The Institute of Actuaries of Japan
Think the Future, Manage the Risk

# Introduction of the novel IML Method MID to Maximize Interpretability of Black-Box Models

Yoshihiro Matsumori
SOMPO Risk Management Inc.

March 24, 2026
Convention A

# Overview



- We introduce **MID**, a novel method for making black-box models interpretable, developed by Hirokazu Iwasawa.
- MID is a **global, model-agnostic and pragmatic** interpretation method.
- MID provides interpretability through a **decomposition of the prediction function** into main-effect terms and interaction-effect terms.
- We introduce the **Uninterpreted Ratio (UR)**. MID is defined as the method that **minimizes UR**.
- MID is readily available in practice, supported by a **well-developed R package** and related tools.

# Motivation



- Actuaries would like to use advanced machine learning methods with **high predictive accuracy** in practice.

- However, advanced machine learning models often become **black boxes**, making interpretability a key challenge.

- It is not desirable to rely on structurally simple but less accurate models solely for the sake of interpretability.

- By applying IML methods that make highly accurate black-box models **interpretable**, we aim to enable the **practical use** of advanced machine learning techniques in actuarial work.

# Global Model-Agnostic IML Methods



- In the Interpretable Machine learning (IML) field, there are two distinctions:
  - **local vs global** and **model-specific vs model-agnostic**
- Local methods provide interpretations of predictions for individual instances.
- **Global** interpretation addresses the **overall behavioral structure** of prediction functions.
- **Model-agnostic** methods interpret predictive functions **using only input–output relationships** over the space of possible inputs.
- Representative examples include Partial Dependence (**PD**) and Accumulated Local Effects (**ALE**).

# Pragmaticity



- “Interpretation should depend only on the **observable** behavior of the predictive function”

- This is a natural requirement for global model-agnostic interpretations.

- Two prediction functions $f_A$ and $f_B$ are **black-box equivalent**, denoted $f_A \equiv f_B$, if $P(f_A = f_B) = 1$.

- A decomposition method is **pragmatic** if

  $f_A \equiv f_B \Rightarrow$ Decomposition of $f_A$ = Decomposition of $f_B$.

# Notation



- Feature variables: $x_1, \dots, x_d$ denoted by capital letters when treated as r.v.
- $D := \{1, \dots, d\}$
- For non-empty $J \subseteq D$, $\mathbf{x}_J := \{x_j \mid j \in J\}$
- For $J \subseteq D$, $\setminus J := D \setminus J$
- $f: \mathbb{R}^d \to \mathbb{R}$ a given prediction function to be interpreted
- Functional decomposition:

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

# Partial Dependence (PD)



- The 1-dim PD can be viewed as a functional decomposition method whose 1st order terms are defined as:

$$f_{\emptyset} + f_j(x_j) = \mathrm{PD}_j(x_j) := \mathbb{E}\left[f\left(x_j, \mathbf{X}_{\setminus j}\right)\right]$$

- The computational cost is extremely high.

- Performs poorly when the features are strongly correlated.

- Not pragmatic.

# Accumulated Local Effects (ALE)



- Also, the 1-dim ALE can be viewed as a functional decomposition method whose 1st order terms are defined as:

$$f_j(x_j) := \int_{x_{\min,j}}^{x_j} \mathbb{E}\left[\frac{\partial f}{\partial x_j}\middle| X_j = z_j\right] dz_j - c_j,$$

where $c_j$ is defined so as to satisfy $\mathbb{E}[f_j(X_j)] = 0$.

- The meaning of “accumulation” is not intuitive.

- Not pragmatic.

# Maximum Interpretation Decomposition (MID)



- **Uninterpreted Ratio (UR) of order $\boldsymbol{k}$** is defined as follows:

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

- The $k$th-order MID is defined as the functional decomposition

$$f(\mathbf{x}_D) = \sum_{|J|\leq k} f_J(\mathbf{x}_J) + f_D(\mathbf{x}_D)$$

  that minimizes $\mathrm{UR}_k$.

- Computationally efficient.

- **Pragmatic**.

Thank you for your attention.

We welcome feedback and hope many of you will try MID!