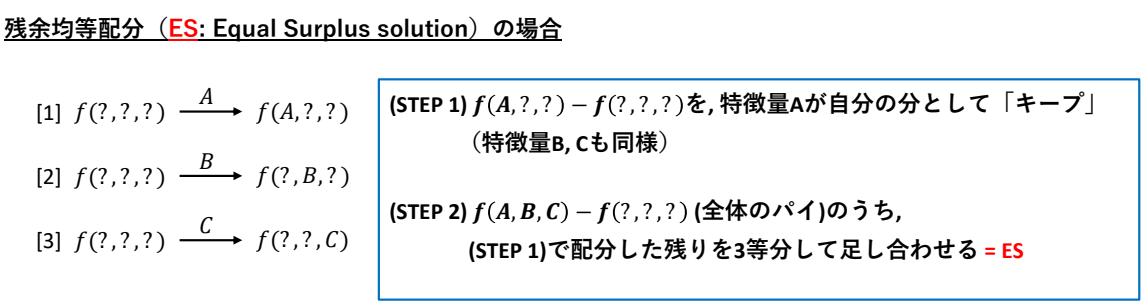
次に, 残余均等配分解 (ES) についてみてみよう (図 6). ES でも, まず, 全ての特徴量が未知の状態である $f(\emptyset)$ からスタートする. しかし, 3 つの特徴量全てが加わる順列を考えるのではなく, 考慮するのは $f(\emptyset)$ から 1 つめの特徴量が加わる状態のみである. 図 6 で [1] とあるケースにおいては, すべての特徴量が未知である場合の予測値 $f(\emptyset)$ から, 特徴量 A が既知になったときの予測値 $f(A)$ の変化幅 $f(A) - f(\emptyset)$ を, SHAP 同様に A の予測の「限界貢献度 (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

[*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.

[*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.

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．

図 6: **ES** (残余均等配分解を用いた **AFA**) の計算方法



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 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 [28], 森下 [40] 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, 2.2 節以降では, 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．

2.2 準備

観測値を t 個, 特徴量の数を n 個とし ($N = \{1, ..., n\}$ および $T = \{1, ..., t\}$), 特徴量のベクトルを $t \times n$ 次元ベクトル $X = (X_1, ...X_j, ..., X_n)$ とする. j 番目の特徴量のベクトルを $X_j = (x_{1,j}, ..., x_{t,j})'$, また, τ 番目の観測値の特徴量のベクトルを $x_\tau = (x_{\tau,1}, ..., x_{\tau,j}, ..., x_{\tau,n})$ とする. f を学習済モデルとし, $Y = (y_1, ..., y_t)'$ 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. $N = \{1, ..., n\}$ 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 $v_\tau : 2^N \longrightarrow R$ 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:

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

(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, $E\left[f(x_{\tau,S}, X_{N \setminus S})\right] = (1/t) \sum_{\tau' \in T} f(x_{\tau,S}, x_{\tau',N \setminus S})$ 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.

$v_\tau(N) = E\left[f(x_{\tau,1}, ..., x_{\tau,n})\right] = f(x_{\tau,1}, ..., x_{\tau,n})$ かつ $v_\tau(\emptyset) = E[f(X_1, ..., X_n)] = E[f(X)]$ 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.

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

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:

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

$v_4(\emptyset)$ 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, $v_\tau(N) - v_\tau(\emptyset)$

[*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 [6], Lundberg and Lee [25]). 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. [20] がある. Chen et al. [8] 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 [28]).

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 $\Psi_\tau(j) : N \longrightarrow R$ 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, $\Psi_\tau = (\Psi_{\tau,1}, ..., \Psi_{\tau,n})$ とする. Ψ_τ が以下の (3) 式を満たすとき, Ψ_τ を AFA と呼ぶ:

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

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.

2.3 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, 2.1 節で取り上げた 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:

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

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

$\Psi_{\tau,j}^{SHAP}$ は, 2.2 節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 [33]) 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. [10] 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.

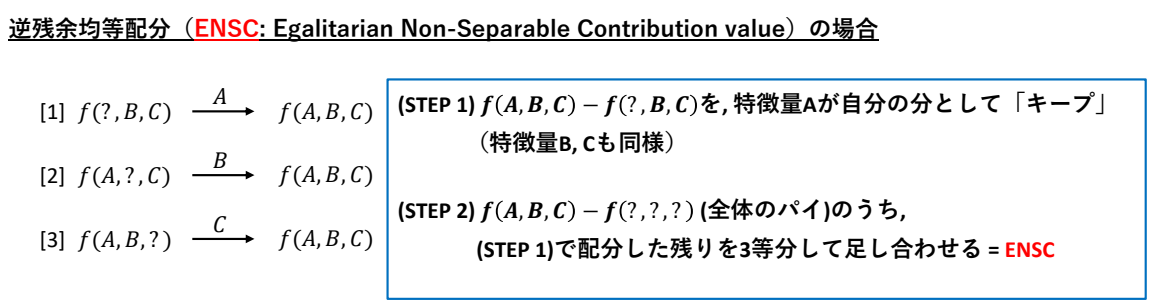
このうち 1 つめの AFA は, 残余均等配分解に対して,「逆残余均等配分解」とも呼ぶべきものである. 協力ゲーム理論においては, Egalitarian Nonseparable Contribution (ENSC) という名で, 1990 年代に定式化・公理化が行われている (Dragan et al. [13] および Drissen and Funaki [14] 参照).

2.1 節における, 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,

[*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.

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）.

図 **7: ENSC** (逆残余均等配分解を用いた **AFA**) の計算方法



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.

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

(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.

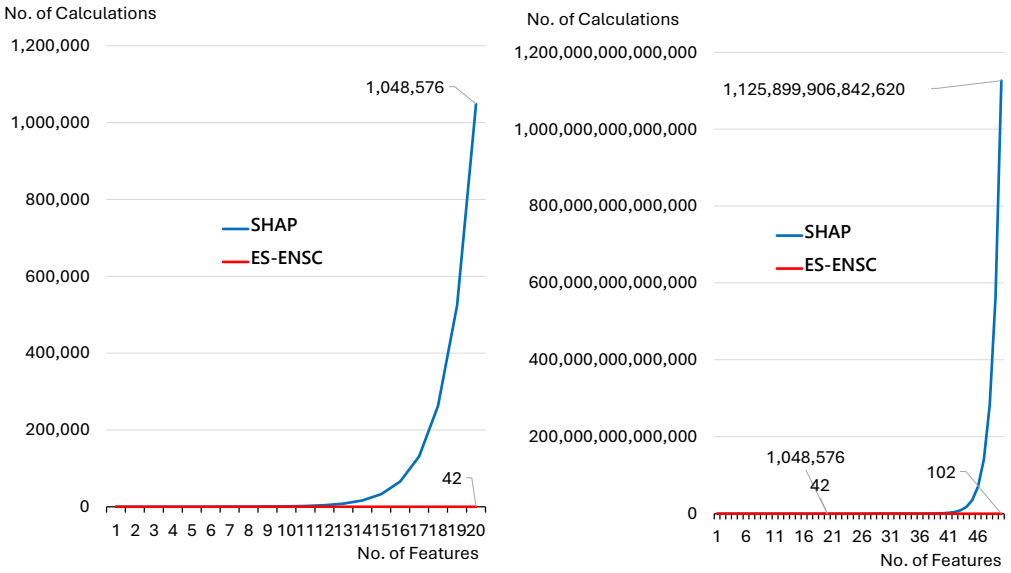
本項で新たに提示する 2 つめの AFA は, ES と ENSC を按分した $\Psi_{\tau,j}^{ES-ENSC}$ であり, 以下の (7) 式によって定義される.

