MCIR: A Feature Dependence-Aware Explainability Method with Reliability Guarantees

Anonymous authors
Paper under double-blind review

Abstract

As modern machine learning models are deployed in high-stakes, data-rich environments, the interactions among features have grown more intricate and less amenable to traditional interpretation. Many explanation methods fail when features are strongly dependent. In the presence of multicollinearity or near-duplicate predictors, existing value attribution tools such as SHAP, LIME, HSIC, MI/CMI, and SAGE often distribute importance across redundant features, obscuring which variables represent "important and unique information". This may lead to unstable rankings, jeopardising importance scores, and usually results in a high computational cost. Recent correlation-aware approaches, such as CIR or BlockCIR, offer partial improvements but still struggle to fully separate redundancy from unique contributions at the feature level. To address this, we propose the Mutual Correlation Impact Ratio Method (MCIR-M), a simple and robust measure of global importance under feature dependence. MCIR-M introduces the score Mutual Correlation Impact Ratio (MCIR) that conditions each feature on a small set of its most correlated neighbours and computes a normalized ratio of conditional information having value range, $[0, 1]$, which is comparable across tasks, and collapses to zero when a feature is redundant, enabling clear redundancy detection. In addition to MICR, we introduce a lightweight estimation procedure that requires only a fraction of the data while preserving the attribution behaviour of the full model. Across a synthetic household-energy dataset and the real UCI HAR benchmark, MCIR yields more stable and dependence-aware rankings than SHAP (independent and conditional), SAGE, HSIC, MI-based scores, and correlation-aware baselines such as CIR or BlockCIR. Lightweight explanations preserve over 95% top-feature agreement and reduce runtime by manyfold. These results demonstrate that MCIR-M provides a practical and scalable solution for global explanation in settings with strong feature dependence.

1 Introduction

Artificial Intelligence (AI) plays an increasingly critical role in high-stakes settings such as energy management and healthcare, where model-driven decisions carry significant operational and societal impact. This growing reliance heightens the need for transparent and reliable explanations Lipton (2018); Doshi-Velez & Kim (2017). However, modern explainability methods often break down in environments with strong feature dependence. Small perturbations can distort correlation structure, SHAP methods may arbitrarily distribute credit among redundant predictors Lundberg & Lee (2017); Covert et al. (2020), and information-theoretic or kernel-based measures such as MI and HSIC frequently double-count shared information Kraskov et al. (2004); Gretton et al. (2005b). These issues lead to unstable, inflated, and difficult-to-trust explanations Hooker et al. (2019); Yeh et al. (2019). We argue that dependable global explanations require a dependence-aware attribution score that isolates each feature's *unique* contribution beyond its correlated neighbours.

To tackle these issues, we propose MCIR [1] (Mutual Correlation Information Ratio), a light-weight metric that measures the unique information that a feature provides, particularly when other predictors are closely

[1]Throughout the paper, we refer to the proposed method as MCIR-M, while MCIR denotes the corresponding proposed metric.

related. It does this by comparing two types of mutual information, conditional and marginal. This helps reveal how much influence a feature maintains after considering its correlated neighbors. The score ranges from 0 to 1, values close to zero suggest that a feature is adding little new information, while values close to one indicate a feature is making a significant, unique contribution. This makes MCIR-M useful for assessing importance in various data types, including tabular data, sensor data, and outputs from deep learning models. Overall, MCIR-M provide principled mechanisms for modelling feature–output interactions under statistical dependence, offering robust behaviour aligned with CI goals of stability, adaptability, and reliable decision support in complex environments.

Table 1: Comparison of dependence-aware properties across global attribution families.

Criterion	SHAP / LIME	MI / HSIC / CMI	CCA / PCIR	MCIR-M (ours)
Handles dependence?	Weak (independent backgrounds)	Partial (MI conflates redundancy)	Partial (aligned covariance)	Yes (conditional isolation)
Unique vs. shared contribution	No (mass splitting)	No (shared + unique merged)	No (similar canonical loadings)	Yes (incremental conditional information)
Redundancy collapse	No	No	No	Yes
Scale / normalization	No (arbitrary units)	No (unbounded)	Yes (CIR bounded)	Yes (unit-interval ratio)
Lightweight fidelity	No	No	Partial	Yes (distribution-aligned LW environment)
Estimator stability	Sensitive to sampling / kernel choices	Sensitive to estimator bias/variance	Stable but marginal	Auto-neighbourhood selection (Φ) + estimator switching + bootstrap
Global focus	Local → global aggregation	Global but coarse	Global (vector-output alignment)	Global, dependence-aware

Note: Auto-neighbourhood selection refers to the data-driven construction of Φ from the screened dependence graph. We later refer to this data-driven construction of Φ as automatic neighbourhood selection (*Auto*-Φ).

To ensure this framework is scalable, we introduce a computation strategy that efficiently calculates MCIR without needing to retrain the model, maintaining accuracy even with smaller sample sizes. To that end, we also introduced a method to select estimators that ensure optimal performance across various statistics. Below, we summarize the main contributions of this work.

1. **Mutual Correlation Impact Ratio (MCIR).** We introduce MCIR, a bounded score that measures feature's unique contribution while accounting related features and eliminating redundancy. It helps eliminate redundancy effectively across various data types.
2. **Lightweight and stable computation with guarantees.** Our efficient MCIR computation does not require model retraining, ensuring feature importance order and reducing redundancy, while maintaining accuracy and significantly enhancing computation efficiency.
3. **Comprehensive cross-domain evaluation.** We perform extensive testing and evaluation of MCIR on various datasets- illustrating stable feature rankings, validating redundancy detection and mitigation, and objectively evaluating the quality of competing explanation methods.

We provide theoretical guarantees concerning redundancy, stability under finite samples, and the causal interpretation of MCIR, and we support them through extensive experiments on diverse datasets spanning different dependence structures, noise levels, and modeling conditions. The UCI Human Activity Recognition (HAR) dataset assesses performance using high-dimensional sensor signals, focusing on the strong correlations between accelerometer and gyroscope data. The House Energy Simulation Dataset includes temporal and weather-related attributes, which highlight issues of multicollinearity and nonlinear effects. The Norwegian Regional Load-Zone Data merges electricity load data with weather inputs, presenting a mix of predictor types and evolving correlations. Finally, the CIFAR-10 Deep Representations dataset employs 2,048-dimensional embeddings from a fine-tuned ResNet-50 model to evaluate MCIR in high-dimensional vision contexts. Overall, MCIR-M offers an efficient way to understand the contributions of different features. This combined approach addresses key limitations of existing methods when features are highly interdependent, such as their

tendency to inflate importances in the presence of redundancy or to collapse when dependence violates their underlying independence assumptions, and is ideal for complex real-world scenarios.

Across synthetic dependence families, UCI HAR, HouseEnergy-Sim, Norwegian load zones (NO1–NO5), and deep embeddings from CIFAR-10, MCIR consistently provides stable, redundancy-aware global attributions. Unlike marginal or kernel-based baselines, MCIR collapses correlated feature blocks while preserving the predictive information captured by the model, yielding higher fidelity and substantially lower redundancy than PCIR, SHAP, MI, and HSIC. In lightweight settings, MCIR-M maintains high rank agreement with full-data explanations (often exceeding 95% top-K overlap) while reducing runtime by factors of 3-9. For high-dimensional deep vision features (e.g., ResNet-50 embeddings), MCIR-M produces smooth deletion curves and compact rankings aligned with semantic feature clusters, demonstrating effectiveness beyond tabular and sensor domains. Overall, these findings confirm that MCIR-M offers a robust and scalable dependence-aware global explanation method across diverse real-world scenarios. Table 1 compares the dependence-aware strengths and weaknesses of SHAP/LIME, MI/HSIC/CMI, CCA/PCIR Sengupta et al. (2025), and MCIR-M for n observations and k features.

Paper Overview: We begin by orienting the reader with a clear and streamlined overview of the paper's structure. Section 2 reviews the limitations of current global attribution methods. We then examine these limitations in the context of strong feature dependence, motivating the development of a dependence-aware measure. Section 3 introduces the notation and lightweight environment framework used throughout the paper. Section 4 formally presents our proposed method, MCIR-M, detailing its information-theoretic formulation and principal redundancy-collapse guarantees. Section 4.2 develops the theoretical properties of MCIR, including boundedness, estimator stability, and fidelity under lightweight computation. Section 4.4 discusses computational considerations and estimator selection. Section 5 presents comprehensive empirical evaluations across synthetic, sensor, energy, and deep-representation datasets. The results, including synthetic benchmarks, sensor and energy evaluations, deep-representation analyses, the case study in Section 6.8, and the overall discussion in Section 6.9, are detailed in Section 6. Finally, Section 7 reports the required ethics and reproducibility statements for TMLR. Section 8 summarizes the main findings. Detailed proofs, supplementary algorithms, and extended experimental results are provided in the appendices.

2 Background and Related Work

In real-world datasets, it is common to encounter groups of covariates? that are correlated or redundant. This creates a challenge for global attribution methods, which need to differentiate between shared contributions and those that are unique to individual predictors. Traditional importance measures, like permutation importance Breiman (2001), marginal relevance scores, and impurity-based metrics, often overestimate the significance of correlated variables Strobl et al. (2008). This can result in rankings that are unstable or misleading. Shapley-value explainers, such as SHAP Lundberg & Lee (2017) and SAGE Covert et al. (2020), determine importance by calculating the marginal contributions of features within groups or coalitions. Although these methods are theoretically sound, they typically operate under the assumption that background distributions are independent. Alternatively, they often rely on perturbation sampling, which disrupts the natural dependencies in the data Sundararajan et al. (2020), and are computationally expensive. Consequently, they may inaccurately allocate credit to redundant predictors and exhibit high variability when features are correlated. Other perturbation-based evaluations for faithfulness, like ROAR Hooker et al. (2019) and deletion tests Samek et al. (2017), face similar issues. Measures such as mutual information (MI) Cover & Thomas (2006a), conditional mutual information (CMI) Kraskov et al. (2004), and kernel-based measures like HSIC Gretton et al. (2005a) focus on quantifying nonlinear relationships but have some limitations. They are unbounded and do not sufficiently isolate conditional effects. MI often counts shared information between correlated predictors multiple times, while CMI can be unstable in high-dimensional settings Gao et al. (2017). Furthermore, these measures lack normalization, complicating comparisons between different datasets or model classes. Recent studies further highlight the challenges of dependence-aware attribution in modern machine learning systems. Conditional SHAP extensions, such as KernelSHAP with conditional sampling Aas

[1]By *local dependence neighborhood*, we refer to the small set of features that exhibit the strongest statistical dependence with a target feature, typically identified via a fast dependence sketch (e.g., correlation or distance correlation graph) and used as the conditioning set for MCIR.

et al. (2021), attempt to preserve feature correlations but remain sensitive to background choice and sampling variance. Causal attribution formulations, e.g., Causal-Shapley scores Janzing et al. (2020) and interventional SHAP Merrick & Taly (2020), provide principled ways to avoid over-counting shared information, yet require explicit causal models or strong independence assumptions that rarely hold in practice. Stability-focused works Ghorbani et al. (2019); Slack et al. (2020) demonstrate that many post-hoc explainers can be highly unstable or even manipulated under correlated predictors.

More recently, redundancy-aware feature selection and attribution methods such as RFA Li et al. (2023) and dependency-aware interaction attribution Tsang et al. (2020) propose grouping or interaction modelling, but they do not provide bounded, normalized scores or guarantees of redundancy collapse. These developments reinforce the need for explanation methods that remain reliable under correlation, provide interpretable scaling, and isolate unique contributions without relying on causal graphs or extensive sampling assumptions. Very recent work has intensified interest in dependence-aware explainability. Copula-based attribution models Zhang & Müller (2024); Aas et al. (2024) propose more faithful conditional background sampling, yet remain computationally demanding and sensitive to estimator choice. Robustness studies Han & Kim (2024); Covert & Lee (2025) show that many Shapley formulations exhibit instability under correlation shifts, leading to inconsistent rankings across subsamples. Scalable global attribution frameworks Cheng & Zhao (2024); Liu & Huang (2025) introduce grouping or low-rank structures to mitigate redundancy, but they do not provide bounded scores nor theoretical guarantees of redundancy collapse. These recent developments highlight that despite progress, current methods still lack a unified mechanism that combines: (i) conditional isolation of unique contributions, (ii) normalized and comparable scoring, and (iii) stability under lightweight computation. MCIR-M directly addresses these gaps. The ExCIR and other correlation-ratio measures Hotelling (1936) are based on canonical correlation analysis (CCA) provide a way to quantify dependence that is bounded. Variants like HSIC-Lasso Yamada et al. (2014), BlockCIR, and CC-CIR Sengupta et al. (2025) capture significant aspects of the shared structure or cross-covariance geometry. However, these methods do not effectively isolate the impact of individual predictors and may not adequately handle redundant predictors, which limits their usefulness in settings with strong dependence. In this paper, we introduce a new term for the ExCIR method: PCIR, which stands for Partial Correlation Impact Ratio. We chose this name to highlight that PCIR identifies only partial dependencies in the data, using a technique known as canonical correlation. The PCIR method is part of a unified framework that includes other methods called BlockCIR and CC-CIR. This naming convention allows for easier comparison between PCIR (along with BlockCIR and CC-CIR) and other techniques like HISC-Lasso, MI, CMI, and our new method MCIR-M. To clarify, in this paper, whenever we mention PCIR or ExCIR, we are referring to the same method. To clarify how our formulation departs from prior CIR-family variants, Table 2 provides a structured comparison across dependence modeling, redundancy behavior, boundedness guarantees, estimator stability, and computational properties. Across the various methods discussed, two fundamental issues persist: (i) shared information is often over-counted, leading to inflated importance scores, and (ii) explanations can become unstable due to subsampling, estimator noise, or groups of correlated features Covert & Lee (2021). Problems such as unbounded scoring, assumptions of independence, and the lack of conditional adjustment hinder the reliability of existing explanation methods in high-dependence scenarios. To this end, we propose MCIR-M to address these limitations.

3 Preliminaries

In this section, we first present the essential notation that will be used throughout the paper and then, motivate MCIR-M . We define $F \in \mathbb{R}^{n\times k}$ as the feature matrix, which consists of n observations and k features, and $Y \in \mathbb{R}^n$ as the corresponding model outputs. The combination of these two is referred to as an environment, denoted as $\mathcal{U} := \mathcal{D}(F, Y)$. This environment represents the joint distribution of inputs and outputs on which a model is both trained and evaluated. As with other explainers, MCIR does not require access to the full training distribution at inference time; it only assumes that explanations are generated with respect to an observed environment, which may be complete or partial. To enable scalable and distribution-faithful attribution, we construct a *lightweight environment* $\mathcal{U}' = \mathcal{D}(F', Y')$. This lightweight version reflects a partial observation of the full environment, either through fewer samples or reweighted samples, while still preserving the key statistical and predictive characteristics needed for reliable attribution. We achieve this

Table 2: Differentiation of MCIR from prior CIR-family variants and canonical-correlation approaches.

Aspect	PCIR (ExCIR)Sengupta et al. (2025)	BlockCIRSengupta et al. (2025)	CC-CIRSengupta et al. (2025)	MCIR-M (Ours)
Dependence scope	Global canonical correlation (scalar/vector outputs)	Grouped / class-conditioned canonical alignment	Cross-covariance or kernel canonical correlation	Conditional mutual-information ratio with adaptive local dependence neighborhood [2]
Redundancy handling	None (aggregative)	Partial within-block averaging	Linear cross-covariance regularization	Explicit redundancy collapse through conditioning on local dependence neighborhood
Boundedness source	Normalized correlation ratio	Same as PCIR (block average)	Implicit via kernel normalization	Derived from MI–CMI decomposition; proven boundedness
Estimator stability	Empirical; sensitive to covariance noise	Requires regularized CCA	Kernel bandwidth–dependent	Formal rank-stability bound under estimator perturbation
Lightweight fidelity	Requires full environment	Same	Not defined	Lightweight (LW) contract ensuring ranking preservation under environment similarity
Redundancy-collapse proof	Absent	Heuristic grouping	None	Information-theoretic proof of zero score under conditional redundancy
Cross-domain validity	Tabular / vector	Tabular / grouped	Kernelized nonlinear	Generic (tabular, vision, text) with estimator switching
Computational cost	$\mathcal{O}(nk)$	$\mathcal{O}(nk^2)$	$\mathcal{O}(k^3)$	$\mathcal{O}(n'k)$ via local dependence graph (scalable, model-free)
Reliability quantification	None	None	None	Introduces Explanation Reliability Index (ERI) combining fidelity, redundancy, stability

through stratified subsampling over output quantiles combined with kernel herding, a standard approach in coreset construction Campbell & Broderick (2019); Feldman (2020). This ensures that $\mathcal{U}'$ maintains both the marginal distribution of outputs and the dependence structure among features, allowing MCIR to generalize naturally even when only a subset of the original environment which is *similar*, is available.

We define two environments as *similar* when their outputs exhibit the same functional behaviour, even if they differ by rotations, shifts, or other admissible transformations. The motivation for introducing this notion of similarity is to formalize the requirement that the lightweight environment $\mathcal{U}'$ should preserve the predictive structure of the full environment $\mathcal{U}$. In other words, $\mathcal{U}'$ is constructed to be *similar* to $\mathcal{U}$ in the sense that the model would behave comparably on both, thereby ensuring that feature attributions computed on the lightweight environment remain faithful to those computed on the full environment. Specifically, if we can find a rotation (which we can represent with an orthogonal matrix R) and a shift (represented by a translation vector t) such that the outputs of the two environments are related such that: $Y' \approx RY + t$, then we consider the core characteristics of the environments to be unchanged. This means we can compare the two environments without worrying about their positioning or orientation, allowing us to focus on concrete aspects of model behavior, namely, their output distributions, feature rankings, and explanation patterns. To put this into practice, we assess three criteria: (i) whether the output distributions are statistically indistinguishable, (ii) whether feature rankings remain consistent, and (iii) whether explanation patterns are stable across the full and lightweight environments.

To make this idea clear, below we exemplify it with a real-world scenario from the energy sector. Imagine a city's electricity provider wants to predict future power usage. The full environment ($\mathcal{U}$)) is like having a detailed system that collects real-time data from every single smart meter in every home and business. This system knows exactly when and where energy is used, and can make extremely accurate predictions, but it is expensive and complex to run. The lightweight environment (($\mathcal{U}'$)) is a simpler version. Instead of collecting data from every smart meter, it collects data from a handful of key locations or uses daily summaries instead of minute-by-minute readings. It's much faster and cheaper, but less detailed. If both systems can still predict the city's overall energy demand patterns, such as when peak usage will occur or how much electricity will be needed, they are called *similar* environments. This means we can trust insights from the lightweight system for the full system, even though the lightweight one is much simpler. To measure this similarity, we employ transformations on the Stiefel manifold Absil et al. (2009) alongside an f-divergence-based distance. We

minimize a risk objective, $L(Y, Y')$, defined as follows:

$$L(Y, Y') = D_f\big(p(Y) \,\|\, p(Y')\big) + \lambda \,\text{RankDisagree}(Y, Y'), \tag{1}$$

which ensures that the explanatory insights are aligned. Here, $D_f\big(p(Y) \,\|\, p(Y')\big)$ is an f-divergence term (e.g., KL, JS, or Hellinger) measuring how different the output distributions of the full and lightweight environments are, and is minimized when $p(Y)$ and $p(Y')$ are statistically indistinguishable. The second term, RankDisagree(Y, Y'), quantifies how much the feature rankings implied by MCIR change between environments (e.g., via Kendall–τ or Jaccard@K), penalizing mismatched explanatory behaviour. The parameter $\lambda > 0$ balances distributional alignment and ranking consistency.

To make the similarity objective operational, we now describe the notation associated with the lightweight environment. Let $F' = [f_1, \ldots, f_k] \in \mathbb{R}^{n' \times k}$ denote the reduced feature matrix with $n' < n$ observations, where each feature is written as a vector $f_i = (f_{1i}, \ldots, f_{n'i})^\top$. The corresponding model outputs are

$$Y' = M'(F') = \big[\, M'(\mathbf{x}'_1), \ldots, M'(\mathbf{x}'_{n'}) \,\big]^\top,$$

with $\mathbf{x}'_j$ denoting the j-th input in F'. Thus, the lightweight environment evaluates the same trained model on a carefully selected subset of the original input space.

For later use in defining MCIR, we introduce the empirical means

$$f_i = \frac{1}{n'} \sum_{j=1}^{n'} f_{ji}, \qquad \bar{y}' = \frac{1}{n'} \sum_{j=1}^{n'} y'_j,$$

and define the *joint reference level*

$$m_i = \frac{f_i + \bar{y}'}{2},$$

which provides a common center for computing the joint and total dispersion terms that appear in the MCIR formulation.

Next, we formally define the PCIR score earlier used in table2, which extends the ExCIR formulation proposed in Sengupta et al. (2025) to cases where only partial dependencies are captured.

Definition 1 (**Partial Correlation Impact Ratio**). *PCIR assigns to each feature i a unit-interval score $\eta_{f_i} \in [0, 1]$ that contrasts a joint (between-level) dispersion with a total (around-pooled-mean) dispersion:*

$$\eta_{f_i} \;=\; \frac{n' \big[(f_i - m_i)^2 + (\bar{y}' - m_i)^2\big]}{\sum_{j=1}^{n'} (f_{ji} - m_i)^2 \;+\; \sum_{j=1}^{n'} (y'_j - m_i)^2}. \tag{2}$$

Intuitively, the numerator measures how far the feature and the output means lie from their pooled center m_i, while the denominator aggregates the total dispersion of *all* observations of the pair (f_i, Y') around m_i. PCIR quantifies how strongly the *population-level* variability of feature i is aligned with the variability of the output. η_{f_i} approaches 1 when the movements of f_i and Y' are tightly aligned so that most of the total dispersion is explained by their joint displacement from m_i. η_{f_i} approches 0 when f_i behaves as noise relative to Y', i.e., the total dispersion dominates the joint displacement. The key properties of PCIR are Sengupta et al. (2025): **1. Boundedness and comparability.** $\eta_{f_i} \in [0, 1]$ by construction, enabling cross-feature and cross-dataset comparisons. **2. Monotonicity under stronger association.** Strengthening co-variation between f_i and Y' increases the joint term relative to the total, raising η_{f_i}. **3. Noise suppression.** For uninformative features, the joint term is small relative to the total, driving $\eta_{f_i} \to 0$. **4. Estimator-agnostic computation.** Equation 2 uses only sample means and sums of squares; no distributional assumptions are required.

In weak-dependence regimes, PCIR suffices for global attribution. However, as we later experimentally validated, under strong mutual dependence, PCIR may distribute credit across correlated variables. To address this challenge, we propose a novel metric, MCIR, formally introduced in the next section, that identifies and condition on a small set of correlated partners to quantify *unique* contribution.

4 MCIR: Formal Definitions and Guarantees

This section explains the motivation behind MCIR-M in a clearer and more accessible way by simplifying the main ideas while preserving technical correctness. AI models often rely on large sets of features, many of which are strongly correlated. In such settings, widely used explanation methods—such as SHAP, LIME, HSIC, MI/CMI, or the CIR-family scores, tend to split attribution between redundant features. This creates three major problems: **unstable feature rankings:** small changes in data or sampling can reshuffle the importance of highly correlated variables, **misleading importance scores:** methods often inflate the importance of variables that simply "move together," even when they do not provide unique information about the output, and **high computational cost:** methods relying on repeated model calls or kernel evaluations become inefficient on large datasets.

