Integrated Gradients in Finance: Baseline Effects and Stability across CNNs and Transformers

Research By: Jonathan Chen & Pin-Ju(Ruby) Chen
Supervised By: Dr. Branka Hadji Misheva, Dr. Ali Hursa, Miao Wang

Abstract

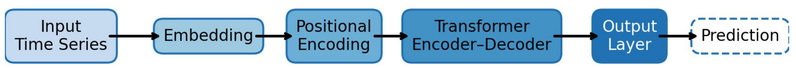
We study how to explain deep models that predict asset returns by asking a simple question: which factors matter, and when? We train a CNN and a Transformer on standard factor inputs and compare gradient-based explanations across market regimes split by K-means and Bayesian change-points. Our key finding: Integrated Gradients depends strongly on the chosen baseline—different baselines can flip signs or rankings. Averaging across several reasonable baselines improves stability. Overall, we recommend: pick baselines carefully, check results across regimes, and validate IG with simpler gradient checks.

Data & Task

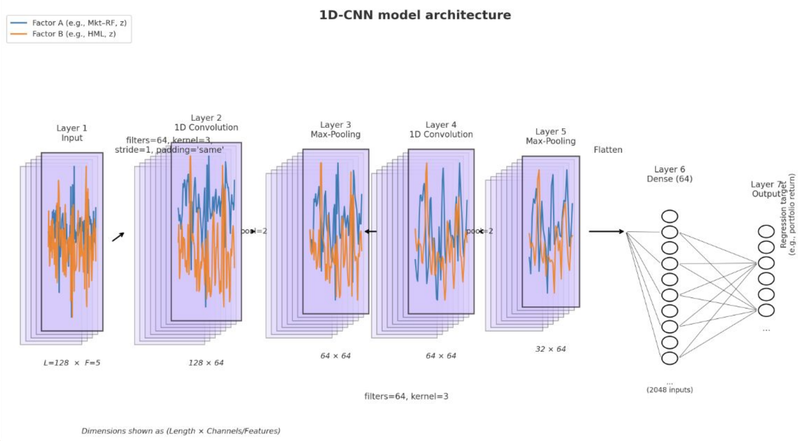
The dataset contains 738 rows of monthly returns of decile portfolios and **Fama–French five factors model** from 1963 to 2024. In this study, we focus on predicting the return of Hi10 (the highest decile portfolio) using the following explanatory variables: Mkt-RF, SMB_constructed, HML, RMW, and CMA.

Model Architectures

Transformer:
The model converts time series into a higher-dimensional representation, adds positional information to capture order, and uses a Transformer to learn patterns across time. The final step predicts the next value based on the learned sequence.



Convolutional Neural Network (CNN):
We built a compact 1D CNN that takes five financial factors as input. It uses two convolution layers (kernel size = 3, channels = 64, padding = 1) to detect how groups of nearby factors interact, and global average pooling with a small fully connected layer (64→1) to keep the model lightweight (~12.7k parameters). The model is trained with mean squared error loss (MSE), Adam optimizer (learning rate = 0.003), batch size = 32, and 10–15 training passes (epochs), which provides stable learning while reducing the risk of overfitting on financial data.
(sample diagram of what it looks like shown below)



Regime Segmentation

We tested four regime segmentation methods: Point-Changing Detection, Spectral Clustering, Bayesian, and K-means. The first two produced unusable splits, while Bayesian and K-means yielded reasonable three-regime partitions with solid performance.Therefore, we proceed with these two methods for our XAI analysis.

- **Bayesian** is a **probabilistic approach** that estimates regime membership based on posterior probabilities.
- **K-means** is a **distance-based** clustering method that groups data around centroids.