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

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．

2.4 LIME とカーネル

HIS 論文では, LIME（Ribeiro et al. [30], 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．2.4 節と 2.5 節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

[*10] 例えば, 分析対象の観測値を x = (20代, 大卒, 東京居住) とすると, x' = $(1,1,1)$ であり ((20代, 大卒, 東京居住) = $h_x((1,1,1))$), z = (40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) = $h_x((1,1,0))$).

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:

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

ここで, $g(z) = \phi_0 + \sum_{i=1}^{n} \phi_i z_i$ (ただし $\phi_i \in R$), 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, $\phi = (\phi_1, ..., \phi_n) \in R^n$ 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るべきであるため, 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. 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:

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

ここで, 以下の分析のために, (8) 式の最小化問題を, 2.2 節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

[*11] LIME 論文 [30] 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.

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:

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

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, $z=(0,...,0)$ のとき, $\phi_0=f(0,...,0)$, 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.

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

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.

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

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 $\pi_{x_\tau}(S)$ 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 $\pi_{x_\tau}(S)$ について, 本項では以下の対称性条件を課す:

$$\pi_{x_\tau}(S)=\pi_{x_\tau}(T)\qquad\Big(\forall S,T\in 2^N\ with\ |S|=|T|\Big) \tag{12}$$

(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, ということを意味する. 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:

$$\sum_{S\in 2^N: j\in S} 2\left(\sum_{i\in S}\phi_i - \{v_\tau(S) - v_\tau(\emptyset)\}\right)\pi_{x_\tau}(S) = 0. \tag{13}$$

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:

$$\sum_{S\in 2^N: i\in S}\left(\sum_{k\in S}\phi_k - \{v_\tau(S) - v_\tau(\emptyset)\}\right)\cdot\pi_{x_\tau}(S) = \sum_{S\in 2^N: j\in S}\left(\sum_{k\in S}\phi_k - \{v_\tau(S) - v_\tau(\emptyset)\}\right)\cdot\pi_{x_\tau}(S) \tag{14}$$

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

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

(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 $\phi_i\Sigma_{S\subseteq N\setminus\{i,j\}}\pi_{x_\tau}(S\cup\{i\})$, 第 2 項 $\phi_j\Sigma_{S\subseteq N\setminus\{i,j\}}\pi_{x_\tau}(S\cup\{j\})$ において, $\Sigma_{S\subseteq N\setminus\{i,j\}}\pi_{x_\tau}(S\cup\{i\}) = \Sigma_{S\subseteq N\setminus\{i,j\}}\pi_{x_\tau}(S\cup\{j\}) = 1$ 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.

2.5 節以降では, 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:

$$\phi_1 - \sum_{S:1\in S, S\neq N}\pi_{x_\tau}(S)\cdot v_\tau(S) = \cdots = \phi_n - \sum_{S:n\in S, S\neq N}\pi_{x_\tau}(S)\cdot v_\tau(S). \tag{17}$$

さらに, (13) 式より, 以下が成り立つ:

$$\left(\sum_{S\in 2^N: j\in S}\pi_{x_\tau}(S)\right)\phi_j + \sum_{i\in N: i\neq j}\left(\sum_{S\in 2^N: i,j\in S}\pi_{x_\tau}(S)\right)\phi_i = \sum_{S\in 2^N: j\in S}\pi_{x_\tau}(S)\cdot\Big(v_\tau(S) - v_\tau(\emptyset)\Big). \tag{18}$$

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:

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

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

$$\phi_j = \sum_{S : j \in S \neq N} \pi_{x_\tau}(S) \cdot v_\tau(S) + \frac{T - \sum_{i \in N} \left\{ \sum_{S : i \in S \neq N} \pi_{x_\tau}(S) \cdot v_\tau(S) \right\}}{n} \quad (20)$$

ここで, T は以下の通りである:

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

(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:

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

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 $\phi = (\phi_1, ..., \phi_j, ..., \phi_n)$ は, 以下で表される:

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

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

[*12] なお, Ruiz et al. [32] 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.

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 2.5 節では, 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.

2.5 カーネル関数に基づく AFA の理論的比較

ここでは, 2.4 節で得られた AFA とカーネル関数の関係を基に, まずは, 2.3 節で取り上げた, 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, 2.5.1 節). 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が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, 2.5.2 節).

2.5.1 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]

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

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

[*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.

[*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:

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

(24) の右辺は (23) の右辺を定数倍 ($n/(n-1)$ 倍) したものなので, 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.

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.

(24) 式において, $|S| = 0$ または $|S| = n$ のとき $\pi_{x_\tau}^{SHAP}(S) = \infty$ 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_τ} 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.

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

$\pi_{x_\tau}^{ES}$ も, $|S|$ に関する増加関数という条件を明らかに満たしてない. すなわち, 2.3 節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.

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

$\pi_{x_\tau}^{ENSC}$ 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 [13] 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.

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

[*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.

$\pi_{x_\tau}^{EN-ENSC}$ 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 [22] 参照).

2.5.2 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が持つべき条件である「既知の特徴量の数に関しての増加関数」– すなわち, $|S|$ についての増加関数という条件 – 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]

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

(28) を (22) に代入して, 以下の AFA を得る:

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

$\Psi_{\tau,j}^{LnK}$ 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　2.4 節で触れたように, LIME 論文 [30] 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:

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

[*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. 2.4 節の (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 $\pi^{ExK}_{x_\tau}$ を得る:

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

$\pi^{ExK}_{x_\tau}$ 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 $\Psi^{ExK}_{\tau,j}$ は (32) 式の通りとなる.

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

$\Psi^{ExK}_{\tau,j}$ 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]

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

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

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

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.

[*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.

■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]

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

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:

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

$\Psi_{\tau,j}^{PNucl}$ 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. [31][32]) と一致することが証明できる.[*19] カーネル関数は特徴量の数から独立 (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.