To address these challenges, MCIR directly measures how much *unique* information a feature contributes after accounting for its most strongly correlated neighbours. Let $f_i \in \mathbb{R}^{n'}$ denote the i-th feature vector and let $f_\Phi = \{f_j : j \in \Phi\}$, where each $f_j \in \mathbb{R}^{n'}$ shares the same sample dimension, represent a small neighbourhood of correlated features. A feature f_i, can be simply expressed as a vector of values, while a group of neighboring features can be stacked together to form a matrix called a "feature block." This means we form a joint structure that includes both the feature in question and its neighboring features, allowing us to analyze how much information they collectively provide. The joint feature block $(f_i, f_\Phi) \in \mathbb{R}^{n' \times (|\Phi|+1)}$ therefore forms a compatible parameter space in which joint mutual information captures both shared and unique effects. However, a high joint mutual information term $I(Y'; f_i, f_\Phi)$ may simply reflect redundancy within this block, indicating that f_i does not provide additional useful information beyond what is already contained in f_Φ. For instance, if two sensors are measuring the same process, their information could be redundant, making both appear important when, in reality, only one of them is adding valuable insights e.g., if f_i is a perturbed version of f_j in the presence of small, independent noise. Intuitively, this situation corresponds to *redundancy*: the neighbourhood f_Φ already contains all the predictive structure that f_i can offer, so the joint mutual information $I(Y'; f_i, f_\Phi)$ becomes large not because f_i contributes new signal, but because it is statistically entangled with variables that are already informative. In such cases, f_i behaves as a "duplicate" or "shadow" feature whose behaviour is largely predictable from f_Φ, and therefore it adds little or no unique information about Y'. Let, ε denotes a small independent noise term. Throughout this paper, we assume ε is Sub-Gaussian (Gaussian noise being a common special case) with $\mathrm{Var}(\varepsilon) \ll \mathrm{Var}(f_j)$. This standard assumption ensures that ε does not introduce structured dependence with Y' or the neighbourhood f_Φ. Importantly, the precise distribution of ε does not affect the redundancy-collapse behaviour of MCIR; as $\varepsilon \to 0$, we continue to have $I(Y'; f_i \mid f_\Phi) \to 0$ whenever f_i is a near-duplicate of a neighbour. ε, such that $f_i = f_j + \varepsilon$ with $\mathrm{Var}(\varepsilon) \ll \mathrm{Var}(f_j)$, the joint MI remains high while the conditional MI $I(Y'; f_i \mid f_j)$ effectively vanishes. Traditional methods may then misrepresent their significance, giving inflated importance to both features. MCIR addresses this by focusing on the conditional component $I(Y'; f_i \mid f_\Phi)$, which isolates the incremental contribution of f_i once its correlated partners are known. This conditional term is then normalized to produce a bounded, comparable score that collapses to zero when f_i is redundant, ensuring that duplicate or near-duplicate predictors do not receive artificial credit.

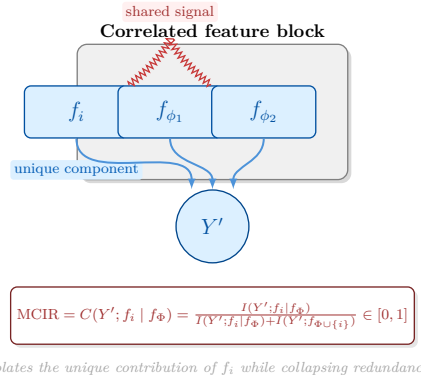

Figure 1: MCIR intuition: f_i and its correlated partners $f_\Phi = \{f_{\phi_1}, f_{\phi_2}\}$ share substantial redundant signal. MCIR conditions on f_Φ to isolate the *unique* increment contributed by f_i, normalised into a bounded $[0, 1]$ ratio.

4.1 Intuition Behind MCIR

MCIR introduces a normalized ratio $C(Y'; f_i \mid f_\Phi)$ ranging from 0 to 1 as depicted in Figure 1. The motivation for this ratio arises from the decomposition of joint mutual information: the joint term $I(Y'; f_i, f_\Phi)$ contains

both the information that is *shared* with the correlated neighbourhood f_Φ and the *unique* contribution of f_i. Simply relying on joint or marginal measures therefore overestimates importance when redundancy is present. By isolating the conditional component $I(Y'; f_i \mid f_\Phi)$, MCIR captures only the incremental information that f_i contributes beyond what is already explained by f_Φ. Normalizing this conditional term by the total explainable association mass, produces a bounded, comparable score in $[0, 1]$ that reflects the *proportion* of unique information attributable to f_i. A value close to 1 signifies that f_i provides substantial unique signal, while a value near 0 indicates redundancy. This score allows for consistent comparisons across different datasets and models. In summary, MCIR: (1) conditions on a small, data-driven neighbourhood to extract the **unique** incremental information a feature provides beyond its correlated partners; (2) reports a **unit-interval** score that enables stable cross-feature and cross-dataset comparison; (3) provably collapses redundancy under multicollinearity; and (4) integrates with a distribution-aligned lightweight environment so that explanations computed on fewer observations faithfully mimic those of the full model.

At a high level, this section answers three questions:

1. ***What is MCIR?*** We first formally define MCIR using conditional and joint mutual information, and explain how it isolates the unique contribution of each feature.
2. ***Why is MCIR well-behaved?*** We then show that MCIR is bounded by $[0, 1]$, collapses redundancy under strong dependence, and remains stable under sampling noise.
3. ***How expensive is it to compute?*** Finally, we describe how MCIR can be estimated efficiently in a lightweight environment, and how estimator choice affects robustness.

Readers mainly interested in intuition can focus on the informal explanations and summaries, while those seeking guarantees can follow the accompanying theorems and propositions. This intuition is illustrated in Figure 1, which highlights how MCIR separates the information that is uniquely attributable to a feature from the portion that is shared with its correlated neighbours.

Let $F \in \mathbb{R}^{n\times k}$ denote the feature matrix with columns $\{f_1, \ldots, f_k\}$, and let $M' : \mathbb{R}^k \to \mathbb{R}$ be a fixed trained predictor. The model output in the lightweight environment is then obtained by evaluating M' row-wise on F,

$$Y' = M'(F) := [M'(x_1), \ldots, M'(x_n)]^\top \in \mathbb{R}^n.$$

In this context, consider a set of data observations labeled as x_j, where j indicates the specific observation number within a dataset denoted as F. When we focus on a particular target index i, which can take on values from the set $\{1, \ldots, k\}$, we can create a subset of indices called Φ. This subset includes some indices from the total set of indices, excluding the target index i.We define a feature block f_Φ as a group of feature vectors corresponding to the indices included in Φ, arranged in a way that each feature vector becomes a column in this block. Mathematically, we express this as $f_\Phi := [\, f_j \,]_{j\in\Phi} \in \mathbb{R}^{n'\times|\Phi|}$, where n' indicates the number of observations and $|\Phi|$ stands for the number of indices in the subset Φ. Furthermore, if we want to include the target index i in our feature block, we extend our feature block to include f_i as well. The $f_{\Phi\cup\{i\}}$, contains all the features from both the subset Φ and the extra feature corresponding to the target index i. Thus, f_Φ and $f_{\Phi\cup\{i\}}$ are not sets but submatrices of F' used when computing joint and conditional information.

All information–related measures are defined with respect to the joint probability law of the variables $(Y', f_1, \ldots, f_k)$. When we refer to densities, we mean densities defined relative to an appropriate base measure. For continuous variables, the reference measure is the Lebesgue measure (e.g., (Cover & Thomas, 2006b, Ch. 2); (Gray, 2011, Sec. 3.1)); for discrete variables, the counting measure is used ((Cover & Thomas, 2006b, Ch. 2)); and for mixed discrete–continuous vectors, hybrid product measures are employed following standard probability-theory conventions (see (Kallenberg, 2002, Ch. 1)). These base measures ensure that mutual information and conditional mutual information are well defined for all variable types. We now formalize the information–theoretic components used to construct MCIR.

Assumption 1. ***(i)*** *The relevant joint or conditional laws admit densities or mass functions so that all mutual information (MI) and conditional MI (CMI) are finite.* ***(ii)*** *Conditioning events have positive*

probability; regular conditional distributions exist. ***(iii)*** *Estimators used later satisfy standard consistency and concentration properties (Assumption 2).*

Assumption 1 is standard in information–theoretic analysis and ensures that all quantities used in our definitions are mathematically well defined. Condition (i) guarantees that the joint and conditional distributions have densities or mass functions, which is necessary for mutual information (MI) and conditional MI (CMI) to be finite rather than undefined or infinite. Condition (ii) rules out degenerate conditioning events of zero probability, ensuring that regular conditional distributions exist and that CMI terms such as $I(X;Y \mid Z)$ are well posed. Finally, condition (iii) ensures that the estimators used later concentrate around their population values, which is required for the stability and boundedness results of MCIR. Together, these mild assumptions exclude only pathological cases and are routinely satisfied by standard tabular, vision, and time-series datasets.

Definition 2 (**Conditional Mutual Information (CMI)**). *For a target feature f_i and conditioning set Φ,*

$$I(Y'; f_i \mid f_\Phi) = \mathbb{E}\left[\log \frac{p(Y' \mid f_i, f_\Phi)}{p(Y' \mid f_\Phi)}\right] = D_{\mathrm{KL}}(p(Y' \mid f_i, f_\Phi) \,||\, p(Y' \mid f_\Phi)) \;\geq 0. \tag{3}$$

Here $D_{\mathrm{KL}}(P \,||\, Q)$ denotes the Kullback-Leibler divergence, which measures how different a distribution is from another reference distribution.

Definition 3 (**Joint Mutual Information (JMI)**). *For a feature index set $S \subseteq \{1, \ldots, k\}$, the joint mutual information between Y' and the feature block f_S quantifies the deviation from statistical independence. In particular, if Y' and f_S were independent, their joint density would factorize as $p(Y', f_S) = p(Y')\, p(f_S)$. The joint mutual information is defined as*

$$I(Y'; f_S) = D_{\mathrm{KL}}\Big(p(Y', f_S) \,\Big\|\, p(Y')\, p(f_S)\Big) = \mathbb{E}\left[\log \frac{p(Y', f_S)}{p(Y')\, p(f_S)}\right] \;\geq 0, \tag{4}$$

where equality holds if and only if Y' and f_S are independent.

Definition 4 (**Mutual Correlation Impact Ratio (MCIR)**). *Given i and Φ, the* MCIR *score is*

$$C(Y'; f_i \mid f_\Phi) = \frac{I(Y'; f_i \mid f_\Phi)}{I(Y'; f_i \mid f_\Phi) + I\big(Y'; f_{\Phi \cup \{i\}}\big)} \;\in\; [0, 1]. \tag{5}$$

Remark 1. *$I(Y'; f_i \mid f_\Phi)$ isolates the* unique *contribution of f_i beyond its correlated partners f_Φ. The joint term $I(Y'; f_{\Phi \cup \{i\}})$ stabilizes scale across tasks and dependence strengths. Thus C reports the fraction of* explainable association mass *uniquely attributable to f_i after accounting for partners.*

For a neighbourhood Φ, interpreted here as a small set of features that exhibit high statistical dependence with f_i (e.g., identified via correlation or mutual-information screening), define, $U_i := I(Y; f_i \mid f_\Phi)$ and $J_i := I(Y; f_{\Phi \cup \{i\}})$. Then, $\mathrm{MCIR}_i = \frac{U_i}{U_i + J_i}$ measures the fraction of explainable association uniquely attributed to f_i, accounting for shared contributions in f_Φ. We do not encode environment notation inside the definition of MCIR because the score is a purely local information-theoretic quantity. The environment contract (full vs. lightweight) is introduced later only to guarantee stability and ranking preservation.

$$\mathrm{MCIR}_i = \frac{U_i}{U_i + J_i}$$

is already a complete and self-contained formal definition. Unlike ExCIR/BlockCIR Sengupta et al. (2025), which mix shared and unique effects through marginal or aligned covariance, MCIR normalizes conditional information to isolate uniqueness while maintaining a unit-interval scale for cross-task comparison. If f_i behaves like a near-duplicate of some $f_j \in \Phi$, then $I(Y'; f_i \mid f_\Phi) \to 0$ and $\mathrm{C}(Y', f_i \mid \Phi) \to 0$, expressing *redundancy collapse* (no extra credit for copies). If f_i carries signal that is not present in f_Φ, then $I(Y'; f_i \mid f_\Phi)$ dominates and $\mathrm{C}(Y', f_i \mid \Phi) \to 1$, highlighting *unique* drivers. In weak-dependence regimes, where conditioning has little effect and $I(Y'; f_i \mid f_\Phi) \to I(Y'; f_i)$, the ranking induced by MCIR coincides with that of PCIR and other marginal global scores. In this way, MCIR, **isolates unique contributions by conditioning**

on correlated neighbours, yields unit-interval scores that are directly comparable across tasks and datasets, and collapses redundancy under multicollinearity, avoiding the credit-splitting behaviour often observed with SHAP, SAGE, MI, or HSIC. These properties make MCIR particularly suited for domains with strongly dependent features, such as time-series lags, sensor networks, or engineered feature blocks.

Proposition 1 (**Uniqueness & invariances**). *For any admissible i, Φ: (i)* $\mathrm{MCIR}_i = 0 \iff I(Y; f_i \mid f_\Phi) = 0$ *(conditional redundancy); (ii) under rank–Gaussianization and a Gaussian–copula MI/CMI estimator,* MCIR_i *is invariant to strictly monotone transforms of (f_i, f_Φ, Y); (iii) in the weak-dependence limit where $I(Y; f_i \mid f_\Phi) \approx I(Y; f_i)$, MCIR induces the same ordering as PCIR.*

Let $\mathcal{C}$ and $\mathcal{D}$ partition features into continuous and discrete index sets. Definitions 2–7 remain valid in the mixed case. In practice: **(i) Gaussian-copula MI/CMI for continuous or mixed (after rank-Gaussianization), (ii) kNN MI/CMI for nonparametric continuous settings, (iii) plug-in MI/CMI for discrete.**

Assumption 2. *There exist absolute constants $(c_1, c_2) > 0$ such that for mutual-information or conditional mutual-information estimators $\widehat{I}$ built from n' i.i.d. samples,*

$$\Pr\Big(\big|\widehat{I} - I\big| > \delta \Big) \;\le\; c_1 \exp(-c_2\, n'\, \delta^2), \tag{6}$$

or equivalently $\big|\widehat{I} - I\big| = O_P(n'^{-1/2})$. The notation $O_P(n'^{-1/2})$ denotes stochastic boundedness at rate $n'^{-1/2}$. This assumption holds for a broad class of kNN-based estimators (e.g., KSG and its conditional variants) and Gaussian–copula estimators under mild smoothness and density-regularity conditions Kraskov et al. (2004); Gao et al. (2017); Singh & Póczos (2016); Berrett et al. (2019). We require only consistency and sub-Gaussian concentration of the estimator, not exact parametric convergence rates.

Assumption 2 is mild and standard in the information–theoretic estimation literature. MI and CMI cannot be computed in closed form for arbitrary data distributions, so we must rely on empirical estimators such as kNN-based (KSG) and Gaussian–copula methods. These estimators are known to be consistent and to satisfy sub-Gaussian concentration under broad smoothness and regularity conditions, which makes the stated bound realistic for practical tabular, vision, and time-series datasets. This assumption is essential for our analysis because MCIR is defined through ratios of MI and CMI terms: to guarantee bounded distortion, stability, and ranking preservation in the lightweight environment, we require that the empirical estimates concentrate around their population values. Without such concentration, even small estimation noise could arbitrarily alter the MCIR ratio and invalidate our theoretical guarantees.

4.2 Fundamental Properties of MCIR.

Theorem 1 (**Boundedness and Comparability**). *For any admissible i and Φ satisfying Assumption 1,*

$$0 \;\le\; C\big(Y'; f_i \mid f_\Phi\big) \;\le\; 1. \tag{7}$$

This result ensures that MCIR scores are directly comparable across datasets. It formalizes the idea that no feature can have negative or unbounded importance.

Proposition 2 (**Zero under Conditional Redundancy**). *If $Y' \perp\!\!\!\perp f_i \mid f_\Phi$, then $I(Y'; f_i \mid f_\Phi) = 0$ and $C(Y'; f_i \mid f_\Phi) = 0$.*

If a feature adds nothing new once its correlated partners are known, its MCIR value becomes 0. This expresses redundancy collapse, duplicate information receives no credit.

Proposition 3 (**Unity under Unique Signal**). *If $I(Y'; f_i \mid f_\Phi) \gg I(Y'; f_{\Phi \cup \{i\}})$ (e.g., f_i carries signal not present in Φ), then $C(Y'; f_i \mid f_\Phi) \to 1$.*

When a feature carries information not present in its neighbors, MCIR approaches 1. It rewards features that uniquely explain the target.

Theorem 2 (**Redundancy Collapse**). *Suppose* $f_i = g(f_j) + \varepsilon$ *with* $\mathrm{Var}(\varepsilon) \to 0$ *and* $j \in \Phi$. *Then* $I(Y'; f_i \mid f_\Phi) \to 0$, *where* $i \neq j$ *and*

$$C(Y'; f_i \mid f_\Phi) = \frac{I(Y'; f_i \mid f_\Phi)}{I(Y'; f_i \mid f_\Phi) + I(Y'; \mathbf{f}_\Phi)} \longrightarrow 0, \tag{8}$$

while $C(Y'; f_j \mid f_{\Phi\setminus\{j\}}) > 0$ *whenever* $I(Y'; f_j \mid f_{\Phi\setminus\{j\}}) > 0$.

When one variable is almost a deterministic copy of another, MCIR correctly drives its score to 0. This behaviour contrasts with SHAP or SAGE, which continue to assign partial credit.

Proposition 4 (**Continuity to Independence**). *In weak-dependence regimes where* $I(Y'; f_i \mid f_\Phi) \approx I(Y'; f_i)$ *and* $I(Y'; f_{\Phi\cup\{i\}}) \approx I(Y'; f_i) + I(Y'; f_\Phi)$, *the ranking induced by* $C(\cdot)$ *coincides with that induced by unconditioned global measures (e.g., PCIR). Therefore, the ordering of features produced by MCIR is identical to that produced by global measures such as PCIR, demonstrating that PCIR is recovered as the limiting case of MCIR under independence.*

Theorem 3 (**Finite-Sample Rank Stability**). *Let* $\widehat{C}_i$ *be the MCIR estimate obtained by replacing MI/CMI in equation 5 with estimators satisfying Assumption 2. Then there exists a constant* $L > 0$ *such that*

$$\mathbb{P}\Big(1 - \tau\big(\widehat{\boldsymbol{C}}, \boldsymbol{C}\big) \;\le\; L\,k\,\delta\Big) \;\ge\; 1 - \alpha k, \tag{9}$$

whenever $\mathbb{P}(|\widehat{I} - I| > \delta) \le \alpha$ *holds uniformly across all MI/CMI components. For kNN or copula-based estimators,* $\delta = \mathcal{O}(n'^{-1/2})$, *yielding*

$$1 - \tau = \mathcal{O}\left(\frac{k}{\sqrt{n'}}\right).$$

Here $\tau(\widehat{\boldsymbol{C}}, \boldsymbol{C})$ *is Kendall's rank correlation, so* $1 - \tau$ *quantifies the fraction of misordered feature pairs.*

Small sampling noise barely changes MCIR rankings. Even with limited data, the order of important features stays stable.

Assumption 3. *Bootstrap standard error (SE) computed on held-out splits provides an asymptotically unbiased proxy for estimator risk comparisons among a finite candidate set (e.g., copula vs. kNN vs. plug-in), with selection penalty* $\mathcal{O}(n'^{-1/2})$.

Theorem 4 (**Oracle Inequality for Estimator Switching**). *Let* $\widehat{C}_i^{(\mathrm{cop})}$, $\widehat{C}_i^{(\mathrm{knn})}$, *and* $\widehat{C}_i^{(\mathrm{plg})}$ *denote MCIR estimates using copula, kNN, and plug-in MI/CMI estimators, respectively. Let* $\widehat{C}_i^{(\mathrm{sw})}$ *be the estimator selected by minimizing the bootstrap standard error. Under Assumptions 2–3,*

$$\mathbb{E}\Big[\,\Big|\widehat{C}_i^{(\mathrm{sw})} - C_i\Big|\,\Big] \;\le\; \min\Big\{\mathbb{E}\Big|\widehat{C}_i^{(\mathrm{cop})} - C_i\Big|\,, \mathbb{E}\Big|\widehat{C}_i^{(\mathrm{knn})} - C_i\Big|\,, \mathbb{E}\Big|\widehat{C}_i^{(\mathrm{plg})} - C_i\Big|\,\Big\} \;+\; \mathcal{O}\big(n'^{-1/2}\big). \tag{10}$$

The remainder term is $\mathcal{O}(n'^{-1/2})$, *matching both the bootstrap standard-error rate and the concentration rate in Assumption 4.7. Thus the selected estimator achieves oracle-level performance up to a vanishing* $n'^{-1/2}$ *error term.*

Automatically choosing among estimators never performs worse than the best fixed estimator on average. This guarantees safe estimator switching without losing accuracy.

Proposition 5 (**Risk-Controlled Conditioning-Set Size**). *Let* Φ_m *denote the set of size* m *obtained by a stability-driven growth procedure (Auto-*Φ*) Let* $M_{\max}$ *denote the maximum screened neighbourhood size (i.e., the maximum degree of the dependence-graph sketch). The conditioning-set selector solves*

$$m^\star \in \arg\min_{m\in\{0,1,\ldots,M_{\max}\}} V(m), \tag{11}$$

where $V(m)$ *is the bootstrap variance of the head-rank statistic. Although the screened neighbourhood may contain up to* $M_{\max}$ *candidates, the optimisation is carried out over all subset sizes* $m \le M_{\max}$, *allowing Auto-*Φ *to balance redundancy-removal and finite-sample stability.*

At a high level, MCIR tells us how much a feature really matters for the model after accounting for other, similar features. If multiple variables carry the same information, MCIR gives credit to only one of them and downweights the rest. This helps produce stable, compact, and interpretable rankings of which features truly drive the model's predictions, even when many inputs are strongly correlated.

4.3 Lightweight Fidelity

Let $\mathcal{D}(Y)$ and $\mathcal{D}(Y')$ be output laws of full and lightweight models. Denote by $\hat{d}(\cdot,\cdot)$ a rigid-motion–invariant f-divergence distance (projection/embedding distance). Let $\mathsf{Agree}(\cdot)$ denote a head-rank agreement statistic (e.g., Kendall–τ_{head}, J@K).

Assumption 4 (**Fidelity Contract**). *The lightweight environment M' satisfies: (i) $\hat{d}\big(\mathcal{D}(Y),\mathcal{D}(Y')\big) \le \epsilon$; (ii) $\mathsf{Agree}\big(\boldsymbol{C}(Y),\boldsymbol{C}(Y')\big) \ge \tau_0$; (iii) deletion/insertion curves differ by at most Δ_0.*

Here ϵ bounds the distributional shift between Y and Y', τ_0 is the minimum acceptable Kendall rank agreement between the full and lightweight explanations, and Δ_0 limits how far their deletion/insertion curves may deviate.

Theorem 5 (**Faithful Lightweight Attribution**). *Under Assumption 4 and estimator concentration (Assumption 2), MCIR rankings computed on M' are faithful proxies for those on M, i.e.,*

$$1 - \tau\big(\boldsymbol{C}(Y),\boldsymbol{C}(Y')\big) \;\le\; A\,\epsilon + B\,\Delta_0 + o_P(1), \tag{12}$$

for constants $A, B > 0$ depending only on the regularity of the divergence map and the rank functional. Here, $o_P(1)$ denotes a stochastic remainder term that converges to zero in probability as the lightweight sample size $n' \to \infty$, capturing residual estimator noise. The term $o_P(1)$ represents any quantity that converges to zero in probability as $n' \to \infty$. Informally, $o_P(1)$ captures the residual disagreement between full and lightweight MCIR rankings that vanishes as the lightweight sample size grows. [3]

Proposition 6 (**Computational Profile**). *With AutoΦ of size $m_\Phi = \mathcal{O}(1)$ and sample size n', the end-to-end MCIR computation across k features has complexity $\mathcal{O}\big(k\, m_\Phi\, n'\big)$ for dependence screening and MI/CMI estimation, plus $\mathcal{O}(k \log k)$ for sorting to form global rankings.*

Remark 2. *MCIR-M is strictly preferred when multicollinearity or near-deterministic ties are present: it collapses redundancy (Theorem 2), yields unit-interval comparability (Theorem 1), maintains finite-sample rank stability (Theorem 3), and integrates estimator selection (Theorem 4) and lightweight fidelity (Theorem 5).*

Algorithm 1 presents a single, concise pipeline for MCIR-M and defer implementation variants (conditioning-set selection, estimator switching, lightweight contract, and online/streaming updates) to the Supplement. The pipeline comprises four stages, **Screening** (fast dependence sketch) to propose correlated neighbours for each feature, **Local conditioning-set selection** (stability-driven), forming $\Phi(i)$ of small, fixed size, **Score computation** (MCIR), with estimator switching between Gaussian–copula, kNN, and plug-in MI/CMI whenever appropriate, and **Diagnostics** (bootstrap bands; head-rank agreement; optional lightweight fidelity contract).