Integrated Gradients – Baseline Choices

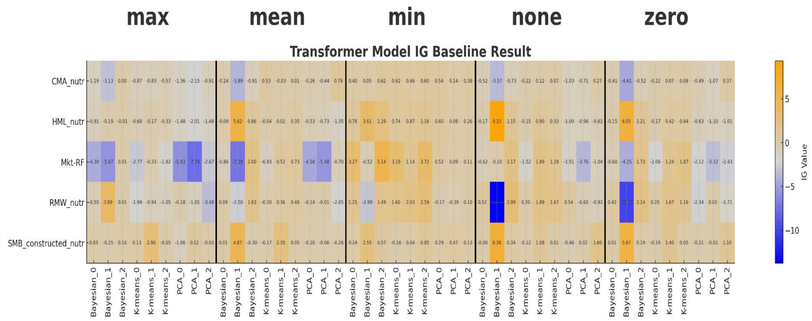
Formula $IG(x)=(x-x')\times 01\partial f(x'+\alpha(x-x'))\partial x d\alpha$

Where **x'** can be…

IG (zero baseline - default)	x' = 0 Relative to no exposure
IG (mean baseline)	x'=E[x] (dataset mean, or regime mean)
IG (random baseline)	x' =xi for a random training point
IG (Min baseline)	x' =min(X) Relative to lowest observed exposures
IG (Max baseline)	x' = max(X). Relative to highest observed exposures
IG (Extreme baseline)	x'=sign(E[X])c(e.g., c=3). Relative to a stress/anchor state

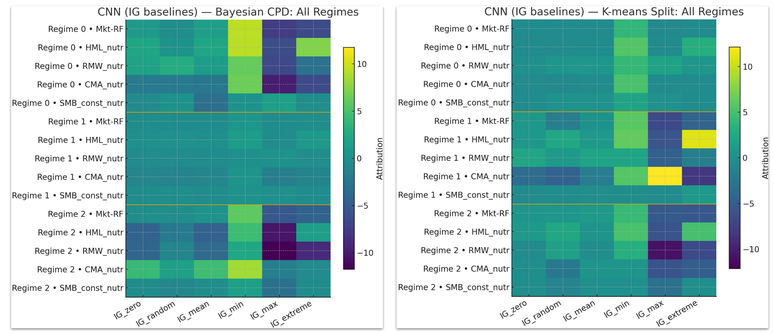
Results

Transformer:



As shown in the heatmap, when the IG baseline is set to the **minimum value**, factor contributions remain stable and consistent across different regimes and segmentation methods. The direction of each factor's influence on the Transformer model does not vary across regimes, and no extreme attribution values are observed, unlike with other baselines. This indicates that the minimum baseline provides the most reliable and interpretable results for IG in this setting.

Convolutional Neural Network (CNN):



Across All Regime

- Lowest impact factor was **SMB**
- Baseline effects:
  - **IG_min** → positive bias
  - **IG_max / IG_extreme** → negative bias
  - **IG_zero / IG_mean / IG_random** → stable
- **Consensus drivers: HML & CMA** explain most variance; **RMW** matters later (often negative).

Bayesian Change Point Method

- **Bayesian—Regime 0: HML↑, CMA↓; Mkt** mixed; **RMW** modest
- **Bayesian—Regime 1: Low-signal** (near-zero local gradients)
- **Bayesian—Regime 2: CMA↑ (robust); RMW↓, HML↓; Mkt** small

K-mans Splitting Technique

- **K-means—Regime 0: HML↑, RMW↑; CMA/Mkt** mixed
- **K-means—Regime 1: High-signal**; **CMA** largest but **sign-split**; **HML↑**
- **K-means—Regime 2: HML↑, RMW↓, CMA↓; Mkt** tends negative

Conclusion and Future Work

Conclusion:
This study highlights the decisive role of **baseline choice in Integrated Gradients (IG)**. For the Transformer model, using the **minimum-value baseline** produced the most stable factor attributions across regimes and segmentation methods, with consistent trend directions and fewer mixed signals. In contrast, for the CNN model, **zero, mean, and random baselines** yielded more reliable results, while min/max choices amplified sign flips and magnitude swings. These findings suggest that **baseline selection must be model- and regime-aware**, and that stability can only be achieved by carefully matching baselines to the underlying architecture and segmentation setting.

Future Research Direction:

- Explore alternative target variables beyond Hi10, to see if the best baseline change with different target
- Investigate instability in other XAI methods
- Use larger datasets to revisit change point detection and spectral clustering