2.6　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 番目は既知の特徴量の数 $|S|$ (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 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. 具体的には, 3.1 節でわが国の 10 年国

[*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 [25]	シャープレイ値 [33]	U 字型
$\Psi_{\tau,j}^{ES}$	(5)	(25)	Condevaux 他 [10]	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) [13][14]	増加関数
$\Psi_{\tau,j}^{ES-ENSC}$	(7)	(27)	本分析	ES と ENSC の按分 [22] 等	U 字型
$\Psi_{\tau,j}^{LnK}$	(29)	(28)	HIS 論文 [16]	–	線形増加
$\Psi_{\tau,j}^{ExK}$	(32)	(31)	HIS 論文 [16]	–	指数関数的増加
$\Psi_{\tau,j}^{CvK}$	(34)	(33)	HIS 論文 [16]	–	対数関数的増加
$\Psi_{\tau,j}^{PNucl}$	(36)	(35)	HIS 論文 [16]	LS プレ仁 [32]	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.

債利回り, 3.2 節でわが国の失業率を対象にする.[*20]

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.[*21] また, 補論 3 では, 平木ほか [37] 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.

3.1 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 [38] 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.[*22]

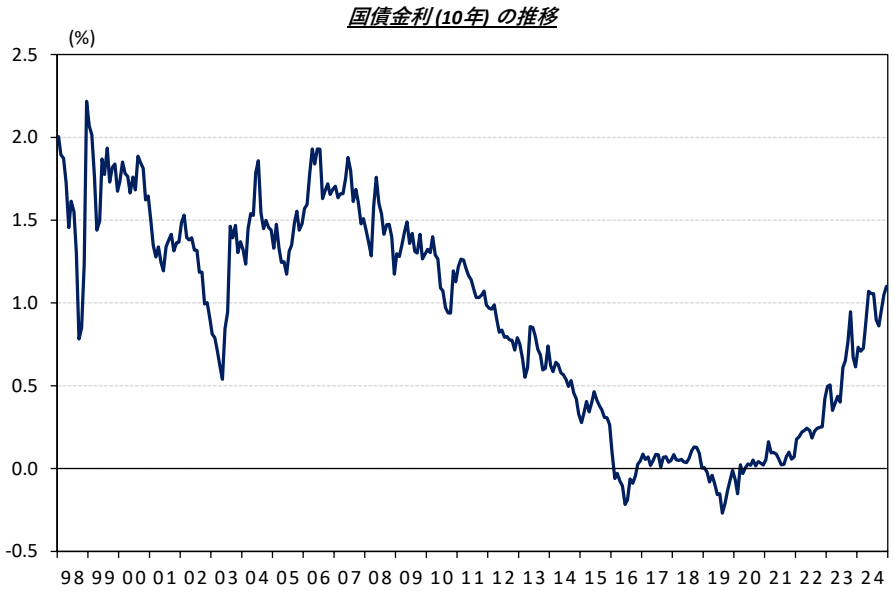
[*20] 既述のとおり, 本稿で取り上げた AFA は, 時系列データだけではなく, 様々な数値データ, 画像や音声データといった全てのタイプのデータに適用可能である. 本稿では, 著者らのドメイン知識も活用して AFA 評価を行う観点から, わが国の経済・ファイナンスに関する時系列データを対象とした.

[*21] 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.

[*22] 日本銀行 [38] 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

- 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)

図 **9:** 長期金利 **(10** 年国債利回り**)** の推移



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

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.

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.

図 10: 長期金利の AFA 分解 (1)

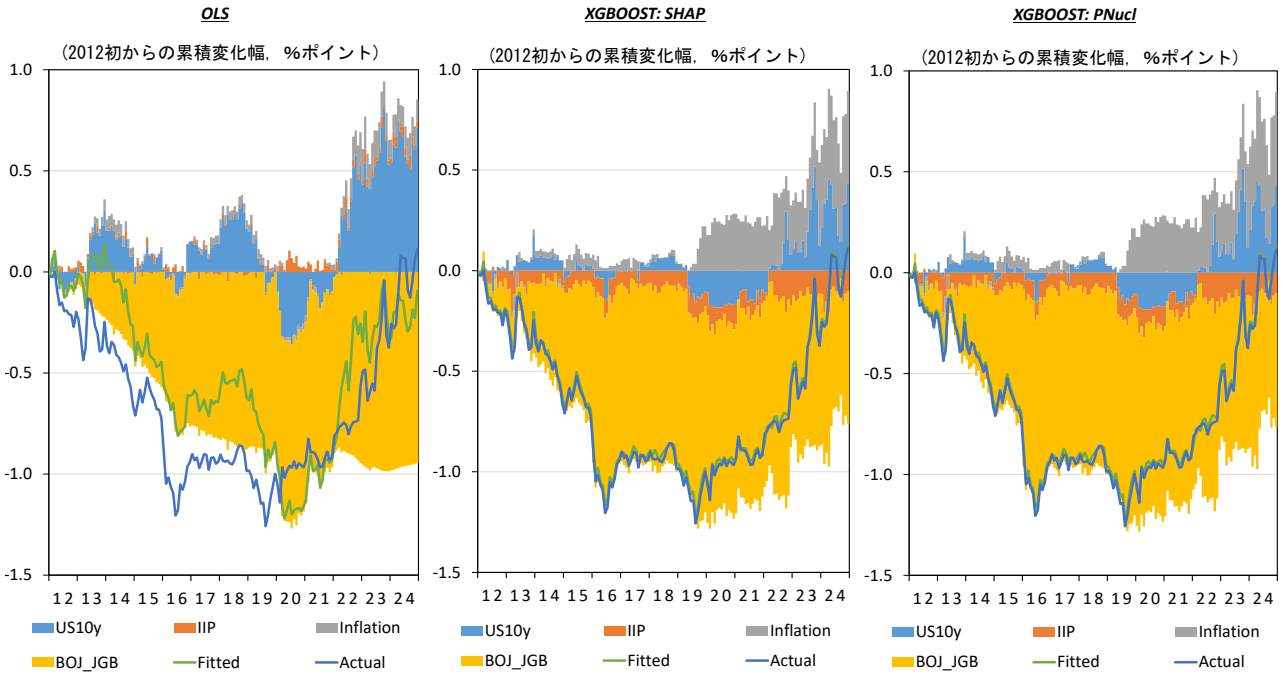


図 11: 長期金利の AFA 分解 (2)