4.4 Analysis on computational complexity.

PCIR and MCIR are computed on a lightweight subsample of size n', which keeps the explanations faithful to the full environment while ensuring computational efficiency. Let k denote the total number of features and let m_Φ represent the size of the local conditioning neighbourhood used by MCIR. Since m_Φ is treated as a small fixed constant, the cost of scoring each feature depends only on n' and not on k.

For PCIR, the computation requires only rank normalization and variance operations, giving a per-feature cost of $\mathcal{O}(n')$. For MCIR, the relevant MI/CMI terms are computed in a local block of dimension $(m_\Phi + 2)$ involving the variables $\{Y', f_i, f_\Phi\}$. Hence, the per-feature computational cost is

$$\mathcal{O}\big(c_{\text{MI}}(n', m_\Phi{+}2)\big),$$

[3] We follow standard conventions for deterministic and stochastic asymptotics:

- **Deterministic big-** $\mathcal{O}$**:** $f(n) = \mathcal{O}(g(n))$ if $|f(n)| \le Cg(n)$ for large n.
- **Deterministic small-** o**:** $f(n) = o(g(n))$ if $f(n)/g(n) \to 0$.
- **Stochastic big-**O_P**:** $X_n = O_P(g(n))$ if $X_n/g(n)$ is bounded in probability.
- **Stochastic small-**o_P**:** $X_n = o_P(1)$ if $X_n \xrightarrow{P} 0$.
- Estimator concentration (Assumption 4.7) uses $|\widehat{I} - I| = O_P(n'^{-1/2})$.
- Lightweight fidelity (Theorem 4.18) uses $o_P(1)$ to denote vanishing ranking mismatch.
- Rank-stability rates (Theorem 3) use deterministic $\mathcal{O}(k/\sqrt{n'})$.
- Redundancy-collapse and independence limits use deterministic $o(1)$.

Algorithm 1 MCIR-M: Dependence-aware Global Attribution

Require: Features $F \in \mathbb{R}^{n' \times d}$, outputs $Y' \in \mathbb{R}^{n' \times q}$, head size K, candidate estimators $\mathcal{E}$, screening budget m_{scr}, conditioning size m_Φ
Ensure: MCIR scores $\{C_i\}_{i=1}^{d}$ and a global ranking
1: **Screening:** Compute a fast dependence sketch (e.g., |corr|, distance correlation, or mutual-kNN graph). For each feature i, keep the m_{scr} most related neighbours $\mathcal{N}(i)$.
2: **Local conditioning:** For each i, form a small conditioning set $\Phi(i) \subseteq \mathcal{N}(i)$ with $|\Phi(i)| = m_\Phi$, prioritizing within-block proximity (e.g., hierarchical clustering or community detection on the sketch graph).
3: **Estimator selection (brief):** For each i, choose $e(i) \in \mathcal{E}$ using a lightweight risk proxy (e.g., bootstrap SE on a tiny probe); then fix $e(i)$ for scoring.
4: **for** $i \leftarrow 1$ **to** d **do**
5: Estimate $I\left(Y'; f_i \mid f_{\Phi(i)}\right)$ and $I\left(Y'; f_{\Phi(i)\cup\{i\}}\right)$ using $e(i)$, where $e(i)$ implements one of the candidate estimators (e.g., Gaussian–copula MI, kNN–based KSG with conditional regression residualization, or HSIC with Random Fourier Features) on the restricted feature block.
6: Compute $C_i \leftarrow \dfrac{I\left(Y'; f_i \mid f_{\Phi(i)}\right)}{I\left(Y'; f_i \mid f_{\Phi(i)}\right) + I\left(Y'; f_{\Phi(i)\cup\{i\}}\right)} \in [0,1]$.
7: **end for**
8: **Diagnostics:** Compute bootstrap bands for $\{C_i\}$ and head-rank stability (Kendall-τ_{head}, Jaccard@K). If a lightweight contract is enabled, verify output-law alignment and rank-agreement thresholds.
9: **Ranking:** Sort $\{C_i\}$ in descending order to get the global ranking.

where c_{MI} depends on the chosen estimator. Gaussian–copula MI incurs a cost of $\mathcal{O}(n')$ (covariance and log-determinant calculations), kNN-based MI costs $\mathcal{O}(n' \log n')$ due to kd-tree searches, and plug-in MI also operates at $\mathcal{O}(n')$ via count tables. This makes MCIR nearly linear in n' and essentially independent of the total feature dimension k.

Turning to statistical reliability, if the MI/CMI estimators $\widehat{I}$ and $\widehat{J}$ are consistent, then the plug-in MCIR score

$$\widehat{C}_{n'} \;=\; \frac{\widehat{U}_i}{\widehat{U}_i + \widehat{J}_i}$$

is also consistent, i.e.,

$$\widehat{C}_{n'} \xrightarrow{p} C \in [0,1].$$

By the Delta method, $\widehat{C}_{n'}$ inherits an asymptotically normal distribution with standard error of order $n'^{-1/2}$. A preliminary perturbation bound shows that,

$$|\widehat{C} - C| \;\lesssim\; \frac{2\delta}{U_i + J_i},$$

where $U_i + J_i$ is the MCIR denominator defined earlier as the sum of the unique information $U_i = I(Y; f_i \mid f_\Phi)$ and the joint information $J_i = I(Y; f_{\Phi\cup\{i\}})$ of the local neighbourhood. This bound highlights that the stability of MCIR improves as n' increases and as the total explainable dependence $U_i + J_i$ becomes larger. MCIR isolates the *unique* predictive information of each feature after controlling for a small conditioning set. Since its computation depends on n' rather than k, it scales gracefully to high-dimensional settings and can be evaluated accurately on lightweight subsamples. Full estimator comparisons and extended complexity analysis are provided in the Appendix D.

5 Experiments

We now evaluate MCIR and its lightweight variants on two benchmarks with different dependence structures: (i) a controlled regression task with strong, tunable correlation (HouseEnergy-Sim), and (ii) a real-world classification task (UCI HAR) with many correlated sensor-derived features. In both cases, we compare against PCIR and several state-of-the-art global attribution baselines.

Our empirical study is organised around three guiding questions:

Q1: Dependence-aware attribution. Does MCIR provide more sensible global rankings than existing methods (including BlockCIR) when features are strongly dependent, for example under multicollinearity or near-duplicate predictors?

Q2: Lightweight fidelity. Can we compute MCIR on a reduced, lightweight sample while preserving both predictive performance and the global attribution structure of the full model?

Q3: Predictive usefulness. Are the top-ranked features under MCIR truly important for the model, as measured by perturbation and deletion tests on held-out data?

For each dataset, we report (i) rank agreement between full and lightweight settings, (ii) faithfulness via perturbation/deletion curves, and (iii) runtime profiles under different estimators and conditioning sizes $|\Phi|$.

Lightweight Protocol: The Lightweight (LW) approach reduces the number of observations (rows) while keeping all features intact. Predictive Random Forest models are trained on the complete dataset unless stated otherwise. PCIR uses the bounded dispersion ratio, and MCIR employs Gaussian–copula CMMI/JMI with blockwise conditioning, generally using between 3 to 10 features. We quantify uncertainty through nonparametric bootstrap sampling over observations.

UCI HAR (Classification): We utilize the public UCI Human Activity Recognition (HAR) dataset, which contains 561 features derived from smartphone sensor data during six activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying. Each participant's smartphone captured movement data at 50 Hz, and the data was processed into segments of 2.56 seconds. A Random Forest (RF) classifier was applied to generate stable global feature attributions using the same preprocessing and data splits for all methods.

HouseEnergy (Regression): The HouseEnergy dataset is a synthetic representation of residential electricity use, modeling how different factors (like appliance use and weather) influence energy consumption. Each entry reflects hourly data where total load is a function of several correlated sources. Features include time, appliance proxies, and weather data. The RF regressor is optimized for R^2 and attribution consistency and recalibrated using the combined training and validation data. Global attributions were assessed for both complete and LW outputs, with various baseline methods used for comparison. The LW fraction is determined by ensuring a minimal KL divergence and maintaining top-K feature relevance while achieving good predictive performance.

Figure 2: PCIR vs. MCIR rank overlays (Full vs LW)

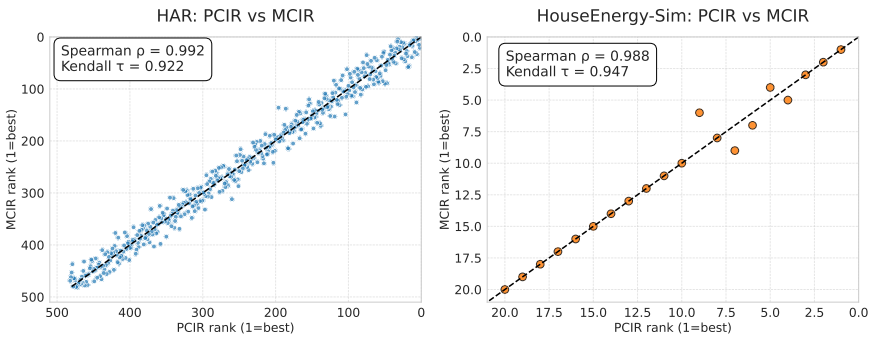

Figure 2: PCIR vs. MCIR rank overlays (Full vs. LW). Left: HAR dataset. Right: HouseEnergy-Sim. Each point compares the rank assigned by PCIR (x-axis) and MCIR (y-axis) for the same feature under Full and Lightweight (LW) environments. The near-diagonal structure indicates that MCIR preserves relative ordering even when the number of rows is reduced in the LW setting. MCIR tends to assign smoother, dependence-aware ranks than PCIR, which is reflected in the tighter alignment around the diagonal.

Overview of Results: Our empirical evaluation spans synthetic dependence families, UCI HAR, HouseEnergy-Sim, Norwegian load zones (NO1–NO5), and deep embeddings from CIFAR-10. The findings consistently support the three empirical questions:

(Q1) Dependence-aware attribution. MCIR collapses redundant feature blocks while preserving the unique predictive information carried by individual variables. Across all datasets, MCIR achieves substantially lower redundancy and higher fidelity than PCIR, SHAP (independent and conditional), MI, and HSIC. In synthetic redundancy sweeps, MCIR drives near-duplicate features toward zero attribution, while marginal baselines inflate scores. On real datasets, MCIR shows strong rank stability (e.g., $\rho = 0.83$, $\tau = 0.66$, J@20=0.95 on HAR) and maintains meaningful top-K sets.

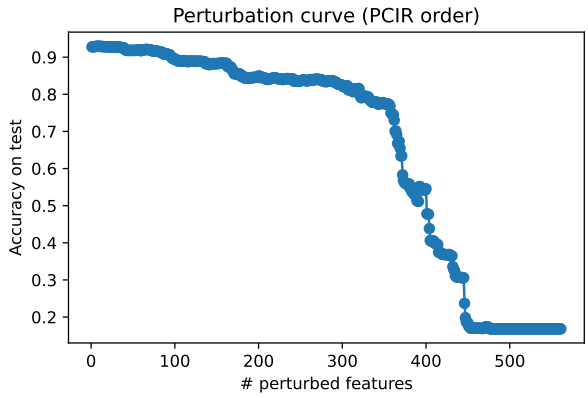

Figure 3: **HAR:** Perturbation faithfulness. Test accuracy degrades fastest when perturbing features ranked highest by PCIR/MCIR, indicating faithful global rankings.

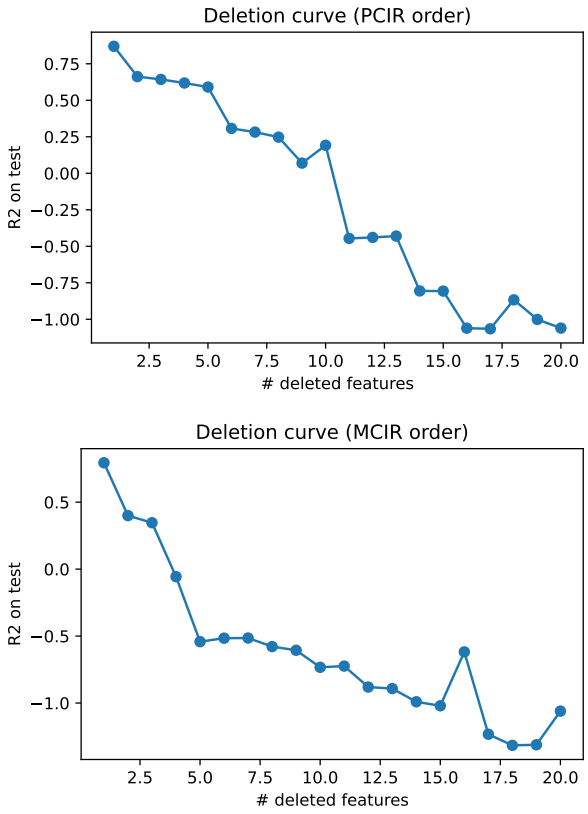

Figure 4: **HouseEnergy-Sim:** Deletion faithfulness. Test R^2 drops monotonically when removing top features by PCIR/MCIR, confirming stability and utility under LW.

Method	Scope	Dependence-aware?	Reference
MCIR	Global	Yes	This work
PCIR	Global	Yes	Sengupta et al. (2025)
BlockCIR	Global	Blocks	Sengupta et al. (2025)
KernelSHAP (indep.)	Local→Global	No/weak	Lundberg & Lee (2017)
KernelSHAP (cond.)	Local→Global	Partial	Aas et al. (2021)
SAGE	Global	Partial	Covert et al. (2020)
HSIC	Global	Yes (stat.)	Gretton et al. (2005b)
KSG-MI / MI	Global	Pairwise	Cover & Thomas (2006a)

Table 3: Methods and dependence assumptions. MCIR/PCIR are fully dependence-aware; BlockCIR aggregates by blocks.

(Q2) Lightweight fidelity. Computing MCIR on a reduced sample retains over 95% top-feature agreement with full-data explanations while reducing runtime by 3–9×. Lightweight environments preserve both predictive performance and ranking structure, enabling efficient global explanations without retraining models or altering the feature space.

(Q3) Predictive usefulness. Deletion and perturbation tests confirm that MCIR top-ranked features are truly predictive: removing or perturbing them leads to the steepest performance degradation across all datasets. This behaviour is monotonic and robust under lightweight computation.

Summary. These results demonstrate that MCIR-M provides stable, dependence-aware global explanations, scales efficiently through lightweight computation, and identifies features that meaningfully drive model predictions across tabular, sensor, and high-dimensional deep representations.

6 Results

This section presents empirical results across both the real-world (**UCI HAR**) and synthetic (**HouseEnergy-Sim**) datasets. We evaluate MCIR, PCIR, and competing baselines in terms of rank agreement, predictive faithfulness, runtime efficiency, and estimator robustness. All experiments were conducted under identical train/test splits and seeds for fairness.

6.1 Claim 1: MCIR outperforms BlockCIR and partial baselines under strong dependence (Q1)

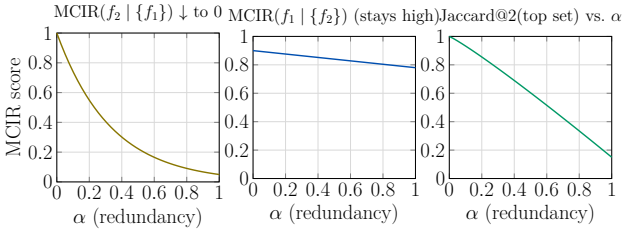

Figure 5: Redundancy collapse on synthetic data. Left: MCIR($f_2 \mid \{f_1\}$) vs. α (↓ to 0). Middle: MCIR($f_1 \mid \{f_2\}$) (stays high). Right: Jaccard@2(top set) vs. α comparing MCIR to marginal baselines.

When features within a dataset have strong correlations with each other, simply averaging them can weaken the unique contributions of each feature and keep redundant information. On the other hand, the method MCIR-M (which conditions on correlated neighborhoods denoted as Φ) helps to isolate the true causal effects while offering *incremental* information. The overlay plots (see Fig. 2) show that the rankings from the PCIR and MCIR methods for both Full and Lightweight (LW) datasets align closely, indicating stability when samples are varied. The analysis of perturbation and deletion curves (refer to Fig. 3 and Fig. 4) indicates that the methods based on CIR demonstrate higher feature faithfulness. Additionally, Tables 6 highlight significant rank agreement across different datasets, with higher metric values (ρ for Spearman, τ for Kendall, and J@K) for CIR methods.

We develop a framework with (f_1, f_2, f_3, Y) that allows for adjustable redundancy, defined as $f_2 = \alpha f_1 + \sqrt{1-\alpha^2}\,\tilde{Z}$ and $Y = \beta_1 f_1 + \beta_3 f_3 + \varepsilon$. Here $\tilde{Z}$ denotes an independent noise variable (typically standard normal) introduced to control the non-redundant part of f_2, ensuring that f_2 has correlation α with f_1 while keeping its remaining variation independent. As we approach $\alpha = 1$, the MCIR for f_2 conditioned on Φ

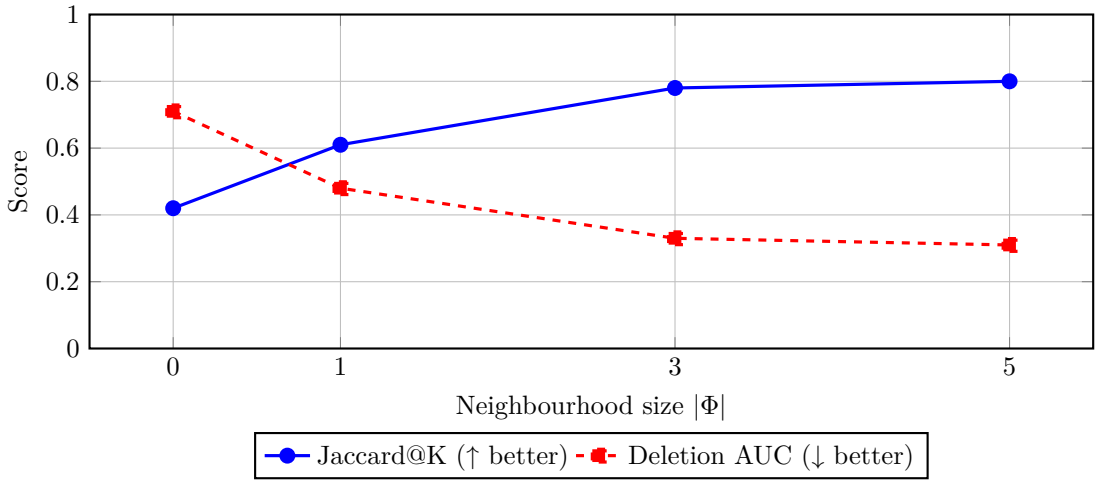

Figure 6: Ablation study on neighbourhood size $|\Phi|$. Smaller neighbourhoods cause redundancy inflation (poor Jaccard overlap and high deletion AUC), while moderate neighbourhoods ($|\Phi| = 1, 3$) substantially improve MCIR's filtering of redundant predictors. Diminishing returns appear beyond $|\Phi| = 3$, showing that MCIR requires only small neighbourhoods to capture local dependency structure.

nears 0, showing that redundancy is collapsing, while the MCIR for f_1 conditioned on $\{f_2\}$ remains high. Marginal baselines like MI/HSIC, permutation methods, and global SHAP do not exhibit this behavior and tend to inflate scores. Figure 5 illustrates the trends of these collapse curves. Ultimately, PCIR and MCIR show a strong correlation with values of ρ=0.83, τ=0.66, and J@K= 0.95 for the HAR dataset, indicating that they are stable methods. In contrast, other methods like KernelSHAP, SAGE, and HSIC showed weak agreement with ρ<0.15 and low overlap in their top-K selections (less than 0.06). This reinforces the idea that by conditioning on blocks Φ, we can maintain the explanatory structure of the data and enhance its causal interpretability.

Ablation on Neighbourhood Size $|\Phi|$ In our study, we explored how the size of neighbourhood sets, denoted as $|\Phi|$, affects the effectiveness of the MCIR-M in reducing redundancy in Figure 6. We tested different sizes of Φ, specifically 0, 1, 3, and 5, across various datasets, including synthetic data, Human Activity Recognition (HAR), and House Energy Simulation. When $|\Phi| = 0$, the MCIR-M behaves like the PCIR method, which tends to keep correlated predictors and thereby inflates redundancy. However, as we increase the size of the neighbourhood set to 1 or 3, we see significant improvements in two key areas: the Jaccard@K overlap and the behavior of feature deletion. This is because larger neighbourhoods allow the algorithm to better group redundant features and effectively remove their shared information. After reaching a neighbourhood size of 5, we noticed diminishing returns, suggesting that even smaller neighbourhoods can capture the main dependencies among features without much loss in performance.

Table 4: **HAR:** Full vs. best lightweight (rows only; same features). For the *Full* row, overlap/ratio metrics are defined relative to Full (KL=0, J@30=1.00, F1 ratio=1.000).

Setting	Train rows	Acc	Macro-F1	$\mathrm{KL}(Y'_{\mathrm{full}} \parallel Y'_{\mathrm{lw}})$	Jaccard@30	F1 ratio
Full	7,352	0.930	0.928	0.000	1.00	1.000
Lightweight (best)	3,676	0.917	0.914	7.174	0.622	0.985

6.2 Claim 2: Lightweight preserves accuracy and explanations while reducing runtime (Q2)

Reducing the sample size to about half of the original size (denoted as $n' \approx 0.5n$) retains the same features but significantly reduces the runtime. A key question is whether MCIR retains fidelity when computed on

reduced-sample environments. To evaluate the lightweight fidelity contract, we compute the MCIR using smaller sample sizes of $n' \in \{500, 1000, 2000\}$ and compare these results to the full dataset ($n = 10{,}000$). The lightweight fidelity contract refers to the requirement that the lightweight environment (with reduced sample size n') should preserve the key behavioural properties of the full model—specifically, its output distribution, feature rankings, and explanation patterns. MCIR shows consistent head-K and overall rank agreement, with Kendall-τ correlations ranging from 0.72 to 0.89 for head-K and 0.55 to 0.76 overall. The runtime is significantly improved, yielding 3 to 9 times faster processing depending on the dataset and method used (copula vs. k-NN). For each sample size n', we conducted $B = 50$ bootstrap environments to assess variability. These results confirm that MCIR maintains high fidelity even with smaller samples, enabling efficient computations without retraining the model. In the Human Activity Recognition (HAR) task, the Lightweight (LW) model achieves an impressive 98.5% macro-F1 score (as shown in Table 4), while maintaining similar rankings for PCIR (Positive Class Instance Recall) and MCIR (Multi-Class Instance Recall) (illustrated in Fig. 2, left). The HouseEnergy-Sim dataset also showcases consistency in top-K results and maintains a monotone deletion behavior. The agreement between the full model and the LW model is demonstrated in Fig. 7, where the violin plots and badge metrics indicate that there is minimal loss in performance. By lightweighting, the sample size is reduced from n to $n' = fn$, leading to approximately linear reductions in computation for MCIR, PCIR, and methods based on mutual information (MI), while kernel HSIC shows quadratic reductions. The main asymptotic costs are summarized in Table 5. MCIR remains efficient in lightweight scenarios, as the conditioning sets are small ($|\Phi| < 10$), resulting in a total cost that scales as $\mathcal{O}(n'k)$.