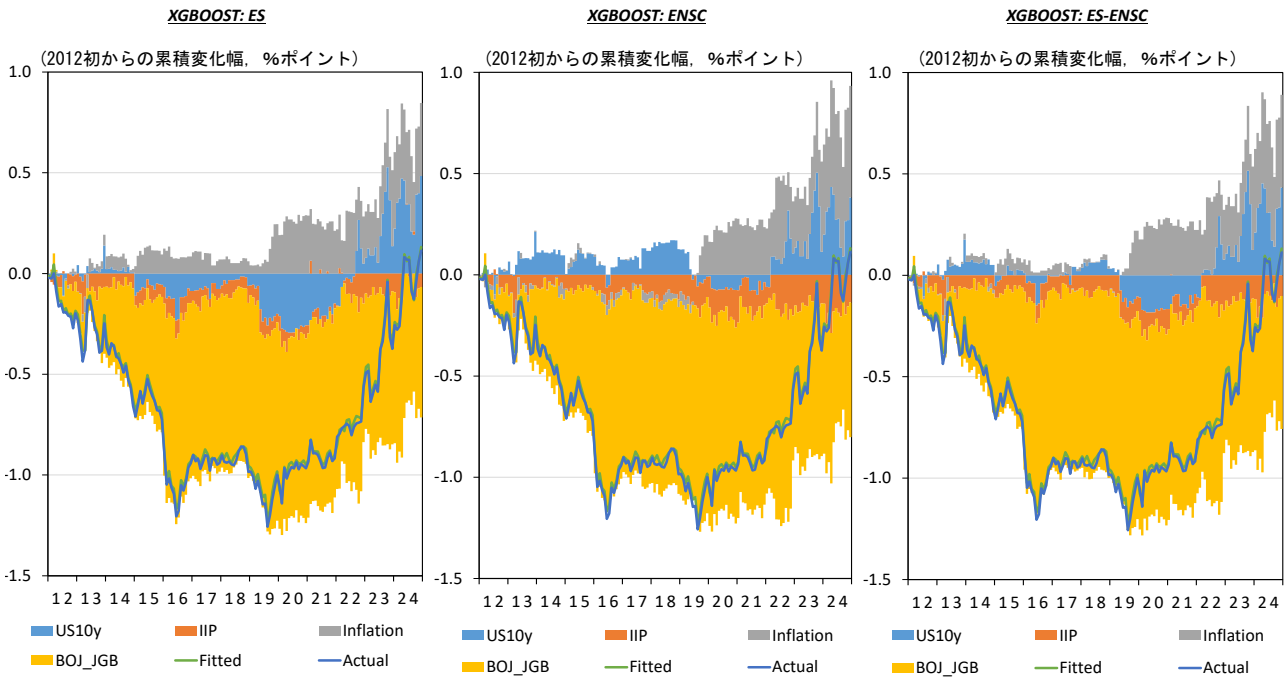
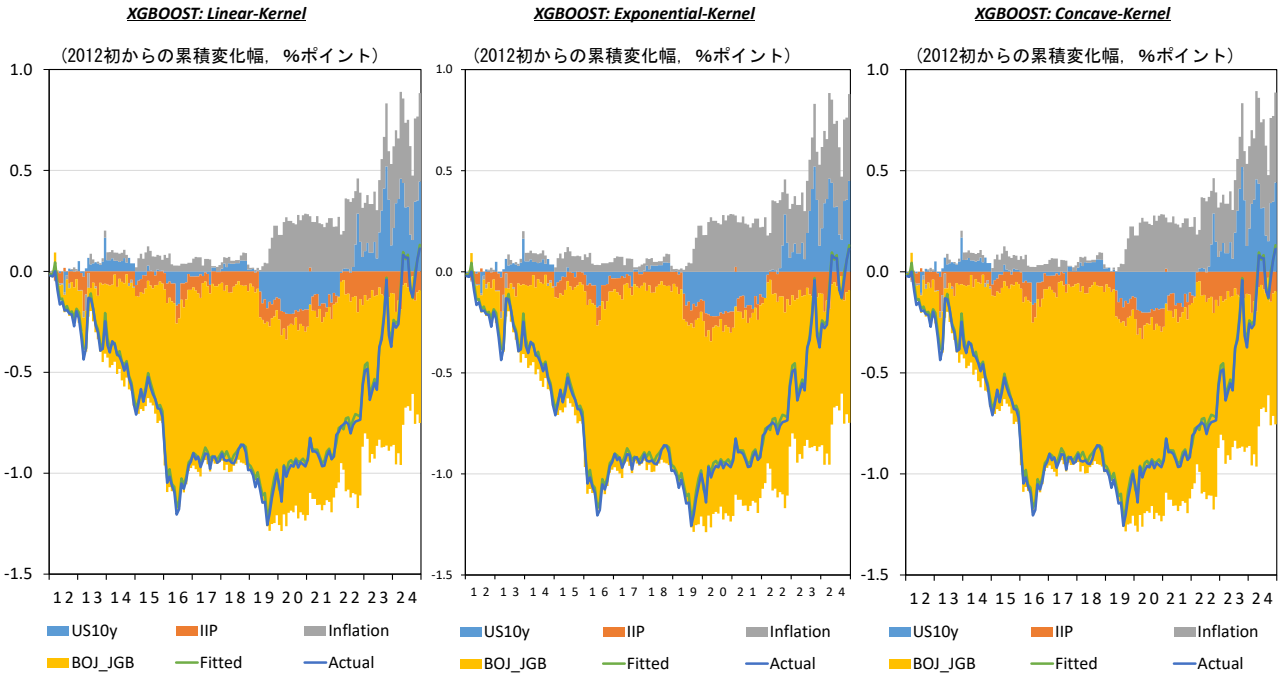


図 12: 長期金利の AFA 分解 (3)



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（$\Psi_{\tau,j}^{CvK}$）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っきりと小さい（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, 既知の特徴量の数が中程度（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. 2.3 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, その差は特徴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 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, 視覚的にはほどんど確認できない. 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徴量について, i 行と j 列に対応した AFA を差の絶対値を計算し, それを全ての特徴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, 右上の領域は常にゼロとなる.

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.

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.

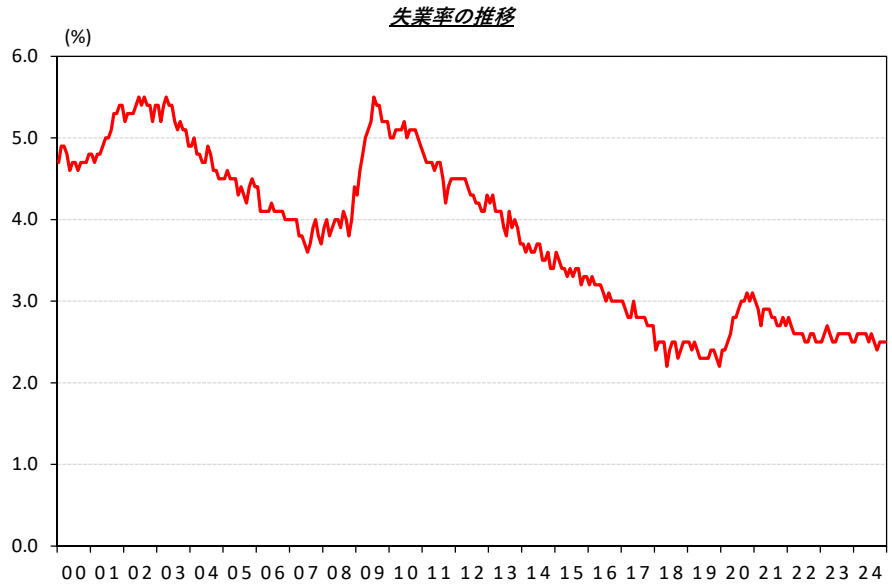
3 点目は, AFA 間の差異の大きさは, 各 AFA に対応するカーネル関数の形状を反映したものとなっている. 例えば, 特徴量の数とは独立の, 一定のカーネル関数を持つ LS プレ仁 (PNucl) をベースにした AFA を基準に, それ以外の AFA との差をみると, 一定の範囲内で PNucl のカーネル関数と類似した形状となる, U 字型のカーネル関数を持つ SHAP との差がもっとも小さくなっている. さらに, 増加関数型のカーネル関数に基づく AFA と PNcul との違いをみると, 傾きが急激に高まる指数関数型のカーネル関数に基づく AFA (Exponential) がも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 PNcul 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 間の間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.

3.2　わが失業率の AFA 分解

Buckmann and Joseph [6] 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.

図 13: 失業率の推移



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 [6] を参考に, 特徴量は以下の 9 つとした. 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 $\Psi_{\tau,j}^{CvK}$ 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.

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 にかけて, 個人消費 (の弱さ) が失業率の押し上げ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, 視覚的にはほどんど確認できず, 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).

図 14: 失業率の AFA 分解 (1)

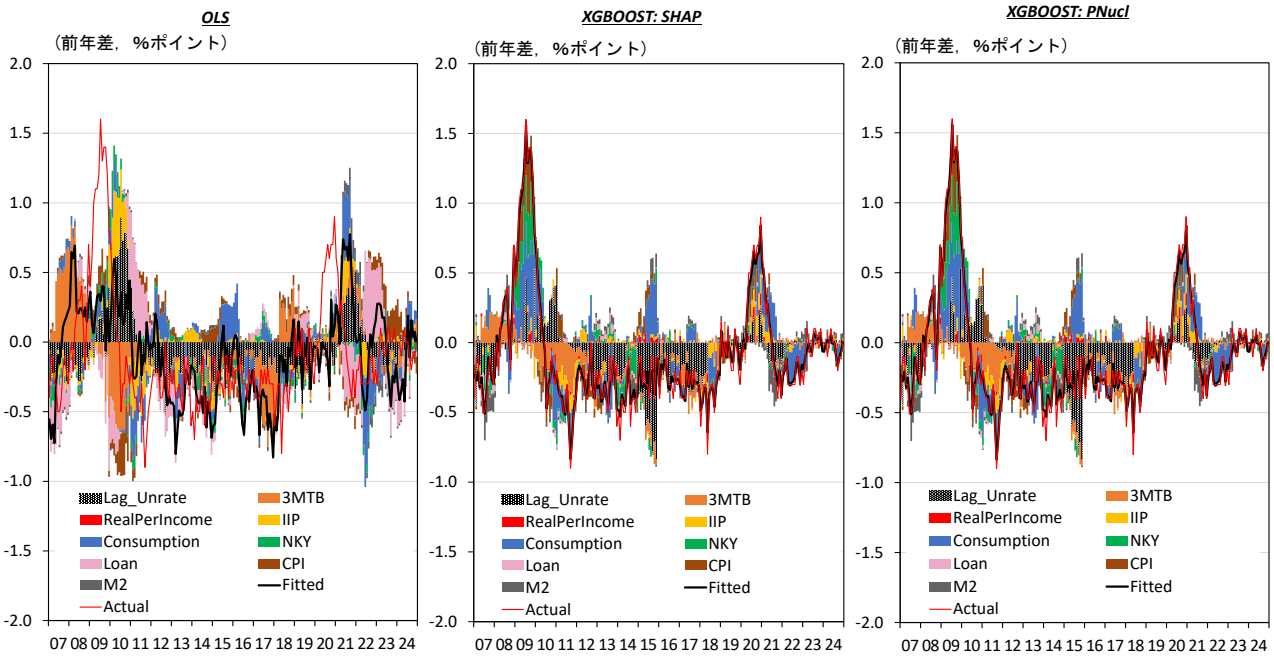


図 15: 失業率の AFA 分解 (2)

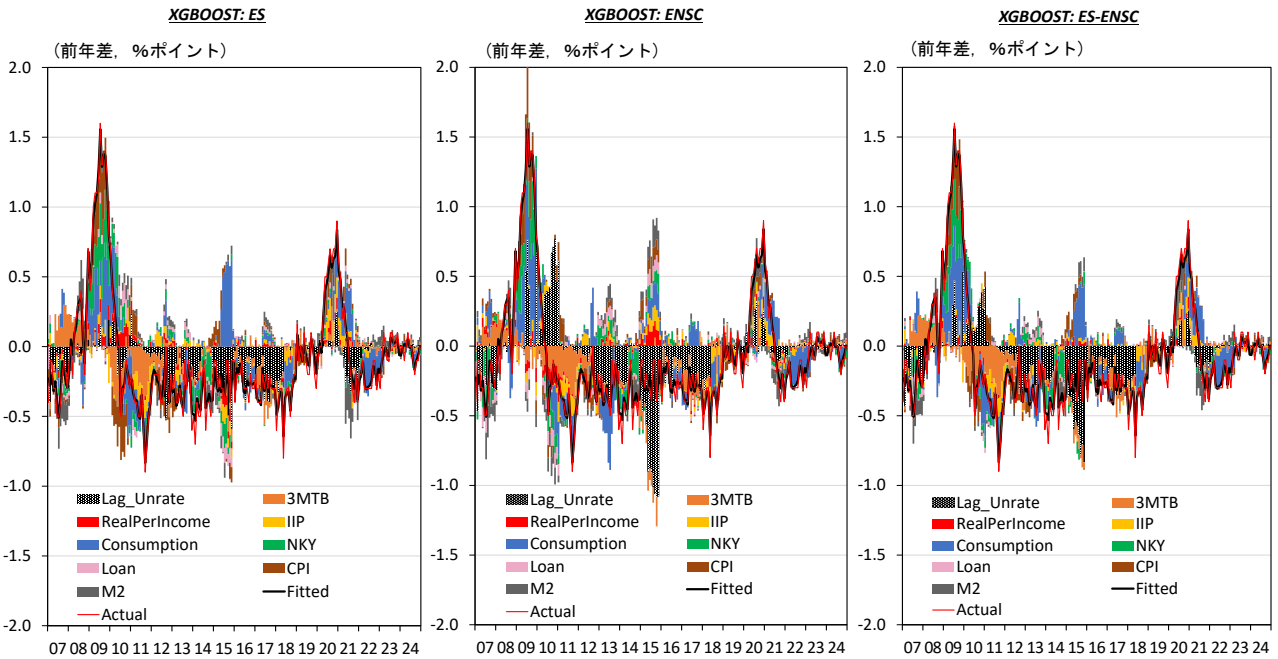
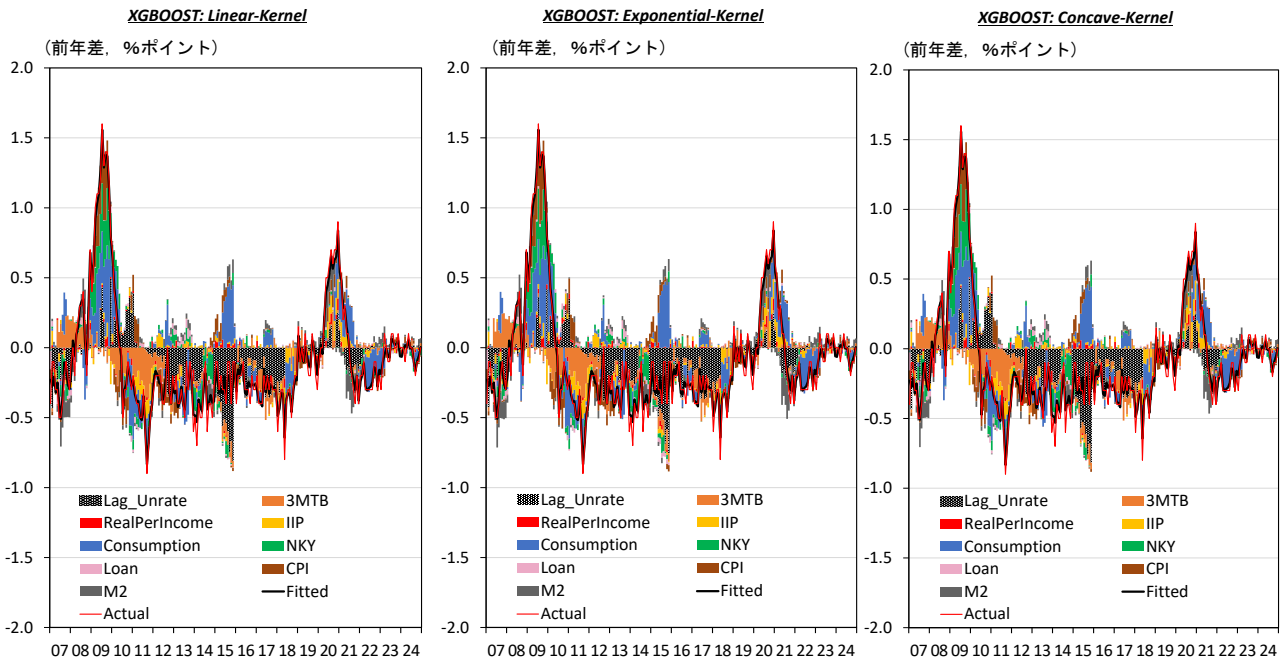