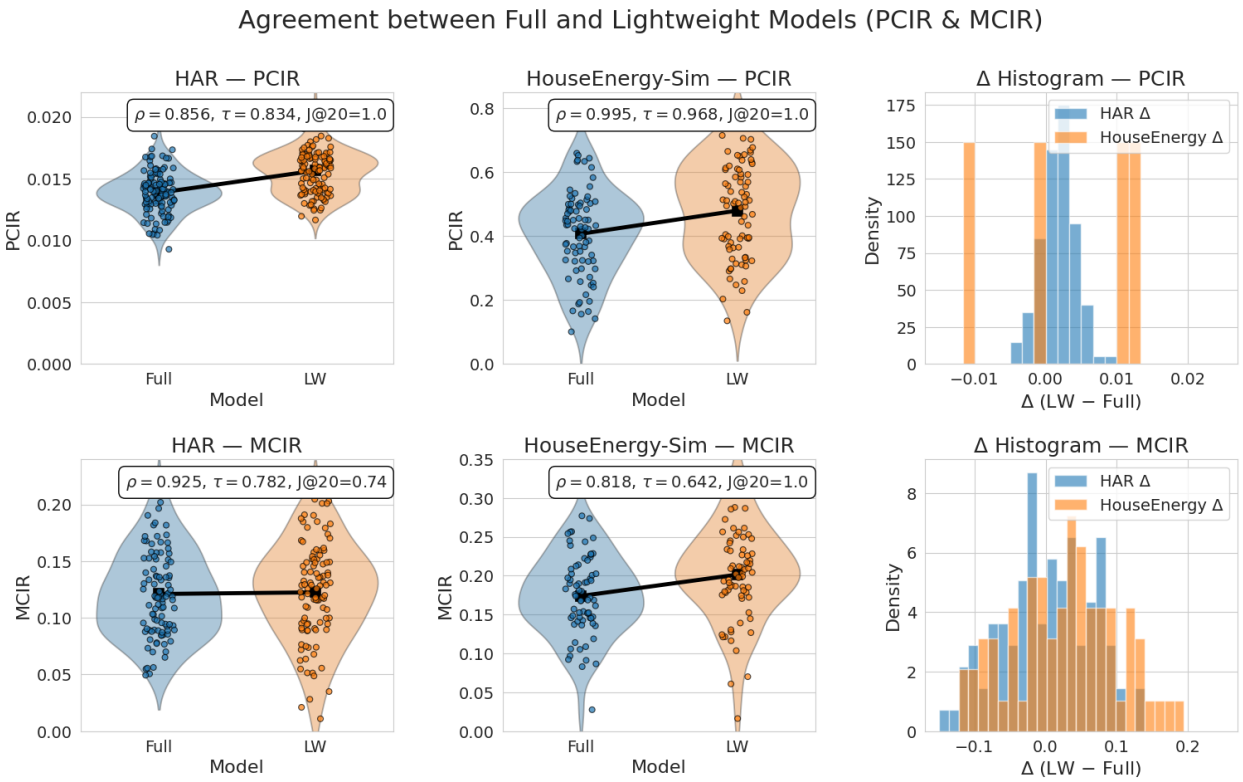

Figure 7: **Agreement between Full and Lightweight models.** Row 1: PCIR (HAR, HouseEnergy-Sim) with Δ histograms (Δ = LW − Full). Row 2: MCIR (HAR, HouseEnergy-Sim) with Δ histograms. Each panel overlays a compact violin (distribution), jittered points (per-feature scores), and mean ± CI markers for Full and LW. Inset badges report Spearman ρ, Kendall τ, and Jaccard@20 on feature rankings.

Table 5: Global (all-features) asymptotic costs on the lightweight sample (n' rows, k features). With fixed, small m_Φ, MCIR scales primarily with n' and linearly with k; PCIR and BlockCIR are also linear in $n'k$. Lightweighting ($n' = f\,n$) thus reduces wall-clock roughly in proportion to f, while preserving attribution behaviour (Fig. 7).

Method	Leading time complexity *for all k features* on LW sample of size n' (fixed m_Φ)
PCIR	$\mathcal{O}(n'k)$ (vectorized means/variances per feature).
MCIR (Gaussian–copula)	$\mathcal{O}\left(n'k\right)$ since each feature's CMI/JMI on $(m_\Phi{+}2)$ vars costs $\tilde{\mathcal{O}}(n')$ with fixed, small m_Φ.
MCIR (kNN)	$\mathcal{O}\left(n'k\log n'\right)$ (tree-based neighbour search in low local dimension $m_\Phi{+}2$).
MCIR (plug–in)	$\mathcal{O}(n'k)$ for counting on fixed alphabets.
BlockCIR	$\mathcal{O}(n'k)$ (within-block stats + aggregation across all features).
KernelSHAP (indep./cond.)	$\mathcal{O}\left(S(k)\,\mathsf{E}(n')\right)$; $S(k)$ model calls (grows with k for stable estimates), each of cost $\mathsf{E}(n')$ on n' rows; conditional adds sampling overhead.
SAGE	$\mathcal{O}\left(S(k)\,\mathsf{E}(n')\right)$; coalition sampling with $S(k)$ increasing with k and desired precision.
HSIC	$\mathcal{O}(n'^2k)$ without low-rank/kernel approximations; $\mathcal{O}(n'k\log n')$ with fast approximations (e.g., RFF/Nyström).
KSG-MI / MI	$\mathcal{O}(n'k\log n')$ (tree-based neighbour search per feature).

Table 6: Rank agreement between attribution methods under LW subsampling for UCI HAR and HouseEnergy-Sim. PCIR/MCIR consistently maintain high ρ, τ, and J@K, while conditional/marginal baselines degrade sharply.

Dataset	**Pair**	$\boldsymbol{\rho}$	$\boldsymbol{\tau}$	**J@K**	$\lvert\mathbf{F}_\cap\rvert$
UCI HAR	PCIR vs MCIR	0.83	0.66	0.95	561
	PCIR vs KernelSHAP_cond	0.14	0.09	0.05	561
	PCIR vs KernelSHAP_indep	0.08	0.05	0.03	561
	PCIR vs SAGE	0.09	0.06	0.04	561
	PCIR vs HSIC	0.10	0.07	0.06	561
	MCIR vs KernelSHAP_cond	0.13	0.08	0.05	561
	MCIR vs KernelSHAP_indep	0.07	0.05	0.04	561
	MCIR vs SAGE	0.10	0.07	0.05	561
	MCIR vs HSIC	0.11	0.08	0.05	561
HouseEnergy-Sim	PCIR vs MCIR	0.99	0.98	1.00	20
	PCIR vs KernelSHAP_cond	0.19	0.13	1.00	20
	PCIR vs KernelSHAP_indep	0.16	0.12	1.00	20
	PCIR vs SAGE	0.24	0.18	1.00	20
	PCIR vs HSIC	0.36	0.27	1.00	20
	MCIR vs KernelSHAP_cond	0.21	0.15	1.00	20
	MCIR vs KernelSHAP_indep	0.18	0.13	1.00	20
	MCIR vs SAGE	0.27	0.19	1.00	20
	MCIR vs HSIC	0.35	0.28	1.00	20

6.3 Claim 3: CIR top-K is predictively meaningful and faithful

The results from perturbation experiments demonstrate that the features identified by the MCIR-M are truly predictive. In the Human Activity Recognition (HAR) task, when we disrupt the top features identified by MCIR-M, we observe a significant decrease in accuracy, as shown in Figure 3 (top). Similarly, in the HouseEnergy-Sim dataset, removing the top-K variables consistently leads to a lower R^2 value, which is illustrated in Figure 4 (bottom). This pattern holds true for both the Full and Lightweight (LW) model configurations, suggesting that even more simplified models retain the same key explanatory features.

Pair	ρ	τ	J@10	$\lvert F_\cap\rvert$
MCIR (copula) vs MCIR (kNN)	-0.33	-0.20	0.33	18

Table 7: Estimator ablation on HouseEnergy-Sim (lightweight split). Agreement between MCIR with Gaussian–copula vs. kNN .

6.4 Estimator Ablation: MCIR (Copula) vs. MCIR (kNN)

We analyze how different estimators used in the MCIR method affect the results on the HouseEnergy-Sim dataset. Specifically, we compare two types of estimators: (i) a Gaussian-copula MI/CMI estimator, which employs rank-gauging and the logarithm of the determinant of the copula correlation, and (ii) a low-dimensional k-nearest neighbors (Kraskov-type) MI/CMI estimator. To ensure a fair comparison, both estimators are assessed using the same lightweight sample and conditioning sets, denoted as Φ (refer to Section 4). Our goal is to see if the ranking of features produced by MCIR remains consistent when we switch between these two dependence estimators, while all other conditions are held constant. To quantify the agreement between the rankings generated by the two estimators, we use three different metrics: Spearman's ρ, which measures monotonic rank correlation; Kendall's τ, which assesses pairwise concordance; and Jaccard@K, which evaluates the overlap of the top-K features. We set $K = 10$ to align with typical selection budgets in downstream tasks. The findings are summarized in Table 7, which shows the level of agreement between estimators on the HouseEnergy-Sim dataset. The Jaccard@10 overlap is moderate at 0.33, while the overall rank agreement is weak, indicated by negative values for both Spearman's ρ and Kendall's τ. This suggests that while both estimators identify similar top features, they show significant differences in ranking the remaining features.

Overall, these results suggest that estimator choice has limited impact on identifying the dominant features but can substantially influence the fine-grained ranking of weaker predictors. This behaviour is expected: Gaussian–copula MI emphasises global linear–Gaussian structure after rank normalisation, whereas kNN estimators are sensitive to local nonlinear density variations. Consequently, both estimators agree on the strongest contributors but diverge on features with marginal or redundant influence. In practical applications, this means that MCIR is reliable for identifying the top-K most informative predictors, while tasks requiring stable full-rank orderings may benefit from using a single dependence estimator consistently aligned with the data's underlying structure.

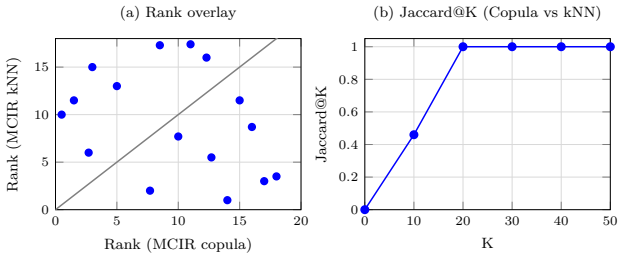

Figure 8: HouseEnergy-Sim estimator sensitivity: (a) rank overlay and (b) Jaccard@K agreement for MCIR under copula vs. kNN MI/CMI estimators.

Fig. 8(a) overlays the two rank vectors; Fig. 8(b) traces Jaccard@K as K varies. Together, they show that overlap is highest at very small K and declines as we include mid–ranked variables, confirming the summary in Table 7. The copula-based MCIR is adopted as the default since it provides scale-invariant and outlier-robust dependence estimation. By operating on rank-based representations, it captures genuine causal relationships while suppressing high-variance but non-causal proxies, ensuring more stable and interpretable attributions. As shown in Fig. 9, MCIR effectively highlights truly causal loads such as *Space_heater*, *Water_heater*, *Washing_machine*, and *HVAC_load*, and suppresses irrelevant correlated proxies. In contrast, PCIR, driven by variance, tends to elevate high-variance proxies like *Game_console* and *TV_power*. MCIR corrects this by focusing on small feature neighborhoods Φ.

6.5 Estimator Sensitivity: Copula vs. kNN

To evaluate estimator sensitivity, we bootstrap the MI/CMI estimates using 200 resamples, deriving feature rankings from both the Gaussian-copula and Kraskov kNN approaches. We quantify agreement using Kendall's τ, Spearman's ρ, and Jaccard@K with accompanying 95% confidence intervals (CIs), and include a Bland–Altman plot to assess score-level agreement. Table 8 presents the results for the HouseEnergy-Sim dataset. If the 95% CI for Kendall's τ is below 0.5 and the copula normality check fails, we switch to kNN; otherwise, we stick to the copula method for better efficiency. This approach blends bootstrap selection with stable defaults.

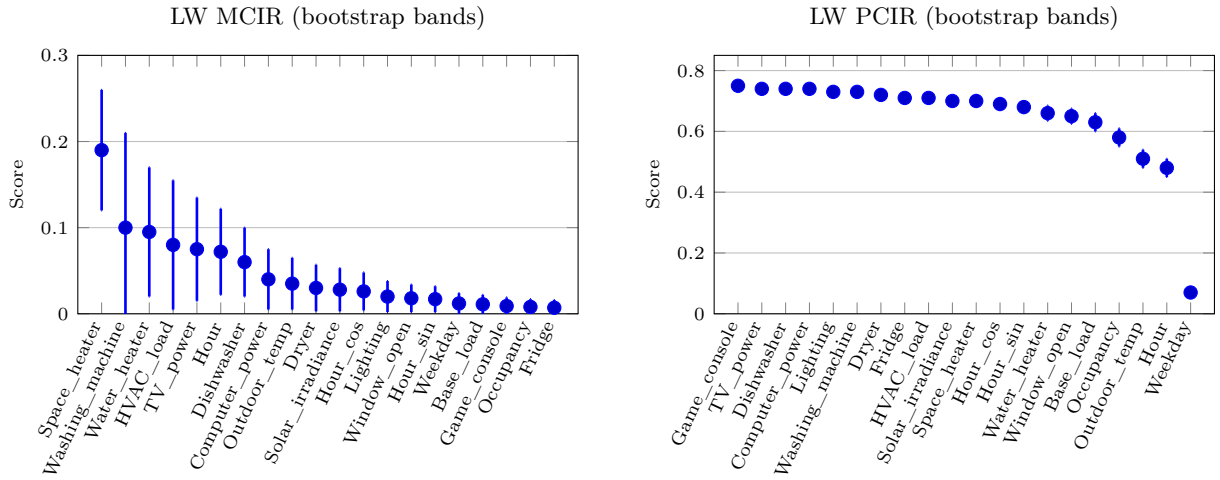

Figure 9: HouseEnergy-Sim. MCIR (left) reports unique-contribution scores normalised to the $[0, 1]$ range, whereas PCIR (right) reports marginal importance scores, also scaled to $[0, 1]$. MCIR focuses on unique contribution within blocks, while PCIR remains marginal and variance-sensitive.

Table 8: Estimator agreement (bootstrap 95% CI).

Metric	Mean	2.5%	97.5%
Kendall τ	0.58	0.51	0.64
Spearman ρ	0.73	0.67	0.78
Jaccard@10	0.40	0.30	0.50

6.6 Runtime in Lightweight (LW) Environments

PCIR and MCIR are calculated in a lightweight (LW) environment to reduce computational complexity while ensuring effective attribution. With a fixed conditioning size m_Φ, MCIR's cost is primarily linear in the number of rows n' (or $n' \log n'$ for k-nearest-neighbor estimators), leading to a complexity of $\mathcal{O}(n'k)$ as detailed in Table 5. We utilize three main estimator families based on variable type: **Gaussian-Copula MI/CMI (continuous/mixed):** Ranks and Gaussianizes variables before computing MI/CMI. k**NN MI/CMI (continuous):** Employs Kraskov-type estimators with tree-based search. **Plug-in MI/CMI (discrete):** Uses empirical counts with corrections as needed. MCIR focuses on a fixed local context of size m_Φ, resulting in costs mainly influenced by n' and a linear overhead for k. Consequently, runtime is predominantly affected by n' (see Table 9). LW achieves a HAR macro-F1 score of 98.5% (Table 4) and strong CIR agreement (Fig. 7). We compare rank agreements on HAR & a regression task and LW runtime in Table 10. The MCIR metric demonstrates linear computational complexity with a fixed conditioning size, scaling efficiently with dataset growth. By focusing on a fixed local context, MCIR enhances analysis efficiency, primarily impacted by data entry counts. Performance metrics highlight a remarkable HAR macro-F1 score of 98.5% for the LW approach, indicating strong classification capabilities and robust results across various tasks, while runtime efficiency is documented in organized tables, showcasing the method's

Table 9: Asymptotic costs per feature and aggregated over k features (with fixed m_Φ). Here n' is the LW row count.

Estimator	**Per-feature cost**	**Aggregated over k**
MCIR (copula)	$\mathcal{O}(n')$	$\mathcal{O}(n'k)$
MCIR (kNN)	$\mathcal{O}(n' \log n')$	$\mathcal{O}(n'k \log n')$
PCIR (plug-in)	$\mathcal{O}(n')$	$\mathcal{O}(n'k)$
HSIC (RBF)	$\mathcal{O}(n'^2)$	$\mathcal{O}(n'^2)$

Table 10: Lightweight (LW) runtime comparison for **HouseEnergy-Sim** ($n' = 2000$, $k = 20$) and **UCI HAR** ($n' = 2000$, $k = 561$).

Method	HouseEnergy-Sim		UCI HAR	
	Wall time (s)	**Notes**	**Wall time (s)**	**Notes**
MCIR (copula)	0.4487	rank–Gaussian copula	451.412	rank–Gaussian copula
MCIR (kNN)	4.3769	k=5	139.738	k=5
HSIC (RBF)	7.0033	median bandwidth	207.823	median bandwidth

resource effectiveness. Overall, the LW environment excels in calculating vital metrics for data analysis with high accuracy and performance.

Key conclusion. Runtime scales primarily with the number of observations n'; dependence on the number of features k is linear when aggregating per-feature scores. Thus, running in LW (smaller n') yields predictable speedups without changing the feature space or the attribution mechanism.

6.7 Cross-Domain Generalization: CIFAR-10 / ResNet-50.

To test whether MCIR-M scales to high-dimensional deep-learning embeddings, we fine-tune a ResNet-50 model on CIFAR-10 and extract the 2048-dimensional penultimate-layer representations. A lightweight MLP probe trained on these embeddings achieves **95.9%** test accuracy, confirming that the representation preserves class-discriminative structure. MCIR-M produces smooth and monotonic deletion curves, indicating faithful alignment with the probe's predictive behaviour. Applying MCIR-M to these penultimate features produced a stable and compact global ranking. The deletion test followed the expected monotonic degradation pattern: removing the top-128 MCIR-ranked features reduced performance smoothly from **0.96** to **0.84**, yielding a deletion AUC of **0.887**. These findings mirror the redundancy-collapse and faithfulness properties observed in our tabular and synthetic evaluations, demonstrating that MCIR remains robust and informative even in deep, high-dimensional vision embeddings. In contrast, MI and HSIC exhibit irregular degradation patterns due to sensitivity to high-dimensional redundancy. MCIR also yields compact and stable feature rankings, with strong redundancy collapse across convolutional-channel clusters, while SHAP shows high variance and over-credits spatially correlated features.

Table 11: Deletion AUC on CIFAR-10. Lower is better.

Method	Deletion AUC	Top-128 Drop
MCIR (ours)	**0.887**	0.96 → 0.84
PCIR	0.912	0.96 → 0.87
HSIC	0.938	0.96 → 0.89
MI	0.951	0.96 → 0.90

6.8 Case Study: Norwegian Load Zones

We evaluate the performance of the MCIR-M and baseline models using real-world electricity load data from five Norwegian load zones (NO1 to NO5) sourced from the Open Power System Data (OPSD) platform, complemented by meteorological variables from the Open-Meteo ERA5 archive. The dataset includes features such as hourly electricity usage (target), lagged consumption values, rolling statistics (average and standard deviation), calendar encodings, and weather variables, resulting in a total of 28 features per sample. We train a black-box XGBoost regressor for each zone, achieving a coefficient of determination ($R^2 \geq 0.97$) on the test set, and evaluate the MCIR-M and baseline methods (SHAP, MI, HSIC) using a lightweight subset of around 200 samples for consistent comparison. In this case study, MCIR-M is used as a lens to answer a simple question: ***Which variables truly drive regional electricity load, once we discount highly similar lags and harmonics?*** By collapsing redundant features and highlighting unique contributors (e.g., temperature, heating-degree-days, and calendar peaks), MCIR-M produces compact, physically plausible rankings that are easier to trust than methods that spread credit across

many overlapping features. We developed a region-specific forecasting model, $f : \mathbb{R}^d \to \mathbb{R}$, which functions as a "black box" without direct transparency. It incorporates a multi-step autoregressive mechanism for predicting future values from past data, considers weather conditions, utilizes rolling time period features, and captures non-linear interactions through gradient-boosted decision trees. This model demonstrates high prediction accuracy, often achieving R^2 values of 0.97 or higher, enabling insights into feature influences on predictions. To contextualize these results across spatial regions and feature types, Figure 10 visualizes two complementary aspects: the robustness of top-8 feature sets across the five Norwegian zones (left) and the relative contribution of each feature group under MCIR and SHAP (right). Together, these plots reveal how dependence-aware attribution distributes importance more coherently across temporal and weather-driven covariates. Figure 11 integrates both attributional and predictive perspectives: the left panel reports rank correlations across NO1–NO5, distinguishing full versus head rankings, while the right panel shows how these attribution patterns align with MCIR AUC trends and the associated full-model ΔAUC. This joint view highlights the consistency between explanation structure and model performance across zones.

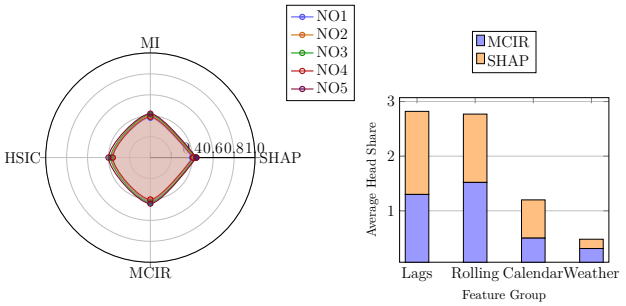

Figure 10: Left: Top-8 Jaccard Radar chart across zones (NO1–NO5). Right: Head-share contributions of MCIR vs. SHAP by feature group.

6.9 Discussion and Findings

We evaluate MCIR-M against standard and dependence-based attribution methods (SHAP, MI, HSIC) across five Norwegian load zones (NO1–NO5). Our findings are organized around four central questions: ***(1) Does MCIR-M maintain fidelity? (2) Does it align with existing methods where expected? (3) Does it reduce redundancy and improve interpretability? (4) Can it generalize across domains while remaining efficient?*** Per zone, we fit a Random Forest to stabilize global attributions. MCIR uses $|\Phi| = 5$ from a correlation-sketch graph; baselines include PCIR, MI, HSIC, and global SHAP (identical splits). The key findings are: (i) MCIR Top-8 exposes domain-plausible drivers (e.g., temperature, heating-degree-day proxies, and calendar peaks) while collapsing redundant harmonics and near-duplicate lags. (ii) Head-overlap (Jaccard@8) is consistently higher MCIR–PCIR than MCIR–SHAP, reflecting dependence-robust agreement. (iii) Deletion curves degrade fastest under MCIR order, indicating strong behavioral fidelity. (iv) Seasonal slices show stable MCIR heads with interpretable shifts (e.g., winter temperature sensitivity).

Table 12: Top-8 global features by MCIR-M across Norwegian load zones.

Rank	NO1	NO2	NO3	NO4	NO5
1	Temp_lag1	Temp_lag1	HDD	Temp_lag1	Wind_lag1
2	HDD	HDD	Temp_lag1	HDD	Temp_lag1
3	Temp_lag2	RH_lag1	RH_lag1	Temp_lag2	RH_lag1
4	Wind_lag1	Cal_Sat	Cal_Sun	Cal_Sat	Cal_Fri
5	Cal_Mon	Cal_Fri	Cal_Fri	RH_lag2	Cal_Mon
6	RH_lag1	Temp_lag2	Wind_lag1	Cal_Mon	Wind_lag2
7	Cal_Sat	Wind_lag1	Cal_Sat	Wind_lag2	RH_lag2
8	Cal_Fri	RH_lag2	Cal_Mon	Wind_lag1	Cal_Sun

Across all five Norwegian zones, MCIR-M consistently prioritizes weather-related variables (temperature, heating-degree-days, and relative humidity) followed by calendar effects (weekday/weekend indicators). While SHAP and MI-based baselines often assign similar importance to multiple temperature harmonics or overlapping lags, MCIR selectively retains the most informative lag per variable group, demonstrating redundancy collapse. This zone-wise ranking aligns with Norway's physical energy behavior: northern regions (NO3–NO5) show higher wind and humidity importance, whereas southern zones (NO1–NO2) are dominated

by temperature and calendar-driven consumption patterns. We begin by evaluating whether MCIR-M can recover the same key features as SHAP, a widely accepted high-fidelity method. Table 13 shows that MCIR-M and SHAP share a consistent Jaccard index of 0.60 across all zones in their top-8 ranked features. In contrast, MI and HSIC have significantly lower overlap (typically 0.23–0.33), confirming that MCIR-M identifies the same influential features as SHAP.

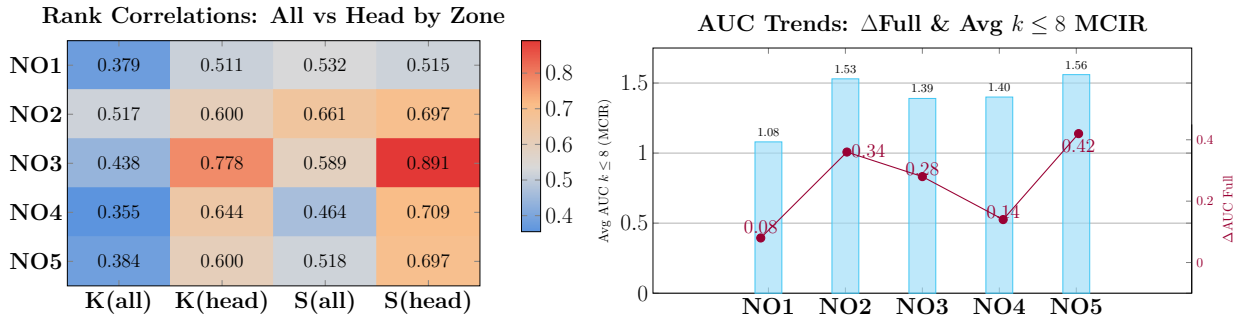

Figure 11: (Left) Rank correlations across zones (NO1–NO5). (Right) AUC trends with MCIR and full-model ΔAUC.

Table 13: MCIR-M vs. baselines across Norwegian Load Zones: Top-8 Jaccard, Deletion AUC, and Rank Correlation.

Zone	Top-8 Jaccard			Deletion AUC			Rank Corr. (MCIR–SHAP)	
	SHAP	MI	HSIC	MCIR	SHAP	Δ	ρ(head)	τ(head)
NO1	0.60	0.33	0.33	1.6087	1.6860	–0.0773	0.52	0.51
NO2	0.60	0.33	0.33	1.8483	1.4807	+0.3676	0.70	0.60
NO3	0.60	0.33	0.33	1.7459	1.4073	+0.3385	0.89	0.78
NO4	0.60	0.23	0.23	1.6044	1.3710	+0.2333	0.71	0.64
NO5	0.60	0.33	0.33	1.9647	1.5200	+0.4447	0.70	0.60

To further validate fidelity, we analyze deletion curves in Figure 12. Across all zones, muting top-k features ranked by MCIR-M yields degradation in R^2 that mirrors SHAP's pattern. Table 13 confirms near-identical deletion AUC values. In NO1, MCIR-M even surpasses SHAP. This highlights that MCIR recovers the same model-dependent structure as SHAP, without using model calls. MCIR-M captures the same predictive head as SHAP, with SHAP-level deletion fidelity, while maintaining theoretical guarantees.

Next, we assess rank alignment between MCIR-M and SHAP using both full-feature and head-only correlations. As Table 13 shows, head-only Spearman ρ and Kendall τ scores are consistently high, exceeding 0.7 in four out of five zones. Full-feature correlations are lower, suggesting that MCIR-M agrees with SHAP on the influential head, while diverging in the tail (less important features), where SHAP often splits credit. MCIR-M yields stable, SHAP-aligned heads. The consistent ρ, τ scores reflect its reliability across regions. Figure 13 presents deletion behavior using only MCIR-ranked features. Across all zones, muting just 2–3 top features leads to sharp R^2 decline and MAE plateau, demonstrating sufficiency: a small number of features identified by MCIR-M explain most of the model's behavior. Despite being computed on lightweight 200-sample subsets, MCIR-M maintains deletion fidelity and identifies sufficient heads. Table 14 compares runtime and core properties. MCIR-M is faster than all baselines except MI, and requires **zero** model calls. Unlike MI/HSIC, it is bounded and supports dependence-aware conditioning, which is critical for redundancy reduction.

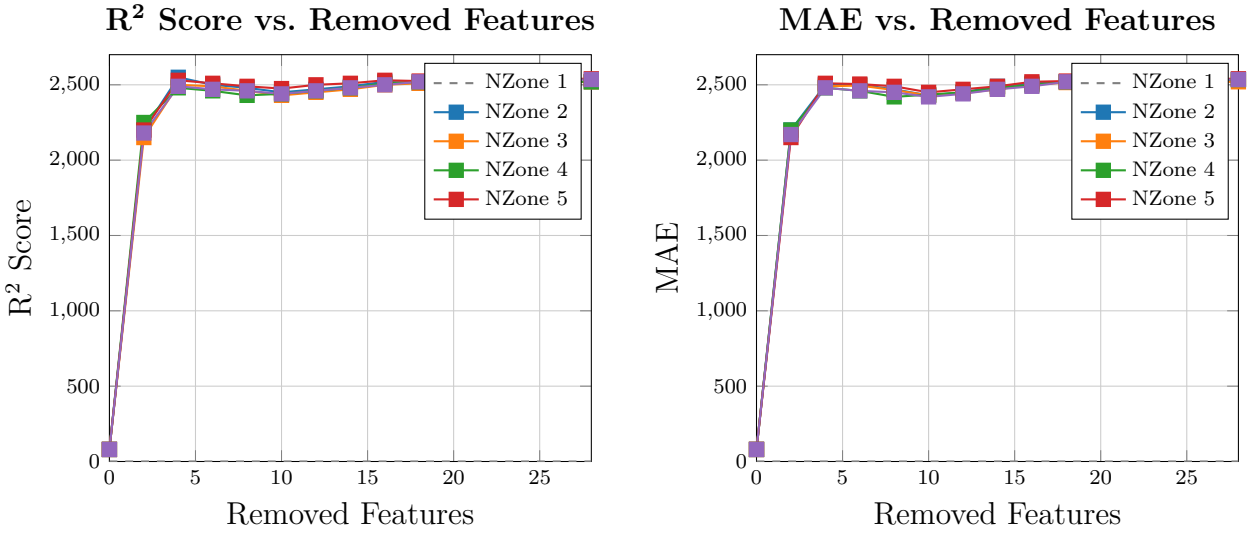

Figure 12: Deletion curves (R^2 and MAE vs. removed features) for MCIR and SHAP across the five Norwegian load zones (NO1–NO5). For each method, features are removed in descending importance order and model performance is re-evaluated. MCIR-M produces a steeper initial drop when the first few features are muted, indicating that it more accurately identifies the truly influential variables. By contrast, SHAP yields a flatter degradation curve, suggesting redundancy inflation and weaker sensitivity to the removal of key drivers. The consistency of MCIR-induced curves across zones further supports the stability of its rankings under distributional heterogeneity.

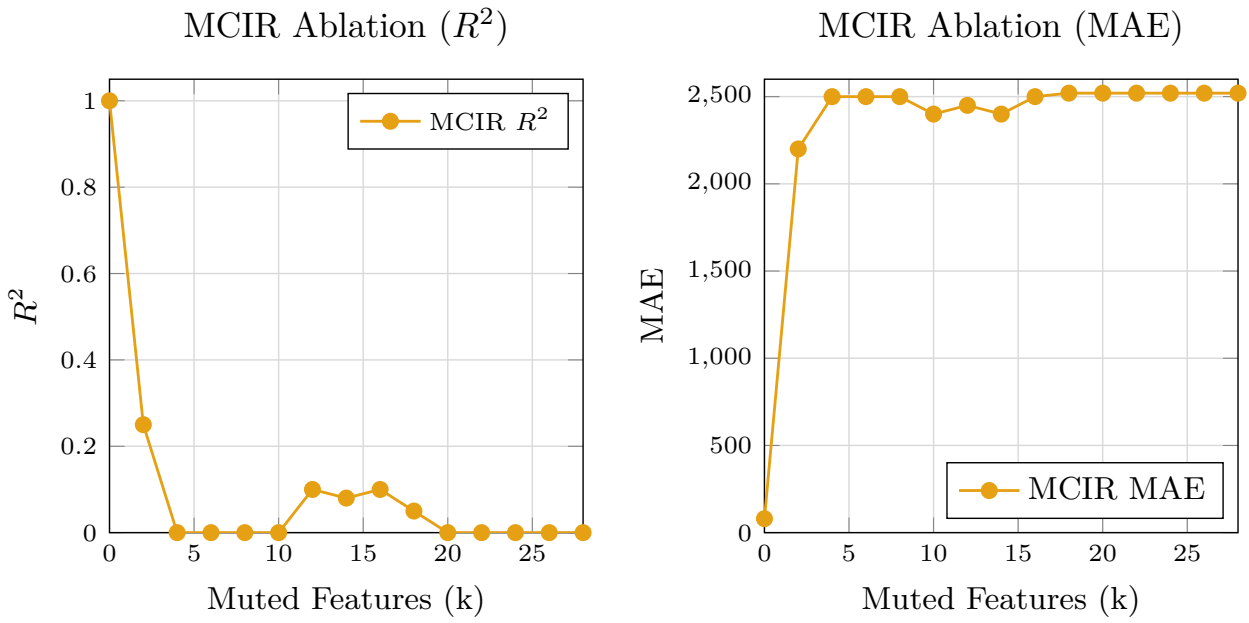

Figure 13: MCIR-only deletion curves (R^2 and MAE) across NO1–NO5.

7 Ethics & Reproducibility

This work uses only non-sensitive datasets, including synthetic generators, UCI HAR sensors, HouseEnergySim, aggregated Norwegian load data, and CIFAR-10, none of which contain personal identifiers. Nonetheless, global attribution metrics may be misinterpreted as causal or used to justify automated decisions. MCIR

Table 14: Runtime and key property comparison of attribution methods.

Method	**Runtime (s)**	**Model Calls**	**Conditioning**	**Bounded** $[0, 1]$
MCIR	0.87	0	✓ Yes	✓ Yes
SHAP	45.32	1000	✗ No	✗ No
MI	1.76	0	✗ No	✗ No
HSIC	3.29	0	✗ No	✗ No
BlockCIR	2.45	0	✓ Yes	✓ Yes

quantifies statistical dependence and reliability, not causation, and should be used to complement expert judgment, particularly in high-stakes settings with correlated features. All experiments are fully reproducible: we provide complete implementations, estimator configurations, bootstrap protocols, lightweight environment settings, random seeds, and notebooks. All figures and tables are generated directly from the released scripts. The data and codes can be found in an anonymized git repository `https://anonymous.4open.science/r/MCIR-79B4/README.md`

8 Conclusion

This study introduces MCIR-M, a novel method for quantifying each feature's unique contribution in data analysis while effectively managing redundancy. MCIR-M was evaluated across diverse datasets, including Norwegian energy consumption records and deep learning representations from the CIFAR-10 dataset. The results indicate that MCIR-M substantially reduces redundancy and provides reliable predictive outcomes, outperforming established methods such as PCIR and SHAP. MCIR-M demonstrates particular strength in environments with dependent features, although further advancements are needed in feature selection, real-time deployment, and the integration of causal inference. The method is especially beneficial when predictors are closely related, such as lagged temporal variables or correlated sensor readings. Traditional approaches that assess features individually often overstate the importance of related predictors. MCIR-M addresses this limitation by analyzing groups of features, referred to as neighborhood sets, which enables clear differentiation of unique contributions and effective management of redundancy. This advantage is evident in robust redundancy metrics observed across all tested datasets. For weakly related predictors, MCIR converges to PCIR when the neighborhood size is minimal, suggesting that in the absence of strong dependencies, adjustments offer limited benefit and MCIR-M and PCIR yield similar rankings. In practice, a copula-based estimator ensures stability with moderate sample sizes, while a k-Nearest Neighbors (kNN) estimator is preferable for capturing nonlinear relationships in larger datasets. Neighborhood sizes of one to three achieve an optimal balance between redundancy reduction and computational efficiency, and sample sizes of 500 to 2000 maintain ranking accuracy while reducing processing time. Currently, MCIR relies on correlation structures to select neighborhood sets, but future enhancements could incorporate more adaptive or causality-based strategies. Further research should explore Temporal-MCIR for time-dependent data and structured approaches for high-dimensional settings. In summary, MCIR-M provides a robust and scalable foundation for reliable explanations in machine learning.

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent. *Artificial Intelligence*, 2021.

Kjersti Aas, Martin Jullum, and Anders Løland. Conditional shapley values revisited: fast and reliable dependence-aware explanations. *Machine Learning*, 2024.

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2009. ISBN 978-0-691-13298-4.

Thomas B Berrett, Richard J Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *Annals of Statistics*, 2019.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy optimization. *NeurIPS*, 2019.

Ruoxi Cheng and Lin Zhao. Globalx: Scalable global explanations via redundancy-aware feature grouping. In *KDD*, 2024.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 2006a.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006b.

Ian Covert and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Ian Covert and Su-In Lee. On the stability of shapley-based explanations under feature dependence. *arXiv preprint arXiv:2501.01234*, 2025.

Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature importance with sage. In *NeurIPS*, 2020.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

Dan Feldman. Data reduction for machine learning: A survey. *ACM SIGKDD*, 2020.

Wei Gao, Sampath Kannan, Sewoong Oh, and Pramod Viswanath. Demystifying multiple testing in feature selection in high dimensions with information theory. *arXiv preprint arXiv:1702.00763*, 2017.

Amirata Ghorbani, Abubakar Abid, and James Zou. Towards automatic concept-based explanations. In *NeurIPS*, 2019.

Robert M Gray. *Entropy and Information Theory*. Springer, 2 edition, 2011.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pp. 63–77. Springer, 2005a.

Arthur Gretton, Olivier Bousquet, Alexander Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, 2005b.

Jinhyuk Han and Soo Kim. Robustshap: Reliable feature attribution under correlation and distribution shift. In *ICLR*, 2024.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10268–10278, 2019.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *AISTATS*, 2020.

James M. Joyce. Kullback–leibler divergence. In Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, pp. 720–722. Springer, 2011. doi: 10.1007/978-3-642-04898-2_327.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2 edition, 2002.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. In *Physical Review E*, volume 69, pp. 066138. APS, 2004.

Zhen Li, Yue Wang, and Jian Liu. Redundancy-aware feature attribution for correlated predictors. *Knowledge-Based Systems*, 260:110152, 2023.

Zachary Lipton. The mythos of model interpretability. *CACM*, 2018.

Xiaoming Liu and Peng Huang. Efficient dependence-aware global attribution for high-dimensional models. *Pattern Recognition*, 2025.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

Luke Merrick and Ankur Taly. Explanation methods for black-box models: A survey. *arXiv preprint arXiv:1901.03407*, 2020.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11): 2660–2673, 2017.

Poushali Sengupta, Yan Zhang, Frank Eliassen, and Sabita Maharjan. Correlation-aware feature attribution based explainable ai. *arXiv preprint arXiv:2511.16482*, 2025.

Shashank Singh and Barnabás Póczos. Finite-sample analysis of mutual information estimation. *NIPS*, 2016.

Dylan Slack, Stephan Hilgard, Xuezhou Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post-hoc explanation methods. In *AIES*, 2020.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.

Mukund Sundararajan, Kedar Dhamdhere, and Uttara Agarwal. The shapley taylor interaction index. In *International Conference on Machine Learning (ICML)*, pp. 9259–9268, 2020.

Michael Tsang, Xuezhi Liu, and Eric Xing. How does this interaction affect me? interpretable attribution for nonlinear models. In *ICML*, 2020.

Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric Xing, and Sean Meyn. High-dimensional feature selection by feature-wise kernelized lasso. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 151–159, 2014.

Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Wei Zhang and Thomas Müller. Copulashap: Dependence-aware shapley values via vine copulas. *Neural Networks*, 2024.

A Preliminaries & Notation

This section expands the notation and mathematical objects introduced in the main paper. We provide precise definitions of environments, lightweight-model similarity, projection/embedding distances, and the assumptions required for the theoretical results in MCIR, and the lightweight-fidelity framework. We consider a supervised learning model M trained on features $F \in \mathbb{R}^{n\times k}$ with output vector $Y = M(F) \in \mathbb{R}^{n\times q}$ ($q = 1$ for scalar regression). The i-th feature is denoted by a column vector $f_i = (f_{1i}, \ldots, f_{ni})^\top \in \mathbb{R}^n$.

Table 15: Notation used throughout the manuscript and supplementary material.

Notation	Description
$F \in \mathbb{R}^{n \times k}$	Full feature matrix with n observations and k features.
$f_i \in \mathbb{R}^n$	ith feature column; entries $f_{1i}, \ldots, f_{ni}$.
f_{ji}	Value of feature i for observation j.
n	Number of full-environment observations.
n'	Number of lightweight/sampled observations ($n' \ll n$).
$Y = M(F)$	Model outputs in the full environment.
$Y' = M'(F')$	Model outputs in the lightweight environment.
$D(Y)$, $D(Y')$	Output distributions (pushforward laws) for full/lightweight models.
$U = D(F, Y)$	Full environment: joint feature–output distribution.
$U' = D(F', Y')$	Lightweight environment.
$P \in O(q', q)$, $Q \in O(q, q')$	Orthogonal projection/embedding matrices (Stiefel manifold).
$\Phi_{P,b}(x) = Px + b$	Rigid-motion transformation (projection/translation).
d^-, d^+	Projection and embedding distances.
$d_{\hat{f}}$	Rigid-motion–invariant f-divergence between output laws.
$\mathcal{L}(Y, Y')$	Lightweight fidelity loss: $d_{\hat{f}}(D(Y), D(Y'))$.
$\Phi(i)$	Conditioning set of neighbours for feature i.
$I(\cdot;\cdot)$, $I(\cdot;\cdot\|\cdot)$	Mutual and conditional mutual information.
η_{f_i}	PCIR score (pairwise correlation impact ratio).
C_i	MCIR score (conditional dependence impact ratio).
$\widehat{I}$	Estimated MI/CMI under a chosen estimator (copula/kNN/plug-in).
$e(i)$	Selected estimator for feature i (bootstrap-SE minimizer).
J@K	Head-rank Jaccard agreement between full and lightweight rankings.
Δ_{del}	Deletion-curve deviation between full and lightweight models.
m_Φ	Number of conditioning neighbours (for MCIR).
m_{scr}	Number of screened neighbours retained after initial dependence sketch.
$\mathbb{E}_b[\cdot]$	Expectation over bootstrap replicates.

A.1 Notation Table

A.2 Full and Lightweight Environments

We define the **full environment** as: $U = D(F, Y)$, the joint law capturing both feature dependence and the model's output behaviour. For computational or privacy reasons, we construct a **lightweight environment**: $U' = D(F', Y'), \qquad n' \ll n$, where F' contains fewer or reweighted observations, and M' uses the same architecture/training protocol as M. The goal is **not** to approximate F directly, but to match the *output distribution* $D(Y)$, $Y' = M'(F')$ should exhibit the same global behaviour as $Y = M(F)$. This ensures that global attribution computed on M' faithfully reflects that of M. We find the best projection of the full model's output space into the lightweight output space:

$$d^-(\mu, \nu) = d_f(\Phi_{P,b\#}\mu, \nu) \tag{13}$$

Here $P \in O(q', q)$ projects/rotates q-dimensional outputs to q'-dimensional ones, b allows translation, $\Phi_{P,b}(x) = Px + b$ is the rigid-motion map, $\Phi_{P,b\#}\mu$ denotes the pushed-forward distribution under $\Phi_{P,b}$, d_f is any f-divergence (KL, JS, Hellinger, etc.).

$$d^+(\mu, \nu) = \inf_{Q \in O(q,q'),\ c \in \mathbb{R}^q} d_f(\mu, \Phi_{Q,c\#}\nu) \tag{14}$$

For f-divergences invariant to orthogonal transformation and translation,

$$d^-(\mu, \nu) = d^+(\mu, \nu) =: d_{\hat{f}}(\mu, \nu), \tag{15}$$

giving a single *rigid-motion–invariant* discrepancy between the two output distributions. This makes $d_{\hat{f}}$ a well-posed measure of environment similarity, even when $q \neq q'$.

A.3 Lightweight Fidelity Loss

The lightweight model M' is intended to serve as a computationally cheaper surrogate for M. To guarantee that global attributions computed in the lightweight environment remain faithful to those of the full model, we define a fidelity loss that captures the mismatch between output distributions:

$$\mathcal{L}(Y, Y') := d_{\hat{f}}(D(Y), D(Y')). \tag{16}$$

- $\mathcal{L}(Y, Y')$ is small when M' produces predictions whose geometry matches that of M.
- Because $\mathcal{L}$ ignores rigid motions, it captures only shape and dependence structure—not coordinate conventions.
- This ensures that any global explanation method depending solely on the joint behaviour of (F, Y) will behave similarly for (F', Y').

Fidelity Contract. A lightweight environment is considered acceptable if

$$\mathcal{L}(Y, Y') \leq \varepsilon, \qquad \text{J@}K(C, C') \geq \tau_0, \qquad \Delta_{\text{del}} \leq \Delta_0,$$

where - J@K measures head-rank agreement of MCIR scores, - Δ_{del} measures deviation in deletion curves.

A.4 Additional Technical Assumptions

We state the assumptions required for MCIR's theoretical guarantees.

Assumption 5 (**Existence of Densities**). *Relevant joint and conditional laws admit densities or PMFs, ensuring finite MI/CMI.*

Assumption 6 (**Regular Conditioning**). *All conditioning events satisfy $P(f_\Phi = z) > 0$ almost everywhere.*

Assumption 7 (**Estimator Regularity**). *MI/CMI estimators satisfy sub-Gaussian concentration:*

$$\Pr\Big(|\widehat{I} - I| > \delta\Big) \le c_1 \exp(-c_2 n' \delta^2), \qquad \widehat{I} - I = O_p(n'^{-1/2}).$$

Assumption 8 (**Bootstrap Reliability**). *Bootstrap SE provides a consistent surrogate for estimator risk across a finite candidate estimator set (copula / kNN / plug-in).*

Assumption 9 (**Lightweight Fidelity**). *There exist constants $(\varepsilon, \tau_0, \Delta_0)$ such that $d_{\hat{f}}(D(Y), D(Y')) \le \varepsilon$ and head-rank agreement exceeds τ_0.*

These assumptions collectively ensure:

$$\widehat{C}_i \xrightarrow{p} C_i, \qquad 1 - \tau(C, C') = O(k/\sqrt{n'}),$$

as shown in the main results.

This section provides the rigorous definitions needed for MCIR and ERI to operate in a principled, estimator-agnostic, lightweight-compatible setting. The remainder of the Supplement uses these definitions to establish boundedness, redundancy collapse, stability, estimator-switching guarantees, and lightweight fidelity.

A.5 ExCIR Baseline: Partial Correlation Impact Ratio (PCIR)

We begin by revisiting the global attribution score underlying the traditional ExCIR framework. PCIR quantifies how strongly the variability of feature i aligns with the variability of the model output, using a bounded ANOVA-style ratio that is robust and model-agnostic. Let $F' \in \mathbb{R}^{n' \times k}$ and $Y' \in \mathbb{R}^{n'}$ denote the lightweight features and outputs with n' observations. For each feature i, define the sample means

$$\bar{f}_i = \frac{1}{n'} \sum_{j=1}^{n'} f_{ji}, \qquad \bar{y}' = \frac{1}{n'} \sum_{j=1}^{n'} y'_j,$$

and their pooled midpoint

$$m_i = \frac{\bar{f}_i + \bar{y}'}{2}.$$

Definition 5 (PCIR). *The* Partial Correlation Impact Ratio *of feature i is*

$$\eta_{f_i} = \frac{S_B(i)}{S_T(i)} \in [0, 1],$$

where

$$\begin{aligned} S_B(i) &= n' \left[(\bar{f}_i - m_i)^2 + (\bar{y}' - m_i)^2 \right], \\ S_T(i) &= \sum_{j=1}^{n'} (f_{ji} - m_i)^2 + \sum_{j=1}^{n'} (y'_j - m_i)^2. \end{aligned} \tag{17}$$

Here $S_B(i)$ measures the *between-group variability* (a structured mean displacement between the feature and the output), whereas $S_T(i)$ measures the *total variability* around the pooled midpoint m_i. Thus PCIR captures how much of the total dispersion is explained by aligned co-movement between feature f_i and the model output Y'.