図 16: 失業率 AFA 分解 (3)



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 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.

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.

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 [16] (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. [30] 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 (2.5.2 節の議論も参照). 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.[*23]

■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. [15] 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

[*23] 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.

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. [29] 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 [17][18] 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.

参考文献

[1] M. J. Ariza-Garzón, J. Arroyo, A. Caparrini, and M.-J. Segovia-Vargas. Explainability of a machine learning granting scoring model in peer-to-peer lending. *IEEE Access*, 2020.

[2] T. G. Bali, H. Beckmeyer, M. Mörke, and F. Weigert. Option return predictability with machine learning and big data. *Review of Financial Studies*, 36(9):3548–3602, 2023.

[3] D. Bianchi, M. Büchner, and A. Tamoni. Bond risk premiums with machine learning. *Review of Financial Studies*, 34(2):1046–1089, 2021.

[4] K. Bluwstein, M. Buckmann, A. Joseph, S. Kapadia, and O. Şimşek. Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics*, 145:103773, 2023.

[5] P. Bracke, A. Datta, C. Jung, and S. Sen. Machine learning explainability in finance: an application to default risk analysis. Staff Working Paper 816, Bank of England, 2021. Bank of England Staff Working Paper No.816.

[6] M. Buckmann and A. Joseph. An interpretable machine learning workflow with an application to economic forecasting. *International Journal of Central Banking*, 19–4:449–522, October 2023.

[7] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock. Explainable machine learning in credit risk management. *Computational Economics*, 57:203–216, 2021.

[8] H. Chen, J. D. Janizek, S. M. Lundberg, and S.-I. Lee. True to the model or true to the data? *arXiv*, pages 973–989, 2020.

[9] L. Chen, M. Pelger, and J. Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.

[10] C. Condevaux, S. Harispe, and S. Mussard. Fair and efficient alternatives to shapley-based attribution methods. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2023.

[11] P. G. Coulombe, M. Leroux, D. Stevanovic, and S. Surprenant. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37:920–964, 2022.

[12] U. Demirbaga and Y. Xu. Empirical asset pricing using explainable artificial intelligence. *SSRN Working Paper*, 2024.

[13] I. Dragan, T. Driessen, and Y. Funaki. Collinearity between the shapley value and the egalitarian division rules for cooperative games. *OR SPEKTRUM*, 18:97–105, 1996.

[14] T. Driessen and Y. Funaki. Coincidence of and collinearity between game theoretic solutions. *OR SPEKTRUM*, 13:15–30, 1991.

[15] S. Gu, B. Kelly, and D. Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33:2223–2273, 2020.

[16] K. Hiraki, S. Ishihara, and J. 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.

[17] S. Ishihara and J. Shino. Some properties of interval shapley values: an axiomatic analysis. *Games*, 14 (3):50, 2023.

[18] S. Ishihara and J. Shino. An axiomatization of the shapley mapping using strong monotonicity in interval games. *Annals of Operational Research*, 345:147–168, 2025.

[19] S. B. Jabeur, S. Mefteh-Wali, and J.-L. Viviani. Forecasting gold price with the xgboost algorithm and shap interaction values. *Annals of Operational Research*, 334:679–699, 2024.

[20] D. Janzing, L. Minorics, and P. Bloebaum. Feature relevance quantification in explainable ai: A causal problem. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

[21] N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath. Fastshap: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2021.

[22] T. Kongo. 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*, 338:973–989, 2024.

[23] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and R. Acharya. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011 – 2022). *Computer Methods and Programs in Biomedicine*, 226:107161, 2022.

[24] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence volume*, pages 56–67, 2020.

[25] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2016.

[26] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, and S.-I. Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2:749–760, 2018.

[27] B. H. Misheva, J. Osterrieder, A. Hirsa, O. Kulkarni, and S. F. Lin. Explainable ai in credit risk management. *arXiv*, 338:973–989, Mar. 2021.

[28] C. Molnar. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Lulu, 3 edition, 2025.

[29] D. Napolitano, L. Vaiani, and L. Cagliero. Efficient neural network-based estimation of interval shapley values. *IEEE Transactions on Knowledge and Data Engineering*, 36:8108–8119, 2024.

[30] M. T. Ribeiro, S. Singh, and C. 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*, pages 1135–1144, New York, NY, USA, 2016. ACM.

[31] L. M. Ruiz, F. Valenciano, and J. 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*, 25:113–134, 1996.