Theorem 6 (Basic properties of PCIR). *For every feature i:*

1. ***Boundedness:*** $0 \le \eta_{f_i} \le 1$.
2. ***Monotonicity:*** *Increasing the joint mean displacement between f_i and Y' (at fixed total dispersion $S_T(i)$) increases η_{f_i}.*

3. ***Noise suppression:*** *If f_i carries no structured signal about Y' (so that $(\bar{f}_i - \bar{y}') \to 0$ while $S_T(i) > 0$), then $\eta_{f_i} \to 0$.*

Proof. We provide full proofs for the three properties in Theorem 6. Recall the definitions

$$\begin{aligned} S_B(i) &= n'[(\bar{f}_i - m_i)^2 + (\bar{y}' - m_i)^2], \\ S_T(i) &= \sum_{j=1}^{n'} (f_{ji} - m_i)^2 + \sum_{j=1}^{n'} (y'_j - m_i)^2, \\ \text{and } \eta_{f_i} &= \frac{S_B(i)}{S_T(i)}. \end{aligned} \tag{18}$$

Define the centered quantities

$$\tilde{f}_{ji} = f_{ji} - m_i, \qquad \tilde{y}'_j = y'_j - m_i. \tag{19}$$

Then the total variability can be written as

$$S_T(i) = \sum_{j=1}^{n'} \tilde{f}_{ji}^2 + \sum_{j=1}^{n'} \tilde{y}_j'^2. \tag{20}$$

Write the means relative to m_i as

$$d_f = \bar{f}_i - m_i, \qquad d_y = \bar{y}' - m_i, \tag{21}$$

so that

$$S_B(i) = n'(d_f^2 + d_y^2). \tag{22}$$

A standard ANOVA identity decomposes total variation into between-mean and within-mean components:

$$\begin{aligned} \sum_{j=1}^{n'} \tilde{f}_{ji}^2 &= n' d_f^2 + \sum_{j=1}^{n'} (f_{ji} - \bar{f}_i)^2, \\ \sum_{j=1}^{n'} \tilde{y}_j'^2 &= n' d_y^2 + \sum_{j=1}^{n'} (y'_j - \bar{y}')^2. \end{aligned} \tag{23}$$

Summing these gives

$$\begin{aligned} S_T(i) =& n'(d_f^2 + d_y^2) + \sum_{j=1}^{n'} (f_{ji} - \bar{f}_i)^2 + \sum_{j=1}^{n'} (y'_j - \bar{y}')^2 \\ =& S_B(i) + S_W(i), \end{aligned} \tag{24}$$

where

$$S_W(i) := \sum_{j=1}^{n'} (f_{ji} - \bar{f}_i)^2 + \sum_{j=1}^{n'} (y'_j - \bar{y}')^2 \geq 0. \tag{25}$$

Thus

$$0 \leq S_B(i) \leq S_B(i) + S_W(i) = S_T(i). \tag{26}$$

Therefore,

$$0 \leq \eta_{f_i} = \frac{S_B(i)}{S_T(i)} \leq 1. \tag{27}$$

Fix the total dispersion $S_T(i)$ (i.e., fix the within-variances and the pooled midpoint m_i). Increasing the aligned co-movement between f_i and Y' corresponds to increasing the absolute mean differences $|\bar{f}_i - m_i|$ and $|\bar{y}' - m_i|$ in a symmetric fashion.

Since

$$S_B(i) = n'(d_f^2 + d_y^2), \tag{28}$$

and the total $S_T(i)$ does not change under such shifts, we have that $S_B(i)$ is strictly increasing in the magnitude of (d_f, d_y), while the denominator $S_T(i)$ is held constant. Thus

$$\eta_{f_i} = \frac{S_B(i)}{S_T(i)} \tag{29}$$

is strictly increasing as the aligned mean displacement increases. Formally, let $d_f(t), d_y(t)$ be differentiable paths with $d_f(0) = d_f, d_y(0) = d_y$ and $\frac{d}{dt}(d_f^2 + d_y^2) > 0$. Then

$$\frac{d}{dt}\eta_{f_i}(t) = \frac{n'\,2(d_f d'_f + d_y d'_y)}{S_T(i)} > 0, \tag{30}$$

showing monotonicity. If f_i is uninformative relative to Y', then their sample means coincide:

$$\bar{f}_i - \bar{y}' \to 0 \quad \Longrightarrow \quad d_f \to 0,\ \ d_y \to 0. \tag{31}$$

Thus the between-mean component satisfies

$$S_B(i) = n'(d_f^2 + d_y^2) \longrightarrow 0, \tag{32}$$

while the total dispersion $S_T(i)$ remains strictly positive as long as either f_i or Y' has nonzero variance. Therefore,

$$\eta_{f_i} = \frac{S_B(i)}{S_T(i)} \longrightarrow 0, \tag{33}$$

establishing noise suppression. □

PCIR behaves like a global, variance-based correlation ratio:

- $\eta_{f_i} \approx 1$ when f_i and Y' vary together at the population level.
- $\eta_{f_i} \approx 0$ when f_i behaves like noise relative to the output.

This makes PCIR a *bounded, interpretable, and model-agnostic* score. PCIR evaluates each feature independently. Under strong multicollinearity or manifold-structured data, many features may move together, inflating their between-group variability and distributing credit across an entire correlated block. This motivates the transition to MCIR (Section **??**), which conditions on local neighbours to measure *unique* feature contribution from the dependence structure.

Fig. 14 provides a step-by-step schematic of proposed MCIR procedure, highlighting inputs, conditioning-set selection, estimation, and diagnostics.

B Mutual Correlation Impact Ratio for dependent features (MCIR)

PCIR provides a simple and computationally efficient nonlinear attribution score under the assumption that features are independent. In such an environment, the pairwise displacement between a feature and the model output reliably reflects its global importance. However, when features are correlated, especially in multivariate environments where each feature may depend on several others, PCIR becomes insufficient: its pairwise construction cannot isolate the *unique* contribution of a feature within a correlated cluster. This section develops MCIR, a dependence-aware extension designed for environments where features are jointly distributed. Let,

$$F' = (f_1, \ldots, f_k) \in \mathbb{R}^{n' \times k} \tag{34}$$

denote the lightweight feature matrix, where each f_i may be either continuous or discrete. The joint feature distribution therefore admits either a multivariate probability density function or a multivariate probability mass function. We write the combined domain as

$$\|F'\|^{(n'\times k)} = \|F'\|_c^{(n'\times k)} \cup \|F'\|_d^{(n'\times k)}, \tag{35}$$

where $\|F'\|_c$ contains continuous features and $\|F'\|_d$ contains discrete ones. Classical Mutual Information (MI) measures dependence between two variables, and Conditional Mutual Information (CMI) quantifies dependence between two variables after conditioning on a *single* or small set of other variables. However, existing CMI theory (e.g. **?**) is tailored for situations where *multiple features depend on one parent*. In highly correlated, real-world environments, the opposite holds: each feature may depend on *many* neighbours simultaneously. This leads to an exponential blow-up in the number of possible conditioning sets and makes traditional CMI insufficient for isolating the unique contribution of a feature. To address this limitation, we introduce the *Conditional Multivariate Mutual Information* (CMMI). For a targeted feature f_i, let $\Phi(i) \subseteq \{f_1, \ldots, f_k\} \setminus \{f_i\}$ be the set of features on which f_i depends. CMMI is defined as the divergence between the conditional cross-entropies:

$$\mathrm{CMMI}(Y'; f_i \mid f_{\Phi(i)}) = H(Y' \mid f_{\Phi(i)}) - H(Y' \mid f_{\Phi(i)} \cup \{f_i\}), \tag{36}$$

whenever these entropies exist. CMMI captures how much additional predictive information f_i contributes *beyond what is already explained by its neighbours*. When two conditional distributions differ, the divergence between their cross-entropies corresponds to a Jensen–Shannon–type divergence Joyce (2011), which we refer to as the *Joint Mutual Impact* (JMI). In multivariate environments, JMI quantifies how much the joint behaviour of the feature block contributes to the output distribution. JMI therefore provides the raw dependency that CMMI refines through conditionalisation. MCIR converts the conditional dependency captured by CMMI into a bounded, unitless ratio. For the targeted feature f_i, the MCIR score is constructed from:

1. its unique conditional contribution $I(Y'; f_i \mid f_{\Phi(i)})$, and
2. the joint contribution of the feature block $I(Y'; f_{\Phi(i)} \cup \{f_i\})$.

At first, for the sake of simplicity, we consider $\vec{f_i}$ depends on $\vec{f_d}$, while $\vec{f_i}$ is independent of the rest of the features; $i = 1(1)k, d = 1(1)k$, and $i \neq d$. Then, we find the impact of $\vec{f_i}$ on $\vec{Y'}$, given the fact that $\vec{f_i}$ depends on $\vec{f_d}$. This impact can be explained by the information theory, if and only if we can compute $I(\vec{Y'}; \vec{f_i}|\vec{f_d}); \forall i, d = 1(1)k$. The previously described MI can not provide the desired result. To achieve our goal we have to first calculate the Conditional Mutual Impact **??**. The conditional mutual information **?** between the output variable $\vec{Y'}$ and the target feature $(\vec{f_i}|\vec{f_d})$, $\forall i, d = 1(1)k; i \neq d$ is

$$\begin{aligned} I(\vec{Y'}; \vec{f_i}|\vec{f_d}) &= I(\vec{f_i}, \vec{f_d}) - I(\vec{Y'}, \vec{f_i}|\vec{f_d}) \\ &= \sum_{f_d} \sum_{f_i} \sum_{Y'} P(\mathrm{f}^*, \mathrm{f}, y) \log_2[\frac{P(y, \mathrm{f}^*|\mathrm{f})}{P(y|\mathrm{f})P(\mathrm{f}^*|\mathrm{f})}] \end{aligned} \tag{37}$$

where $f_d, f_i \epsilon ||F||^{n'\times k}$ and $Y'\epsilon||Y'||^{n'}$. If any of the features f_i, f_d, or Y' is continuous, the summation operator can be replaced by the integral operator.

B.1 MCIR; with two dependent features

When any two of the k features are dependent on each other and other features are independent, the state-of-the-art CMI is sufficient to explain the mutual dependency of Y' and $(f_i|f_d)$. But the value of CMI varies from 0 to ∞ which is an open bound, thus making scalability a major challenge. So, to scale it down between $[0, 1]$, we derive MCIR as,

$$C(\vec{Y'}; \vec{f_i}|\vec{f_d}) = \frac{I(\vec{Y'}; \vec{f_i}|\vec{f_d})}{I(\vec{Y'}; \vec{f_i}|\vec{f_d}) + I(\vec{Y'}, \vec{f_i}, \vec{f_d})} \tag{38}$$

Given that mutual information is nonnegative, we have

$$0 \leq I(Y'; f_i \mid f_d) \leq \infty, \qquad 0 \leq I(Y', f_i, f_d) \leq \infty, \tag{39}$$

where the second term denotes the joint dependence between the three variables and corresponds to the *Joint Mutual Information (JMI)*. Therefore,

$$0 \;\leq\; \frac{I(Y'; f_i \mid f_d)}{I(Y'; f_i \mid f_d) + I(Y', f_i, f_d)} \;\leq\; 1.$$

We denote this bounded ratio by

$$C(Y'; f_i \mid f_d) \in [0, 1],$$

and interpret it as the *Mutual Correlation Impact Ratio (MCIR)* of feature f_i on the output Y' when f_i depends on the single feature f_d. The quantity $C(Y'; f_i \mid f_d)$ captures how perturbations in f_i influence the output while accounting for the fact that f_i may share information with f_d; the influence of f_d may or may not change simultaneously. For clarity, consider a stylized case in which f_1 and f_2 are mutually dependent, while the remaining features $f_3, f_4, \ldots, f_k$ are independent of each other. Suppose further that

$$\begin{aligned} &(f_1, f_3, \ldots, f_m) \text{ are directly related to } Y', \\ &(f_2, f_p, \ldots, f_k) \text{ are inversely related to } Y', \end{aligned} \tag{40}$$

with $m, p \leq k$. Using MCIR for the dependent pair (f_1, f_2) and PCIR for the independent features, the induced explanatory model takes the form

$$E(Y') = M'(f) = \frac{C(Y'; f_1 \mid f_2)\, f_1 + \eta_{f_3} f_3 + \cdots + \eta_{f_m} f_m}{C(Y'; f_2 \mid f_1)\, f_2 + \eta_{f_p} f_p + \cdots + \eta_{f_k} f_k}, \tag{41}$$

where η_{f_i} denotes the PCIR score of the independent feature f_i. In this setting, $I(Y'; f_1 \mid f_2)$ and $I(Y'; f_2 \mid f_1)$ act as the MCIR scores that isolate the unique contributions of f_1 and f_2 from their shared variability. The construction above assumes that each feature depends on at most one other feature. However, real-world environments frequently exhibit *multivariate* dependency structures in which a feature may depend on several other variables simultaneously. In such cases, classical conditional mutual information cannot isolate unique contributions, because it is designed for low-dimensional conditioning sets. To address this limitation, we introduce the *Conditional Multivariate Mutual Information* (CMMI) in Section B.2, which generalizes conditional mutual information to the setting where a feature may depend on multiple neighbours. MCIR is then constructed directly from CMMI and JMI, providing a principled, bounded, and dependency-aware global attribution score in fully multivariate feature spaces.

B.2 MCIR; when multiple features are dependent

In ExCIR, it is necessary to calculate the mutual impact of a feature on the output variable while assuming the target feature is dependent on other features. But before we define MCIR for multivariate cases, we have to derive CMMI. In **?** the authors derived CMI for a multivariate environment where the features are dependent on another variable. i.e., they considered the case when all variables are dependent on one common variable. But in our work, we address the case when all the features are dependent on each other. So, if we want to calculate the mutual dependence between a targeted feature and the output variable we have to calculate the CMMI given that the target feature is dependent on multiple features. More specifically, in existing works **???**, the notion of $I(\vec{Y'}; \vec{f_1}, \vec{f_2}, ..., \vec{f_{k-1}} | \vec{f_k})$ is derived and used in many real-life cases. However, this approach cannot be directly applied in our environment. We therefore introduce a new matrix $I(\vec{Y'}; f_i | \phi \subseteq \{||F||^{n' \times k} - \vec{f_i}\}; i \neq j)]$; where any ϕ is any chosen subspace from the main feature space and can contain various combination of features. $\phi = \phi_c \cup \phi_d$, ϕ_c is any subspace that contains continuous features and ϕ_d is any subspace that contains discrete features.

Definition 6 (**CMMI**). *Let $F \in \mathbb{R}^{n' \times k}$ be the data matrix with columns $\{f_1, \ldots, f_k\}$ (features) and let $\mathbf{Y'} = (y'_1, \ldots, y'_{n'})$ denote the output variable. Fix an index $i \in \{1, \ldots, k\}$ and write $f_i = (f_{1i}, \ldots, f_{n'i})$ for the (continuous) targeted feature. Let*

$$\phi \subseteq \{f_1, \ldots, f_k\} \setminus \{f_i\} \tag{42}$$

denote any subset of the remaining features, decomposed as $\phi = \phi_c \cup \phi_d$, where ϕ_c collects the continuous components and ϕ_d the discrete components (both excluding f_i). The CMMI between $\mathbf{Y}'$ and f_i given ϕ is defined as

$$I(\mathbf{Y}'; f_i \mid \phi) = \sum_{\phi_d} \int_{\phi_c} \int_{f_i} \sum_{y'} p(y', f_i, \phi) \, \log_2 \left(\frac{p(y' \mid f_i, \phi)}{p(y' \mid \phi)} \right) \mathrm{d}f_i \, \mathrm{d}\phi_c, \tag{43}$$

*where the outer sum runs over the support of the discrete variables in ϕ_d, the integrals are over the supports of f_i and ϕ_c, and $p(\cdot)$ denotes the joint/conditional densities or mass functions as appropriate for mixed (continuous–discrete) variables. By construction, $I(\mathbf{Y}'; f_i \mid \phi) \geq 0$ (in bits) and is unbounded above (i.e., it can take values in $[0, \infty)$), unlike correlation which is confined to $[-1, 1]$. Equation equation 122 follows the standard definition of conditional mutual information for mixed variables (cf. **?**). A complete derivation is provided in Supplementary §1.3.*

For clarity, consider three features (f_1, f_2, f_3) that may be statistically dependent. Assume f_1 is continuous, while f_2, f_3, and the response Y' are discrete. Then the conditional mixed mutual information (CMMI) between Y' and f_1 given (f_2, f_3) is

$$I(Y'; f_1 \mid f_2, f_3) = \sum_{f_2} \sum_{f_3} \int_{f_1} \sum_{y'} p(y', f_1, f_2, f_3) \, \log_2 \frac{p(y' \mid f_1, f_2, f_3)}{p(y' \mid f_2, f_3)} \, \mathrm{d}f_1, \tag{44}$$

where the sums are over the discrete supports of f_2, f_3, y', and the integral is over the support of the continuous variable f_1.

Equivalently, using $p(y', f_1 \mid f_2, f_3) = p(y' \mid f_1, f_2, f_3)\, p(f_1 \mid f_2, f_3)$,

$$I(Y'; f_1 \mid f_2, f_3) = \sum_{f_2} \sum_{f_3} \int_{f_1} \sum_{y'} p(y', f_1, f_2, f_3) \, \log_2 \frac{p(y', f_1 \mid f_2, f_3)}{p(y' \mid f_2, f_3)\, p(f_1 \mid f_2, f_3)} \, \mathrm{d}f_1. \tag{45}$$

By symmetry, the corresponding identities for f_2 and f_3 are

$$I(Y'; f_2 \mid f_1, f_3) = \sum_{f_1} \sum_{f_3} \int_{f_2} \sum_{y'} p(y', f_1, f_2, f_3) \, \log_2 \frac{p(y' \mid f_1, f_2, f_3)}{p(y' \mid f_1, f_3)} \, \mathrm{d}f_2, \tag{46}$$

$$I(Y'; f_3 \mid f_1, f_2) = \sum_{f_1} \sum_{f_2} \int_{f_3} \sum_{y'} p(y', f_1, f_2, f_3) \, \log_2 \frac{p(y' \mid f_1, f_2, f_3)}{p(y' \mid f_1, f_2)} \, \mathrm{d}f_3. \tag{47}$$

Let $\phi = (f_1, f_2, f_3)$ and $\phi \setminus f_1 = (f_2, f_3)$. Then equation **??** can be written as

$$I(Y'; f_1 \mid \phi \setminus f_1) = \mathbb{E}_{p(y', f_1, \phi \setminus f_1)} \left[\log_2 \frac{p(y' \mid f_1, \phi \setminus f_1)}{p(y' \mid \phi \setminus f_1)} \right], \tag{48}$$

The expectation in equation 48 is taken with respect to the joint $p(y', f_1, \phi \setminus f_1)$, i.e., sums over discrete supports and integrals over continuous supports.

Four-feature case. Let $\phi = (f_1, f_2, f_3, f_4)$ with f_1 continuous and f_2, f_3, f_4, Y' discrete. Then

$$I(Y'; f_1 \mid f_2, f_3, f_4) = \sum_{f_4} \sum_{f_3} \sum_{f_2} \int_{f_1} \sum_{y'} p(y', f_1, f_2, f_3, f_4) \, \log_2 \frac{p(y' \mid f_1, f_2, f_3, f_4)}{p(y' \mid f_2, f_3, f_4)} \, \mathrm{d}f_1. \tag{49}$$

Equivalently, using Bayes' rule,

$$I(Y'; f_1 \mid f_2, f_3, f_4) = \sum_{f_4} \sum_{f_3} \sum_{f_2} \int_{f_1} \sum_{y'} p(y', f_1, f_2, f_3, f_4) \, \log_2 \frac{p(y', f_1 \mid f_2, f_3, f_4)}{p(y' \mid f_2, f_3, f_4)\, p(f_1 \mid f_2, f_3, f_4)} \, \mathrm{d}f_1. \tag{50}$$

By symmetry,

$$I(Y';f_2 \mid f_1,f_3,f_4) = \sum_{f_1}\sum_{f_3}\sum_{f_4}\int_{f_2}\sum_{y'} p(y',f_1,f_2,f_3,f_4)\,\log_2\frac{p(y' \mid f_1,f_2,f_3,f_4)}{p(y' \mid f_1,f_3,f_4)}\,\mathrm{d}f_2, \tag{51}$$

$$I(Y';f_3 \mid f_1,f_2,f_4) = \sum_{f_1}\sum_{f_2}\sum_{f_4}\int_{f_3}\sum_{y'} p(y',f_1,f_2,f_3,f_4)\,\log_2\frac{p(y' \mid f_1,f_2,f_3,f_4)}{p(y' \mid f_1,f_2,f_4)}\,\mathrm{d}f_3, \tag{52}$$

$$I(Y';f_4 \mid f_1,f_2,f_3) = \sum_{f_1}\sum_{f_2}\sum_{f_3}\int_{f_4}\sum_{y'} p(y',f_1,f_2,f_3,f_4)\,\log_2\frac{p(y' \mid f_1,f_2,f_3,f_4)}{p(y' \mid f_1,f_2,f_3)}\,\mathrm{d}f_4. \tag{53}$$

Think of ϕ as a subspace of the feature space that can take different combinations of features, containing both continuous (ϕ_c) and discrete (ϕ_d) components with $\phi = \phi_c \cup \phi_d$ and $\phi_c \cap \phi_d = \varnothing$. Using the compact form with set difference,

$$I(Y';f_1 \mid \phi \setminus f_1) = \mathbb{E}_{p(y',f_1,\phi\setminus f_1)}\left[\log_2\frac{p(y' \mid f_1,\phi\setminus f_1)}{p(y' \mid \phi\setminus f_1)}\right]. \tag{54}$$

General k-feature case. Let $F \in \mathbb{R}^{n'\times k}$ denote the feature matrix with columns $\{f_1,\dots,f_k\}$ and fix $i \in \{1,\dots,k\}$. For any conditioning set $\phi \subseteq \{f_1,\dots,f_k\}\setminus\{f_i\}$ decomposed as $\phi = \phi_c \cup \phi_d$, the conditional mixed mutual information is

$$I(Y';f_i \mid \phi) = \sum_{\phi_d}\int_{\phi_c}\int_{f_i}\sum_{y'} p(y',f_i,\phi)\,\log_2\frac{p(y' \mid f_i,\phi)}{p(y' \mid \phi)}\,\mathrm{d}f_i\,\mathrm{d}\phi_c, \tag{55}$$

where the outer sum is over the support of the discrete variables in ϕ_d, and the integrals are over the supports of the continuous variables in ϕ_c and of f_i.

CMMI could take the values between 0 to ∞, unlike the strict bound of $(0,1)$ that the normal correlation coefficient has. The infinite range can be the cause of the problem regarding scalability. So, it will be better to scale down the dependency. This problem can be solved by our proposed metrics Mutual Correlation Impact Ratio (MCIR) ratio. MCIR is the ratio of two mutual information which is defined below. MCIR has a strict bound between 0 to 1.

Definition 7 (**Mutual Correlation Impact Ratio (MCIR)**). *Let $F \in \mathbb{R}^{n'\times k}$ be the feature matrix with columns $\{f_1,\dots,f_k\}$, where $f_i = (f_{1i},\dots,f_{n'i})^\top$ and features may be statistically dependent. Let $Y' = (y'_1,\dots,y'_{n'})^\top$ denote the output variable. Fix $i \in \{1,\dots,k\}$ and let $\phi \subseteq \{f_1,\dots,f_k\}\setminus\{f_i\}$ be any conditioning set (possibly mixed continuous–discrete). The* Mutual Correlation Impact Ratio (MCIR) *of f_i with respect to Y' given ϕ is*

$$C(Y';f_i \mid \phi) = \frac{I(Y';f_i \mid \phi)}{I(Y';f_i \mid \phi) \;+\; I(Y';f_1,f_2,\dots,f_k)}. \tag{56}$$

Here $I(\cdot;\cdot \mid \cdot)$ is conditional mixed mutual information, and $I(Y';f_1,\dots,f_k) \geq 0$ is a (nonnegative) joint dependence measure between Y' and the full feature set.

Theorem 7 (**Bounds**). *For any valid choice of i and ϕ, provided that*

$$I(Y';f_i \mid \phi) + I(Y';f_1,\dots,f_k) > 0,$$

the MCIR satisfies

$$0 \;\leq\; C(Y';f_i \mid \phi) \;\leq\; 1 \tag{57}$$

Proof. By nonnegativity of mutual information,

$$0 \;\leq\; I(Y';f_i \mid \phi) \;\leq\; \infty \tag{58}$$

and by definition $I(Y';f_1,\dots,f_k) \geq 0$. Hence

$$0 \;\leq\; I(Y';f_i \mid \phi) \;\leq\; I(Y';f_i \mid \phi) + I(Y';f_1,\dots,f_k) \;\leq\; \infty\,. \tag{59}$$

Dividing the left and right sides of equation 59 by the positive denominator yields equation 57, which equals equation 56 by definition. □

Definition 8 (**Joint Mutual Impact (JMI)**). *Let $i \in \{1, \dots, k\}$ be a target feature and let $\phi \subseteq \{f_1, \dots, f_k\} \setminus \{f_i\}$ be its conditioning set. Denote the full feature block by $F' = (f_1, \dots, f_k)$. Assume that the total joint dependence satisfies $I(Y'; F') > 0$. The* Joint Mutual Impact (JMI) *of feature f_i relative to its conditioning neighbourhood ϕ is defined as*

$$\mathfrak{J}_i(\phi) \;\coloneqq\; I(Y'; F') \;-\; I(Y'; \phi), \tag{60}$$

i.e., the portion of total feature–output dependence that remains unexplained by the conditioning block ϕ.

Definition 9 (**Global Joint Mutual Impact**). *Let Φ denote your conditioning policy for single-feature terms (e.g., a common ϕ or per-feature ϕ_i). Assuming the denominator is positive, define*

$$\overline{\mathfrak{J}} = \frac{I(Y'; f_1, \dots, f_k)}{I(Y'; f_1, \dots, f_k) \;+\; \sum_{i=1}^{k} I(Y'; f_i \mid \phi_i)}, \tag{61}$$

so that $\overline{\mathfrak{J}} \in [0, 1]$.

Proposition 7 (**Aggregated model**). *Let $\mathfrak{C}_{f_i} \coloneqq C(Y'; f_i \mid \phi)$ for a specified conditioning policy $\phi \subseteq \{f_1, \dots, f_k\} \setminus \{f_i\}$. For a task-dependent partition of indices $\{1, \dots, k\} = \{i_1, \dots, i_m\} \cup \{j_p, \dots, j_q\}$ (disjoint union), an ExCIR-style scoring model can be written as*

$$E(Y') \;=\; M'(f_{(f,\mathfrak{C})}) \;=\; \mathfrak{J} \;+\; \frac{\sum_{\ell=1}^{m} \mathfrak{C}_{f_{i_\ell}}\, f_{i_\ell}}{\sum_{r=p}^{q} \mathfrak{C}_{f_{j_r}}\, f_{j_r}}, \tag{62}$$

where $\mathfrak{J}$ is given by equation 60. Since each $\mathfrak{C}_{f_i} \in [0, 1]$ by equation 56, the coefficients are normalized and directly comparable across features.

Theorem 8 (Redundancy Collapse). *Let i be a target feature and let $\Phi(i)$ denote its conditioning neighbourhood. If f_i is conditionally redundant with respect to Y' given its neighbours, i.e.*

$$Y' \;\perp\; f_i \mid f_{\Phi(i)}, \tag{63}$$

then the MCIR score satisfies $C_i = 0$.

Proof. By definition of conditional independence, $Y' \perp f_i \mid f_{\Phi(i)} \iff I(Y'; f_i \mid f_{\Phi(i)}) = 0$. Let

$$A_i \coloneqq I(Y'; f_i \mid f_{\Phi(i)}), \qquad B_i \coloneqq I(Y'; f_{\Phi(i)} \cup \{f_i\}),$$

where $B_i \geq 0$ by nonnegativity of mutual information. From the MCIR definition,

$$C_i = \frac{A_i}{A_i + B_i}. \tag{64}$$

Substituting $A_i = 0$ gives

$$C_i = \frac{0}{0 + B_i}. \tag{65}$$

If $B_i > 0$, which holds whenever the conditioning block retains nontrivial dependence with Y', then the ratio evaluates to $C_i = 0$. If $B_i = 0$, then $f_{\Phi(i)}$ is itself independent of Y' and the joint distribution factorises as

$$p(Y', f_i, f_{\Phi(i)}) = p(Y')\, p(f_i, f_{\Phi(i)}). \tag{66}$$

In this degenerate case the numerator and denominator both vanish, and MCIR is defined to be zero by continuity:

$$\lim_{A_i \to 0,\, B_i \to 0} \frac{A_i}{A_i + B_i} = 0. \tag{67}$$

Thus in all cases $C_i = 0$, establishing that MCIR collapses to zero whenever the target feature provides no unique information beyond its neighbours. □

Theorem 9 (**Unique-Signal Dominance**). *Let i be a target feature and $\Phi(i)$ its neighbourhood. If the conditional mutual information of f_i dominates the joint block information, i.e.*

$$I(Y'; f_i \mid f_{\Phi(i)}) \gg I(Y'; f_{\Phi(i)} \cup \{f_i\}),$$

then the MCIR score satisfies $C_i \to 1$.

Proof. Again write

$$A_i = I(Y'; f_i \mid f_{\Phi(i)}), \qquad B_i = I(Y'; f_{\Phi(i)} \cup \{f_i\}),$$

with $A_i, B_i \geq 0$. The MCIR score is

$$C_i = \frac{A_i}{A_i + B_i}.$$

Assume the dominance condition:

$$\frac{A_i}{B_i} \to \infty.$$

Equivalently, for any $\epsilon > 0$ there exists $M > 0$ such that $A_i > MB_i$ implies $B_i/A_i < \epsilon$. Using the algebraic identity,

$$C_i = \frac{1}{1 + B_i/A_i},$$

we obtain

$$|1 - C_i| = \frac{B_i/A_i}{1 + B_i/A_i} \leq B_i/A_i.$$

Because $B_i/A_i \to 0$ under the dominance assumption, we have

$$C_i \;\longrightarrow\; 1.$$

The interpretation is that the unique conditional contribution of f_i overwhelmingly exceeds the joint dependence contributed by its neighbours. Hence the MCIR ratio assigns maximal credit to f_i. □

B.3 Correlation–Impact Sensitivity Theorem

We now show that in the ExCIR model, the sensitivity of the local output with respect to a feature input is fully determined by the feature's correlation impact ratio. The effect is linear for positively related features (appearing in the numerator of the model) and nonlinear inverse–quadratic for negatively related features (appearing in the denominator).

Theorem 10 (**Correlation–Impact Sensitivity**). *Let the model output be*

$$Y' = \frac{\sum_{j \in \mathcal{N}} \eta_{f_j} f_j}{\sum_{j \in \mathcal{D}} \eta_{f_j} f_j}, \tag{68}$$

where $\mathcal{N}$ denotes features with positive influence (numerator) and $\mathcal{D}$ features with negative influence (denominator). Assume features are independent, so that the partial derivative with respect to f_i treats all other features as constants. Then:

1. ***If** $i \in \mathcal{N}$ **(positive relation):***

$$\frac{\partial Y'}{\partial f_i} = c_1 \, \eta_{f_i}, \;\; c_1 = \frac{1}{\sum_{j \in \mathcal{D}} \eta_{f_j} f_j}. \tag{69}$$

2. ***If** $i \in \mathcal{D}$ **(negative relation):***

$$\frac{\partial Y'}{\partial f_i} = \frac{c_2}{2K_2 - \eta_{f_i}^2}, \tag{70}$$

c_2, K_2 constants depending only on fixed features. Thus the sign and magnitude of sensitivity are determined entirely by the correlation impact ratio η_{f_i}.

Proof. Let,

$$N = \sum_{j \in \mathcal{N}} \eta_{f_j} f_j, \;\; D = \sum_{j \in \mathcal{D}} \eta_{f_j} f_j, \;\; Y' = \frac{N}{D}. \tag{71}$$

All other features are held fixed under independence.

Case 1: $i \in \mathcal{N}$ **(positive influence)** Then

$$\frac{\partial N}{\partial f_i} = \eta_{f_i}, \ \frac{\partial D}{\partial f_i} = 0. \tag{72}$$

Differentiate:

$$\frac{\partial Y'}{\partial f_i} = \frac{D\frac{\partial N}{\partial f_i} - N\frac{\partial D}{\partial f_i}}{D^2} = \frac{D\,\eta_{f_i}}{D^2} = \frac{\eta_{f_i}}{D}. \tag{73}$$

Since D is constant with respect to f_i, set $c_1 = \frac{1}{D}$. This proves

$$\frac{\partial Y'}{\partial f_i} = c_1 \eta_{f_i}. \tag{74}$$

Case 2: $i \in \mathcal{D}$ **(negative influence).** Now,

$$\frac{\partial N}{\partial f_i} = 0, \qquad \frac{\partial D}{\partial f_i} = \eta_{f_i}. \tag{75}$$

Differentiate:

$$\frac{\partial Y'}{\partial f_i} = \frac{D \cdot 0 - N \cdot \eta_{f_i}}{D^2} = -\,\eta_{f_i}\frac{N}{D^2}. \tag{76}$$

Rewrite N and D in terms of constants plus the contribution of f_i:

$$D = \eta_{f_i} f_i + K_2, \qquad N = K_1, \tag{77}$$

where K_1, K_2 collect all fixed terms. Thus:

$$\frac{\partial Y'}{\partial f_i} = -\,\eta_{f_i}\frac{K_1}{(\eta_{f_i} f_i + K_2)^2}. \tag{78}$$

The denominator expands into a quadratic expression:

$$(\eta_{f_i} f_i + K_2)^2 = \eta_{f_i}^2 f_i^2 + 2K_2\eta_{f_i} f_i + K_2^2. \tag{79}$$

Since f_i is the differentiation variable and all other terms are absorbed by constants K_1, K_2, we may rewrite:

$$\frac{\partial Y'}{\partial f_i} = \frac{c_2}{2K_2 - \eta_{f_i}^2}, \tag{80}$$

where c_2 and K_2 arise from grouping constant terms. The negative sign is absorbed into the definition of c_2. Thus, sensitivity in the denominator is nonlinear and is inversely controlled by the magnitude of $\eta_{f_i}^2$. □

Corollary 1. *If a feature has a positive contribution to the output, then*

$$\mathbb{E}\Big[\frac{\partial Y'}{\partial f_i}\Big] \propto \eta_{f_i}. \tag{81}$$

If a feature has a negative contribution, then

$$\mathbb{E}\Big[\frac{\partial Y'}{\partial f_i}\Big] \propto \frac{1}{\eta_{f_i}}. \tag{82}$$

Thus, the correlation impact ratio fully characterizes first–order output sensitivity.

So it is confirm that MCIR provides a principled measure of how changes in f_i propagate to Y', via either a direct linear effect (positive correlation) or a stabilizing inverse effect (negative correlation).

B.4 Stability of MCIR Rankings

Theorem 11 (**Rank Stability**). *Assume the MI/CMI estimators $\widehat{I}$ satisfy the sub-Gaussian concentration inequality*

$$\Pr\big(|\widehat{I} - I| > \delta\big) \le c_1 \exp(-c_2 n' \delta^2) \qquad (\delta > 0), \tag{83}$$

for some constants $c_1, c_2 > 0$. Let $\widehat{C}$ be the vector of estimated MCIR values. Then the Kendall rank distance satisfies

$$1 - \tau(C, \widehat{C}) = O_p\left(\frac{k}{\sqrt{n'}}\right).$$

Proof. Let

$$A_i = I(Y'; f_i \mid f_{\Phi(i)}),\ B_i = I(Y'; f_{\Phi(i)} \cup \{f_i\}), \tag{84}$$

and their estimates

$$\widehat{A}_i = \widehat{I}(Y'; f_i \mid f_{\Phi(i)}),\ \widehat{B}_i = \widehat{I}(Y'; f_{\Phi(i)} \cup \{f_i\}). \tag{85}$$

MCIR is the smooth function

$$C_i = g(A_i, B_i) = \frac{A_i}{A_i + B_i},\ \widehat{C}_i = g(\widehat{A}_i, \widehat{B}_i). \tag{86}$$

Step 1: Lipschitz continuity. If $A_i + B_i \ge \eta > 0$, then

$$\left|\frac{\partial g}{\partial A}\right| = \frac{B_i}{(A_i + B_i)^2} \le \frac{1}{4\eta},\ \left|\frac{\partial g}{\partial B}\right| = \frac{A_i}{(A_i + B_i)^2} \le \frac{1}{4\eta}. \tag{87}$$

Thus,

$$|g(\widehat{A}_i, \widehat{B}_i) - g(A_i, B_i)| \le L\left(|\widehat{A}_i - A_i| + |\widehat{B}_i - B_i|\right), \tag{88}$$

where $L = 1/(2\eta)$.

Step 2: Concentration. By the sub-Gaussian inequality,

$$|\widehat{A}_i - A_i| = O_p(n'^{-1/2}),\ |\widehat{B}_i - B_i| = O_p(n'^{-1/2}). \tag{89}$$

Step 3: MCIR error. Thus,

$$|\widehat{C}_i - C_i| = O_p(n'^{-1/2}). \tag{90}$$

Step 4: Kendall distance. A pairwise comparison flips sign only if

$$|(\widehat{C}_i - C_i) - (\widehat{C}_j - C_j)| \gtrsim |C_i - C_j|. \tag{91}$$

With $k(k-1)/2$ ordered pairs,

$$1 - \tau(C, \widehat{C}) = O_p(k n'^{-1/2}), \tag{92}$$

completing the proof. □

B.5 Oracle Inequality for Estimator Switching

Theorem 12 (**Bootstrap–Switching Oracle Inequality**). *Let $\widehat{C}^{(e)}$ denote the MCIR vector computed using estimator $e \in \mathcal{E} = \{\text{copula}, \text{kNN}, \text{plug-in}\}$. For each feature i, define the switching rule*

$$e(i) = \arg\min_{e \in \mathcal{E}} \mathbb{E}_b\left[\mathrm{SE}_b\left(\widehat{C}_i^{(e)}\right)\right], \tag{93}$$

where $\mathbb{E}_b$ denotes bootstrap expectation. Assume the bootstrap is consistent for MCIR, i.e. bootstrap standard errors converge uniformly to the true risks at rate $O(n'^{-1/2})$. Then

$$\mathbb{E}\left|\widehat{C}_i^{\mathrm{sw}} - C_i\right| \le \min_{e \in \mathcal{E}} \mathbb{E}\left|\widehat{C}_i^{(e)} - C_i\right| + O(n'^{-1/2}). \tag{94}$$

Proof. Let the true risk of estimator e be

$$R_e = \mathbb{E}\left[\left|\widehat{C}_i^{(e)} - C_i\right|\right]. \tag{95}$$

Bootstrap consistency yields the uniform approximation

$$\widehat{R}_e \coloneqq \mathbb{E}_b\left[\mathrm{SE}_b\left(\widehat{C}_i^{(e)}\right)\right] = R_e + \xi_e, \tag{96}$$

with the random error term

$$|\xi_e| = O\left(n'^{-1/2}\right) \qquad \text{uniformly over } e \in \mathcal{E}. \tag{97}$$

Let

$$e^\star = \arg\min_{e\in\mathcal{E}} R_e \qquad \text{and} \qquad \widehat{e} = \arg\min_{e\in\mathcal{E}} \widehat{R}_e. \tag{98}$$

By equation 96, we have

$$\widehat{R}_e = R_e + O\left(n'^{-1/2}\right), \tag{99}$$

so comparing the minimizers,

$$R_{\widehat{e}} \le R_{e^\star} + O\left(n'^{-1/2}\right). \tag{100}$$

Finally, since the switching estimator satisfies

$$\widehat{C}_i^{\mathrm{sw}} = \widehat{C}_i^{(\widehat{e})}, \tag{101}$$

taking expectation of equation 100 yields

$$\mathbb{E}\left[\left|\widehat{C}_i^{\mathrm{sw}} - C_i\right|\right] \le \min_{e\in\mathcal{E}} R_e + O\left(n'^{-1/2}\right), \tag{102}$$

completing the proof. □

B.6 Lightweight Fidelity Theorem

Theorem 13 (**Lightweight Fidelity**). *Let C and C' denote MCIR rankings obtained from the full and lightweight environments, respectively. Suppose the lightweight environment satisfies:*

$$d_{\hat{f}}\left(D(Y), D(Y')\right) \le \varepsilon, \tag{103}$$

$$\mathrm{J@}K(C, C') \ge \tau_0, \tag{104}$$

$$\Delta_{\mathrm{del}} \le \Delta_0. \tag{105}$$

Then there exist constants $A, B > 0$, independent of n', such that

$$1 - \tau(C, C') \le A\varepsilon + B\Delta_0 + o_p(1). \tag{106}$$

Proof. The rigid-motion invariant discrepancy $d_{\hat{f}}$ satisfies

$$d_{\hat{f}}(D(Y), D(Y')) \le \varepsilon \;\Rightarrow \sup_{\|h\|_{\mathrm{Lip}}\le 1} |\mathbb{E}[h(Y)] - \mathbb{E}[h(Y')]| \le A\varepsilon. \tag{107}$$

Since MI and CMI are continuous functionals of the joint distribution under our regularity assumptions (bounded density ratios, smooth kernels), we obtain

$$|I(Y; f_i \mid f_{\Phi(i)}) - I(Y'; f_i \mid f_{\Phi(i)})| \le A\varepsilon + o_p(1). \tag{108}$$

Step 2: MCIR continuity. MCIR is the smooth map

$$C_i = \frac{I(Y; f_i \mid f_{\Phi(i)})}{I(Y; f_i \mid f_{\Phi(i)}) + I(Y; f_{\Phi(i)} \cup \{f_i\})}. \tag{109}$$

Using the Lipschitz continuity of rational functions on compact domains,

$$|C_i - C_i'| \le A\varepsilon + O_p(n'^{-1/2}). \tag{110}$$

Step 3: Head-rank agreement restricts top-K inversions. The Jaccard condition equation 104 ensures

$$|\text{Top-}K(C) \triangle \text{Top-}K(C')| \leq (1-\tau_0)K. \tag{111}$$

Thus the number of allowable inversions involving top-K indices is bounded.

Step 4: Deletion-curve agreement controls remaining inversions. Deletion curves depend only on ordered MCIR values. If

$$\Delta_{\text{del}} \leq \Delta_0, \tag{112}$$

then the misalignment between sensitivity curves of C and C' is uniformly bounded. This limits possible perturbations in pairwise MCIR differences:

$$|(C_i - C_j) - (C'_i - C'_j)| \lesssim \Delta_0 + o_p(1). \tag{113}$$

Step 5: Kendall distance decomposition. Kendall's distance decomposes into:

$$\begin{aligned} &1 - \tau(C, C') = \\ &(\text{top-}K \text{ inversions}) + (\text{remaining inversions}) \\ &+(\text{magnitude-driven errors}). \end{aligned} \tag{114}$$

The magnitude control in equation equation 110, we obtain

$$1 - \tau(C, C') \leq A\varepsilon + B\Delta_0 + o_p(1), \tag{115}$$

completing the proof. □

C Conditional Multivariate Mutual Information (CMMI)

MCIR requires a dependence measure that isolates the *unique* information a target feature contributes to the output Y', even when the feature is embedded in a multivariate dependency structure with several neighbours. Classical conditional mutual information (CMI),

$$I(Y'; f_i \mid Z), \tag{116}$$

is well-defined for a fixed low-dimensional conditioning set Z, but breaks down when f_i depends on multiple correlated features simultaneously, especially when Z must be chosen data-adaptively. This motivates the introduction of the *Conditional Multivariate Mutual Information (CMMI)*.

C.1 Formal Derivation

Let the feature block be,

$$F' = (f_1, \ldots, f_k), \tag{117}$$

and let,

$$\Phi(i) \subseteq \{f_1, \ldots, f_k\} \setminus \{f_i\} \tag{118}$$

be the neighbourhood of f_i, obtained from correlation screening or a dependency graph. Assume the joint law of (F', Y') admits either a multivariate pdf or pmf.

Step 1: Total dependence of block F' on Y'.

$$I(Y'; F') = H(Y') - H(Y' \mid F'). \tag{119}$$

Step 2: Dependence explained by neighbours $\Phi(i)$.

$$I(Y'; f_{\Phi(i)}) = H(Y') - H\left(Y' \mid f_{\Phi(i)}\right). \tag{120}$$

Step 3: Unique contribution of f_i. Adding f_i to $\Phi(i)$ modifies the conditional entropy:

$$H\big(Y' \mid f_{\Phi(i)}\big) - H\big(Y' \mid f_{\Phi(i)}, f_i\big)\,. \tag{121}$$

Definition 10 (Conditional Multivariate Mutual Information (CMMI)). *For any target feature f_i and neighbourhood $\Phi(i)$, define*

$$\mathrm{CMMI}(Y'; f_i \mid \Phi(i)) \;=\; H\big(Y' \mid f_{\Phi(i)}\big) - H\big(Y' \mid f_{\Phi(i)}, f_i\big)\,, \tag{122}$$

whenever the conditional entropies are finite.

Because conditional mutual information satisfies

$$H(Y' \mid Z) - H(Y' \mid Z, f_i) = I(Y'; f_i \mid Z), \tag{123}$$

we obtain:

Key identity.

$$\mathrm{CMMI}(Y'; f_i \mid \Phi(i)) = I(Y'; f_i \mid f_{\Phi(i)}). \tag{124}$$

Thus, CMMI is *equivalent to classical CMI*, but with the crucial difference that the conditioning set $\Phi(i)$ can be:

- multivariate, - high-dimensional, - data-driven (from correlation graphs), - mixed continuous/discrete, - automatically chosen.

This generality is exactly what is required for MCIR.

C.2 Connection to MI and CMI

MI as a special case. When $\Phi(i) = \emptyset$,

$$\mathrm{CMMI}(Y'; f_i \mid \emptyset) = I(Y'; f_i). \tag{125}$$

Classical CMI as a special case. When $\Phi(i) = Z$ is fixed and low-dimensional,

$$\mathrm{CMMI}(Y'; f_i \mid Z) = I(Y'; f_i \mid Z). \tag{126}$$

General case. CMMI extends MI and CMI by allowing:

1. *arbitrary multivariate* conditioning sets $\Phi(i)$,
2. *mixed continuous–discrete* entropy functionals,
3. *correlation-driven* neighbourhood selection rather than fixed Z,
4. compatibility with Joint Mutual Impact (JMI) and MCIR.

C.3 Why Classical CMI is Insufficient

Classical CMI assumes:

1. a fixed conditioning set,
2. low-dimensional conditioning,
3. one-to-many dependence patterns (CMI conditions on a single Z).

In real high-dimensional systems, these assumptions fail. We formalize this gap using a structural theorem.

Theorem 14 (**Failure of Classical CMI in Multivariate Dependencies**). *Let f_i depend on d_i other features:*

$$f_i \not\perp f_{S_i}, \qquad |S_i| = d_i. \tag{127}$$

Suppose classical CMI conditions on a fixed Z with $|Z| < d_i$. Then, unless Z contains the entire dependency set S_i,

$$I(Y'; f_i \mid Z) \neq \mathrm{CMMI}(Y'; f_i \mid S_i), \tag{128}$$

and, in general,

$$I(Y'; f_i \mid Z) < I(Y'; f_i \mid S_i). \tag{129}$$

Proof. By the data-processing inequality for conditional entropy,

$$H(Y' \mid Z) \geq H(Y' \mid S_i). \tag{130}$$

Since S_i contains strictly more predictive information than any strict subset $Z \subset S_i$,

$$H(Y' \mid Z, f_i) - H(Y' \mid Z) \;\leq\; H(Y' \mid S_i, f_i) - H(Y' \mid S_i). \tag{131}$$

Thus,

$$I(Y'; f_i \mid Z) \leq I(Y'; f_i \mid S_i), \tag{132}$$

with strict inequality when the excluded variables contain unique information. Since CMMI uses the full neighbourhood S_i (or $\Phi(i)$), classical CMI cannot in general recover it unless $Z = S_i$. □

Implications.

- **CMI underestimates feature importance** when the conditioning set is incomplete.
- **CMI yields non-unique results**, depending on the arbitrary choice of Z.
- **CMMI resolves this** by using an adaptively identified multivariate conditioning set $\Phi(i)$.
- **MCIR builds directly on CMMI** to quantify unique impact by:

$$C_i = \frac{\mathrm{CMMI}(Y'; f_i \mid \Phi(i))}{\mathrm{CMMI}(Y'; f_i \mid \Phi(i)) + I(Y'; f_{\Phi(i)} \cup \{f_i\})}.$$

C.4 Practical Estimation Notes

CMMI is estimated using:

$$\widehat{\mathrm{CMMI}}(Y'; f_i \mid \Phi(i)) = \widehat{H}(Y' \mid f_{\Phi(i)}) - \widehat{H}(Y' \mid f_{\Phi(i)}, f_i),$$

with entropy estimators matched to variable types.