[32] L. M. Ruiz, F. Valenciano, and J. M. Zarzuelo. The family of least square values for transferable utility games. *Games and Economic Behavior*, 24:109–130, 1998.

[33] L. S. Shapley. A value for n-person games. *Annals of Mathematics Studies*, 28:307–318, 1953.

[34] J. Y. Wei, W. V. D. Heever, RuiMao, E. Cambria, R. Satapathy, and G. Mengaldo. A comprehensive review on financial explainable ai. *Artificial Intelligence Review*, 58(189), 2025.

[35] 和泉潔. 金融分野における因果推論の展開 一統計的手法・因果ＡＩ・自然言語処理の三潮流とその展望一. *SBI Research Review*, 8, 2025.

[36] 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.

[37] 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**, 2025.

[38] 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. 2021.

[39] 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, 2021.

[40] 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**. リックテレコム, 2021.

[41] 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 一原油価格の変動要因分析を例に 一. **日本銀行金融研究所**, 2022.

[42] 鷲見和昭. 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, 2020.

補論 1: 所与の AFA に対応するカーネル関数の導出方法

2.4 節および 2.5 節では, 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[*24]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 $\theta(v) = (\theta_1(v),\ \cdots,\ \theta_n(v))$ が[*25], 重み $w_i(S)$ を用いて以下のように表せるとする.

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

ただし, $w_i(S)$ は,

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

であり, $M(|S|) \geq 0$, $m(|S|) \leq 0$, $\sum_{|S|=1}^{n-1} \left(M(|S|) - m(|S|)\right) \cdot \left({}_{n-2}C_{|S|-1}\right) = 1$ 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.[*26]

$$\pi(S) = M(|S|) - m(|S|). \tag{40}$$

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

[*24] 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.

[*25] $\sum_{i \in N} \theta_i(v) = v(N)$ を満たすとき, 解 $\theta(v)$ は全体合理性を満たすと言う.

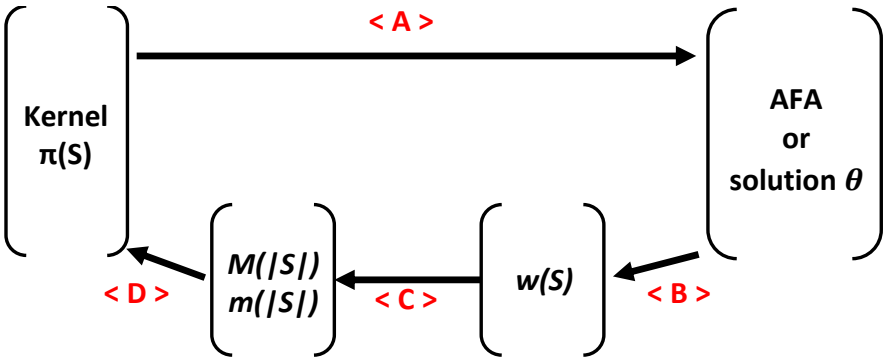
[*26] 証明の概略は以下の通り:

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

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.

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<**B**>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).

図 **17:** 本分析で示したカーネル関数と **AFA** の関係



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,

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

と表すことができる. したがって,

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

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,

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

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

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

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,

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

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:

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

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

このとき,

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

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 $\pi(S)$ は以下のようになる.

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

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*:*

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

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, $\Psi_{\tau,j}^{AFA} = v_{\tau}(N) - v_{\tau}(N \backslash \{j\})$ 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.

$$S \cap T = \emptyset \text{を満たす任意の} S, T \in 2^N \text{について}, v(S \cup T) - v(\emptyset) = \{v(S) - v(\emptyset)\} + \{v(T) - v(\emptyset)\}. \tag{42}$$

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:

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

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, $v_{\tau}(S) = E\left[f(x_{\tau,S}, X_{N \backslash S})\right] = \sum_{k:k \in S} f_k(x_{\tau,k}) + \sum_{l:l \notin S} E\left[f_l(X_l)\right]$. が成り立つ. $a_j = f_j(x_{\tau,j}) - E\left[f_j(X_j)\right]$ とすると, 任意の $S \in 2^N$ について, 以下が成り立つ:

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

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, $a_j = f_j(x_{\tau,j}) - E\left[f_j(X_j)\right]$ とおくと, (22) 式より以下が成り立つ:

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

(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;

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

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

■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:

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

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, $v_\tau(S) = E\left[f(x_{\tau,S}, X_{N\setminus S})\right] = \beta_0 + \sum_{k:k\in S} \beta_k x_{\tau,k} + E\left[\sum_{l:l\notin S} \beta_l X_l\right]$. したがって, $v_\tau(N) - v_\tau(\emptyset) = \sum_{j=1}^{n} \beta_j \left(x_{\tau,j} - E\left[X_j\right]\right)$. ここで $\Psi_\tau^{LM} = (\Psi_{\tau,1}, ..., \Psi_{\tau,n})$ を;

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

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 [40] 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, $a_j = f_j(x_{\tau,j}) - E\left[f_j(X_j)\right] = \beta_j \left(x_{\tau,j} - E\left[X_j\right]\right)$. (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. [19] は, 金価格を 6 つの機械学習モデル (Linear regression, Neural networks, Random forest, Light gradient boosing 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 [37] は, Jabeur et al. [19] 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.

図 18: 金価格の推移

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. [19] 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)
- 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 を適用した際のはっきりとした傾向として捉えることができる.

図 19: 金価格の AFA 分解 (1)

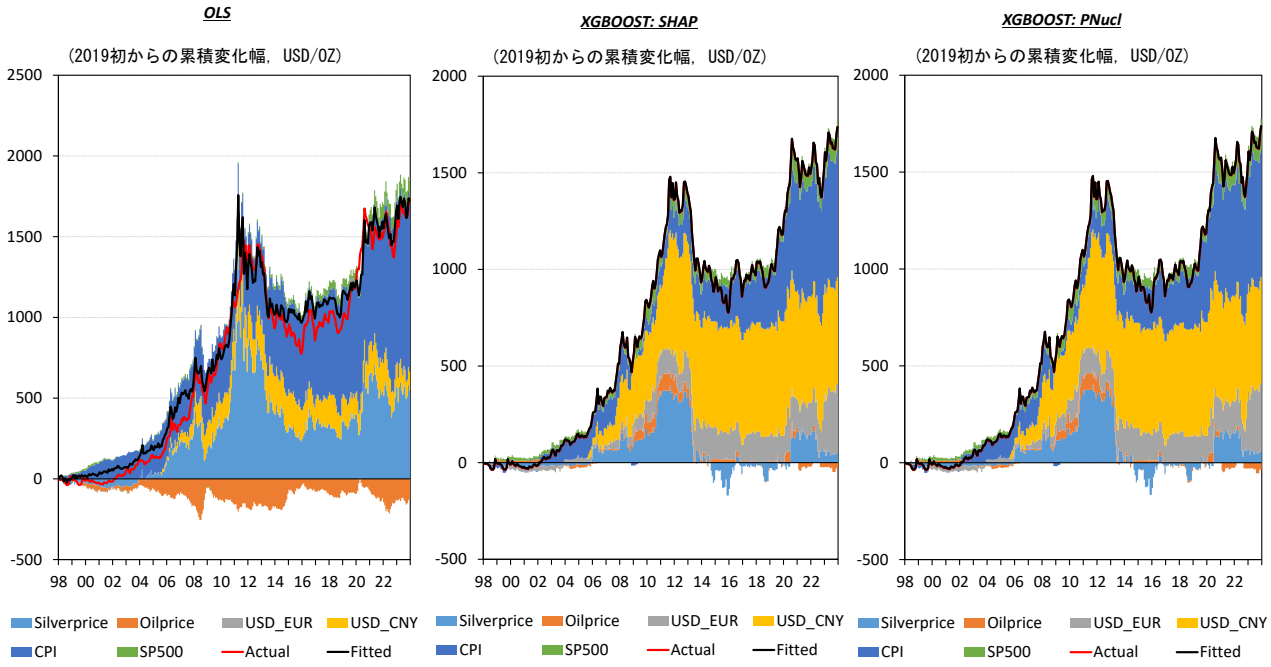


図 20: 金価格の AFA 分解 (2)

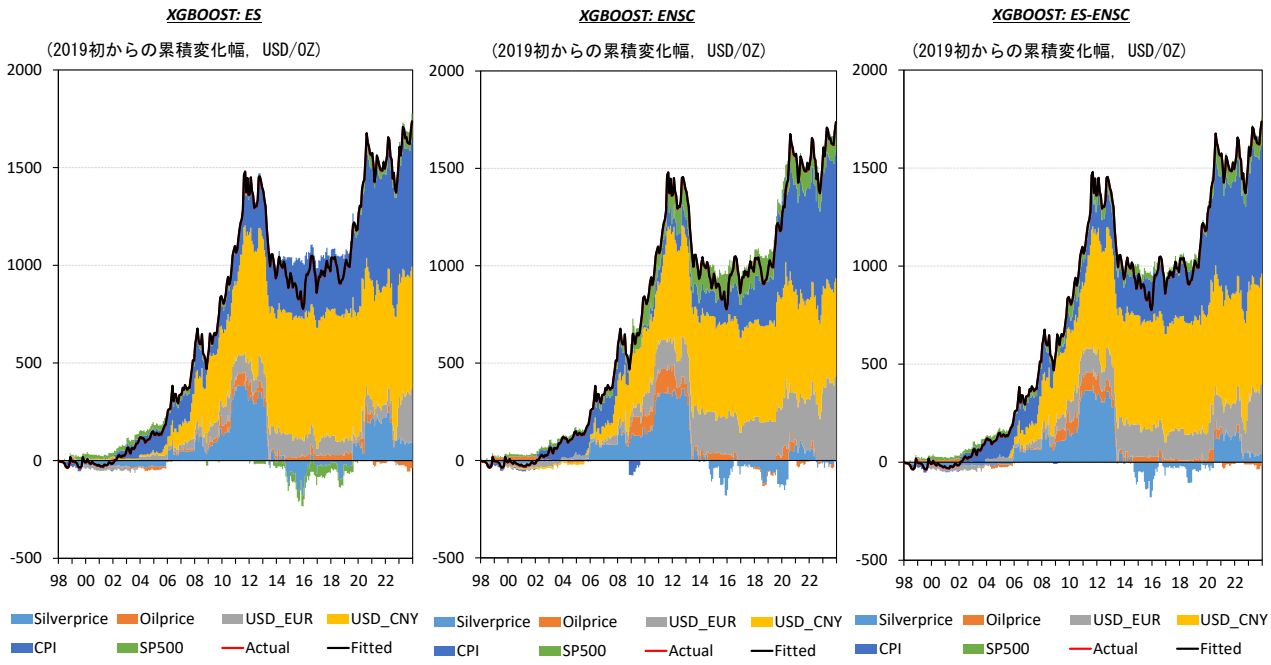


図 21: 金価格の AFA 分解 (3)

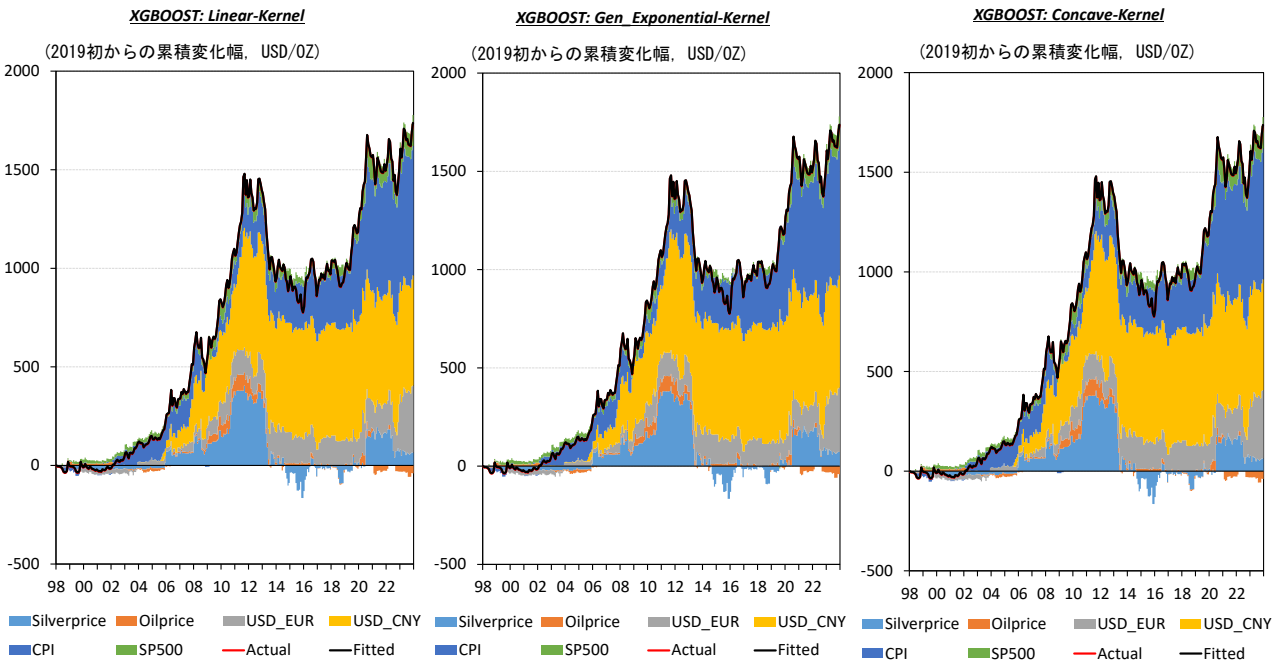


表 4: **AFA 間の相違度 (金価格データ)**

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.