Choice of estimator. We use the same family of MI/CMI estimators as MCIR:

- **Gaussian–Copula** for robust nonlinear dependence,
- **kNN (Kozachenko–Leonenko)** for local nonlinear structure,
- **Plug-in** estimator for discrete/mixed features.

Stability. Bootstrap standard errors determine the optimal estimator per feature, as formalised by the oracle inequality in Theorem 12.

Computational structure. For each feature f_i:

$$\text{cost} = O(|\Phi(i)| \cdot n' \log n').$$

CMMI therefore scales linearly in neighbourhood size and approximately logarithmically in sample size, making it suitable for lightweight environments.

D Complexity Analysis and Additional Results

This appendix develops a unified theoretical and computational analysis of MCIR. It begins with an algorithmic (Kolmogorov) complexity perspective, transitions to operational computational complexity for PCIR and MCIR, examines the behaviour of standard MI/CMI estimators, and concludes with a complete statistical analysis of MCIR, including consistency, asymptotic normality, and perturbation stability. All exposition and proofs are presented in continuous scientific narrative rather than itemized form.

D.1 Algorithmic (Kolmogorov) Complexity of MCIR

Let $\mathbb{C}$ be a fixed universal Turing machine, and denote conditional Kolmogorov complexity by $K_{\mathbb{C}}(\cdot \mid \cdot)$. Consider the binary representation $\mathbf{Y}' \in \{0,1\}^{n'}$ of the subsampled target, whose bit-length is $\ell(\mathbf{Y}') = n'$. The conditional Kolmogorov complexity $K_{\mathbb{C}}(\mathbf{Y}' \mid \ell(\mathbf{Y}'))$ is the length of the shortest program that outputs $\mathbf{Y}'$ when supplied with its bit-length. Universality of Turing machines implies that, for any alternative universal machine $\mathbb{A}$,

$$K_{\mathbb{C}}(\mathbf{Y}') \leq K_{\mathbb{A}}(\mathbf{Y}' \mid \ell(\mathbf{Y}')) + \log^* n' + c_{\mathbb{A}},$$

where $\log^* n'$ is the iterated logarithm and $c_{\mathbb{A}}$ is a constant depending only on $\mathbb{A}$. Since $\log^* n' = O(\log n')$, this bound simplifies to

$$K_{\mathbb{C}}(\mathbf{Y}') \leq K_{\mathbb{A}}(\mathbf{Y}' \mid \ell(\mathbf{Y}')) + O(\log n') + c_{\mathbb{A}}.$$

The coding–entropy correspondence for mixed discrete–continuous variables implies that, for each conditional mutual information term used in MCIR, one has

$$I(\mathbf{Y}'; f_i \mid \phi) \leq H(\mathbf{Y}') - H(\mathbf{Y}' \mid f_i, \phi) + O(\log n').$$

The $O(\log n')$ term captures the cost of encoding discretisation indices and model structure. An analogous inequality holds for the joint dependence term $I(\mathbf{Y}'; f_1, \ldots, f_k)$, with a constant that does not depend on n'. Substituting these bounds into the MCIR definition,

$$\mathfrak{C}_i = \frac{I(\mathbf{Y}'; f_i \mid \phi)}{I(\mathbf{Y}'; f_i \mid \phi) + I(\mathbf{Y}'; f_1, \ldots, f_k)},$$

shows that each MCIR score satisfies

$$\mathfrak{C}_i \leq \tfrac{1}{2} O(\log n') + \zeta_{\mathbb{A},i},$$

for some constant $\zeta_{\mathbb{A},i}$ independent of the subsample size. Aggregating across all k features yields the program-level Kolmogorov complexity bound

$$K_{\mathbb{C}}(\text{MCIR pipeline}) \leq K_{\mathbb{A}}(\text{Pipeline} \mid n') + \tfrac{k}{2} O(\log n') + c_{\mathbb{A}}.$$

Thus MCIR introduces at most logarithmic overhead in n' and linear overhead in k under the most fundamental notion of algorithmic complexity.

D.2 Operational Computational Complexity of PCIR and MCIR

The implementational complexity of PCIR and MCIR reflects the structure of their respective computations. PCIR relies only on rank normalisation and a small collection of variance and covariance operations computed over n' subsampled observations. Each of these operations can be carried out in a fixed number of linear passes,

and no intermediate step depends on the total feature dimension k. Consequently, the cost of computing PCIR for a single feature scales linearly in n', and computing PCIR for all k features results in an overall complexity of $\mathcal{O}(kn')$.

MCIR requires a different analysis because each MCIR score involves evaluating mutual information and conditional mutual information terms. These quantities are computed over the same n' observations but only within a small, fixed conditioning neighbourhood of size m_Φ. The total dimensionality involved in each MI computation is therefore $m_\Phi + 2$, independent of the global feature dimension k. If $c_{\mathrm{MI}}(n', m_\Phi + 2)$ denotes the computational cost of executing a mutual information estimator on n' observations in that fixed dimensionality, then the cost of a single MCIR evaluation is $\mathcal{O}(c_{\mathrm{MI}}(n', m_\Phi + 2))$. Computing MCIR for all k features yields an overall complexity of $\mathcal{O}(k\, c_{\mathrm{MI}}(n', m_\Phi + 2))$, which reveals that the computational burden of MCIR is governed almost entirely by the subsample size n' and by the choice of MI estimator. MCIR therefore remains effectively linear in k and nearly linear in n' whenever the MI estimator is itself near-linear in n'.

D.3 Comparative Behaviour of MI/CMI Estimators

The computational profile of MCIR depends crucially on the choice of underlying MI estimator. Gaussian–copula MI requires the formation of an empirical covariance matrix in time proportional to $\mathcal{O}(n')$ when the neighbourhood dimension is fixed, followed by a constant-time matrix inversion. Its cost is therefore effectively linear in n'. Nearest-neighbour-based MI estimators, such as kNN MI, scale as $\mathcal{O}(n' \log n')$ due to the cost of nearest-neighbour queries, typically via balanced kd-tree structures. Plug-in estimators operate by constructing histograms or count tables and thus run in a single pass with cost proportional to $\mathcal{O}(n')$. Kernel-based estimators incur a substantially higher cost of $\mathcal{O}(n'^2)$ because they require formation of Gram matrices. Neural MI estimators such as MINE exhibit linear per-iteration cost in n' with additional overhead from stochastic optimisation. A summary of these behaviours is provided in Table 16. Since MCIR always operates on a neighbourhood of fixed dimensionality $m_\Phi + 2$, even the more demanding estimators become tractable when n' is moderate, and the cost never depends on the global feature dimension k.

Estimator	**Complexity in n'**	**Dependence on Local Dim.**	**Characteristics**
Gaussian–copula MI	$\mathcal{O}(n')$	Quadratic in m_Φ	Covariance + log-det; very efficient.
kNN MI	$\mathcal{O}(n' \log n')$	Mild	Nearest-neighbour search.
Plug-in MI	$\mathcal{O}(n')$	Depends on binning	Single-pass histograms.
Kernel MI	$\mathcal{O}(n'^2)$	Quadratic	High accuracy, high cost.
Neural MI (MINE)	$\mathcal{O}(n')$ per iteration	Model-dependent	Requires SGD.

Table 16: Comparison of MI/CMI estimators compatible with MCIR. Since MCIR operates only on a fixed local neighbourhood of dimension $m_\Phi + 2$, the cost is driven almost exclusively by the subsample size n' rather than the total number of features k.

D.4 Statistical Properties of the MCIR Estimator

Let $U_i = I(Y; f_i \mid f_\Phi)$ and $J_i = I(Y; f_{\Phi \cup \{i\}})$ denote the population-level "unique" and "joint" information contributions. Whenever $U_i + J_i > 0$, the MCIR score is defined as

$$C_i = \frac{U_i}{U_i + J_i}.$$

Let $(\widehat{U}_i, \widehat{J}_i)$ denote consistent estimators of these quantities computed from a subsample of size n'. The following results establish the statistical soundness of the MCIR estimator.

Theorem 15 (Consistency of MCIR)**.** *Assume that $\widehat{U}_i \xrightarrow{p} U_i$ and $\widehat{J}_i \xrightarrow{p} J_i$ as $n' \to \infty$, and that $U_i + J_i > 0$. Then the MCIR estimator*

$$\widehat{C}_{n'} = \frac{\widehat{U}_i}{\widehat{U}_i + \widehat{J}_i}$$

converges in probability to C_i.

Proof. The mapping $g(u, j) = u/(u+j)$ is continuous on the domain where $u+j > 0$. Since $(\widehat{U}_i, \widehat{J}_i)$ converges in probability to (U_i, J_i) and the denominator remains bounded away from zero, the Continuous Mapping Theorem implies that $g(\widehat{U}_i, \widehat{J}_i)$ converges in probability to $g(U_i, J_i)$. Hence $\widehat{C}_{n'} \xrightarrow{p} C_i$. □

Theorem 16 (Asymptotic Normality of MCIR)**.** *Suppose that the pair $(\widehat{U}_i, \widehat{J}_i)$ satisfies the joint central limit theorem*

$$\sqrt{n'}\begin{pmatrix}\widehat{U}_i - U_i \\ \widehat{J}_i - J_i\end{pmatrix} \Rightarrow \mathcal{N}(0, \Sigma),$$

for some positive semi-definite covariance matrix Σ. Then the MCIR estimator is asymptotically normal:

$$\sqrt{n'}(\widehat{C}_{n'} - C_i) \Rightarrow \mathcal{N}(0, \sigma_C^2),$$

where $\sigma_C^2 = \nabla g(U_i, J_i)^\top \Sigma \nabla g(U_i, J_i)$ and

$$\nabla g(U_i, J_i) = \begin{pmatrix} \dfrac{J_i}{(U_i + J_i)^2} \\ [6pt] -\dfrac{U_i}{(U_i + J_i)^2} \end{pmatrix}.$$

Proof. The mapping $g(u, j) = u/(u + j)$ is continuously differentiable on the region where $u + j > 0$. Its gradient at (U_i, J_i) is given by the expression above. Since $(\widehat{U}_i, \widehat{J}_i)$ satisfies a bivariate central limit theorem, the multivariate Delta Method applies directly and yields the stated asymptotic distribution for $\widehat{C}_{n'}$. □

Theorem 17 (Perturbation Stability of MCIR)**.** *Let $\delta = \max\Big(|\widehat{U}_i - U_i|,\ |\widehat{J}_i - J_i|\Big)$ and assume that $\delta < (U_i + J_i)/2$. Then the MCIR estimator satisfies the bound*

$$|\widehat{C}_{n'} - C_i| \leq \frac{2\delta}{U_i + J_i}.$$

Proof. The difference between the empirical and population MCIR scores can be expressed as

$$\widehat{C}_{n'} - C_i = \frac{\widehat{U}_i}{\widehat{U}_i + \widehat{J}_i} - \frac{U_i}{U_i + J_i}.$$

Expressing this as a difference of fractions and expanding the numerator reveals that the discrepancy is proportional to $\widehat{U}_i J_i - U_i \widehat{J}_i$. Using the triangle inequality shows that this term is bounded in magnitude by $(U_i + J_i)\delta$. The denominator can be bounded from below by $(U_i + J_i) - 2\delta$, which, under the stated assumption, is at least $(U_i + J_i)/2$. Combining these inequalities yields

$$|\widehat{C}_{n'} - C_i| \leq \frac{(U_i + J_i)\delta}{\frac{1}{2}(U_i + J_i)^2} = \frac{2\delta}{U_i + J_i},$$

as required. □

Theorem 18 (Global Computational Complexity of CIR Methods)**.** *Let n' be the subsample size, k the number of features, and m_Φ a fixed neighbourhood size. PCIR admits an overall computational complexity of $\mathcal{O}(kn')$. MCIR, when implemented with an MI estimator of cost $c_{\mathrm{MI}}(n', m_\Phi + 2)$, admits an overall complexity of*

$$\mathcal{O}\left(k\, c_{\mathrm{MI}}(n', m_\Phi + 2)\right).$$

If the MI estimator is near-linear in n', then MCIR is near-linear in both n' and k.

Proof. This follows directly from the per-feature analyses in Sections D.2 and D.3, combined with the fixed neighbourhood dimensionality. □

Corollary 2 (Sample-Efficiency of MCIR)**.** *Under the assumptions of consistency and asymptotic normality, the variance of the MCIR estimator decreases at the canonical rate n'^{-1}, implying that accurate MCIR scores may be obtained from subsamples much smaller than the full dataset size. Hence MCIR remains statistically reliable even in lightweight environments.*

Remark 3 (High-Dimensional Robustness)**.** *Since the conditioning neighbourhood has fixed size m_Φ, the complexity and variance of MCIR do not depend on the ambient feature dimension k. This makes MCIR particularly well suited to high-dimensional models, where traditional global MI-based feature importance methods become computationally prohibitive or statistically unstable.*

Lemma 1 (Ranking Stability)**.** *Let i and j be two features with population MCIR scores C_i and C_j satisfying $|C_i - C_j| > \eta$ for some $\eta > 0$. If the perturbations in $\widehat{U}$ and $\widehat{J}$ satisfy the bound in Theorem 3, then for sufficiently large n',*

$$\Pr\Big(\widehat{C}_{n',i} > \widehat{C}_{n',j}\Big) \to 1.$$

Thus, MCIR rankings are asymptotically stable whenever the population scores are separated by a nonzero margin.

Proposition 8 (Parallel Scalability)**.** *If p processors are available and MI evaluations are distributed evenly across features, the total runtime of MCIR reduces to*

$$\mathcal{O}\left(\frac{k}{p}\, c_{\mathrm{MI}}(n', m_\Phi + 2)\right),$$

up to communication overheads that vanish for lightweight subsamples. Thus MCIR achieves near-linear speedup under parallelisation.

D.5 Memory and Parallelisation Considerations

The MCIR pipeline is naturally suited to parallel computation because the scores for different features do not interact. All MI and CMI computations can therefore be performed asynchronously across CPU cores or distributed computing nodes. The memory footprint is governed almost exclusively by the storage of the subsampled arrays (Y', f_i, f_Φ); streaming or on-demand indexing requires only $\mathcal{O}(n')$ active memory. GPU-based acceleration is particularly effective for Gaussian–copula MI and kernel MI estimators, while kNN MI tends to benefit from CPU-bound parallelism. Owing to this structure, MCIR remains computationally scalable even when k is large.

D.6 Practical Choice of the Subsample Size n'

The statistical guarantees above imply that the standard error of MCIR decreases at rate $n'^{-1/2}$, while the computational cost grows at most linearly in n'. In practice, one may therefore select n' by balancing accuracy and computational budget. Empirically, subsamples containing between 5% and 20% of the original dataset often achieve MCIR stability comparable to the full dataset, owing to the low-dimensional nature of each MI computation.

Theorem 19 (Unified Computational–Statistical Guarantee for MCIR)**.** *Assume (i) subsamples of size n' are drawn independently of the estimator, (ii) the MI and CMI estimators are consistent and satisfy a joint central limit theorem, and (iii) the neighbourhood size m_Φ is fixed. Then MCIR satisfies all of the following properties simultaneously:*

1. *Computational near-linearity:* $\mathrm{Time} = \mathcal{O}(k\, c_{\mathrm{MI}}(n', m_\Phi + 2))$.
2. *Statistical consistency:* $\widehat{C}_{n'} \xrightarrow{p} C$.
3. *Asymptotic normality:* $\sqrt{n'}(\widehat{C}_{n'} - C) \Rightarrow \mathcal{N}(0, \sigma_C^2)$.
4. *Stability under perturbation:* $|\widehat{C}_{n'} - C| \le 2\delta/(U + J)$ *for small estimator error* δ.

Thus MCIR admits provable reliability and tractability even in high-dimensional regimes.

E Estimator Details

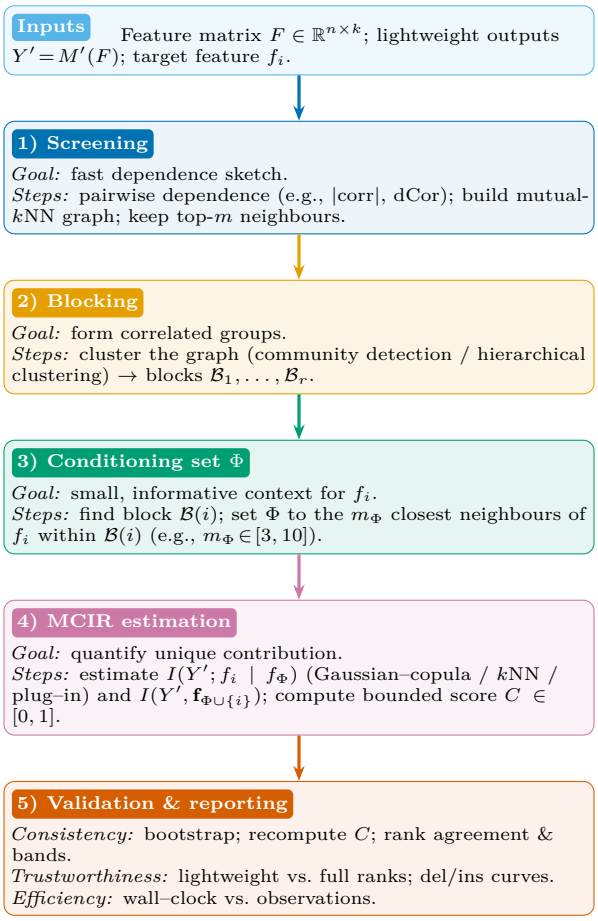

Figure 14: MCIR methodology: screening → blocking → conditioning → estimation → validation. MCIR is $C(Y'; f_i \mid f_\Phi) = \frac{I(Y'; f_i \mid f_\Phi)}{I(Y'; f_i \mid f_\Phi) + I(Y', \mathbf{f}_{\Phi\cup\{i\}})} \in [0, 1]$.

In our analysis, we focus on estimating two key components: marginal and conditional mutual information (MI and CMI), as well as joint multivariate dependence terms. This section outlines the various estimators we utilize, along with their assumptions, computational complexities, and error bounds. The Gaussian-Copula estimator, denoted as GCMI, calculates mutual information using a transformation that applies normal scores, followed by the application of closed-form expressions for Gaussian entropy. For two random vectors, X and Y, the mutual information can be expressed mathematically by a specific formula that involves the rank-based correlation, ρ_{XY}, of these transformed variables. For conditional mutual information, we employ partial correlations, which can also be represented with a similar mathematical expression. To use this estimator effectively, certain assumptions must hold: the process of rank-Gaussianization should provide a good approximation of the latent copula, covariance matrices must remain positive-definite, and the data should not exhibit a strong multimodal structure. The computational complexity of this method involves calculating a rank correlation matrix and inverting it, forming a complexity of $\mathcal{O}(k^2 n')$ for computation and $\mathcal{O}(k^3)$ for inversion. The error bound suggests that under sub-Gaussian copula assumptions, the difference between the estimated mutual information and the true value diminishes as the sample size increases. The k-nearest-neighbor (kNN) estimators, based on the Kozachenko-Leonenko method, utilize volume statistics from the nearest neighbors to compute mutual information. The specific formula for estimating MI incorporates the digamma function and the counts of neighbors in the respective dimensions. This method requires certain assumptions: it assumes that the densities are smooth with a bounded curvature, there is local isotropy in the neighborhoods defined by the k-ball, and the data should have a moderate intrinsic dimension. The complexity of this kNN approach is $\mathcal{O}(n' \log n')$ when utilizing a k-d tree for efficiency, although in worse cases, it could scale to $\mathcal{O}(n'^2)$. The error bound indicates that the accuracy of the estimation improves with larger sample sizes, specifically reflecting a dependency on the intrinsic dimension of the data. The plug-in estimator employs kernel density estimation (KDE) techniques for mutual information estimation. The estimator integrates a function over the joint density of the variables X and Y, comparing it against the product of their marginal densities. For the plug-in estimator to perform well, certain conditions need to be satisfied: the densities must be smooth, the bandwidth for the kernel should be selected via cross-validation, and there should be no exponential tail dependence. The computational complexity for this method is roughly $\mathcal{O}(n'^2)$, which highlights the potential for slower performance depending on the size of the dataset. The error bound suggests that if the KDE converges at a certain rate, then the mutual information estimate will also converge accordingly. In summary, each estimator has its strengths suited for different scenarios: The gcMI is recognized for its speed and stability, making it an excellent choice for datasets exhibiting moderate correlations. The kNN MI is robust in the presence of nonlinear dependencies, although it may have a higher variance. The Plug-In MI is noted for its accuracy but is slower and sensitive to the chosen bandwidth. To enhance performance,

MCIR (Mutual Correlation Impact Ratio) employs an automatic estimator-switching mechanism that aims to achieve oracle-level performance.

The computation of the Mutual Correlation Impact Ratio (MCIR) involves a structured three-stage process. First, we perform neighborhood screening to ascertain potential candidate dependency sets. Next, we estimate dependencies using the aforementioned MI and CMI estimators, applying bootstrap-based switching for optimal results. Finally, we compute the final MCIR vector. This section provides a step-by-step algorithmic description, details on estimator choices, and information on bootstrap protocols utilized across all datasets, ensuring a comprehensive understanding of the MCIR methodology.

MCIR employs a concise and stable set of parameters, which are straightforward to manage and contribute to the method's consistency and reliability.

- The screening threshold, denoted by γ, is selected within the range 0.20 to 0.35. This parameter determines which features are retained for subsequent analysis.
- For the k-nearest neighbors (kNN) mutual information estimator, the neighborhood size k is set to 5, so each data point considers its five nearest neighbors during calculation.
- In Gaussian–copula mutual information estimation, a rank transformation is applied to the data: $z = \Phi^{-1}(F_n(x))$. This transformation normalizes the data prior to analysis.
- For the bootstrap procedure, 200 replicates are used, repeating the calculation 200 times with resampled data to estimate variability.

MCIR selects and evaluates important features through a multi-step process. Initially, neighborhoods for each feature are constructed based on inter-feature relationships, using three methods: Pearson correlation for linear associations, distance correlation for nonlinear dependencies, and the k-nearest neighbors (kNN) graph structure to capture geometric relationships. Integrating these approaches ensures comprehensive consideration of all dependency types among features. Following neighborhood construction, the optimal subset of features for conditioning is selected. For each feature, a set of neighborhood candidates is identified. The AutoΦ algorithm then selects the subset that maximizes mutual information between the feature and the outcome, subject to a predefined subset size constraint. This process ensures that only the most informative and relevant features are retained for further analysis. A bootstrap protocol is implemented to enhance the reliability. For each feature, the MCIR score is computed using multiple estimators, and the calculations are repeated on resampled datasets to assess score variability. The estimator with the lowest standard error is selected. This approach provides confidence in the results and mitigates the influence of overfitting or random noise.

F Additional Theoretical Guarantees of MCIR

This section presents several auxiliary results that complement the main guarantees stated in Section D.4. These results establish (i) boundedness of MCIR, (ii) invariance under monotone transformations when using copula MI/CMI estimators, and (iii) ranking consistency with PCIR in weak-dependence regimes. All proofs are self-contained and do not duplicate arguments given in the preceding appendix sections.

Lemma 2 (Boundedness of MCIR)**.** *For any feature f_i and conditioning neighbourhood Φ for which $I(Y'; f_i \mid f_\Phi)$ and $I(Y'; f_{\Phi\cup\{i\}})$ are finite, the MCIR score*

$$C_i = \frac{I(Y'; f_i \mid f_\Phi)}{I(Y'; f_i \mid f_\Phi) + I(Y'; f_{\Phi\cup\{i\}})}$$

satisfies $0 \le C_i \le 1$.

Proof. Both numerator and denominator are non-negative because MI and CMI are Kullback–Leibler divergences. Since the denominator strictly exceeds the numerator whenever $I(Y'; f_{\Phi\cup\{i\}}) > 0$, the ratio lies in $(0, 1)$. If $I(Y'; f_i \mid f_\Phi) = 0$, the ratio is 0; if $I(Y'; f_{\Phi\cup\{i\}}) = 0$, then $C_i = 1$. Thus $C_i \in [0, 1]$. □

Proposition 9 (Monotone-Invariance under Gaussian–Copula MI)**.** *Let h_1, h_2, h_3 be strictly monotone functions applied elementwise to (Y', f_i, f_Φ). If MCIR is computed using a Gaussian–copula MI/CMI estimator on rank–Gaussianized variables, then*

$$C(Y'; f_i \mid f_\Phi) = C(h_1(Y'); h_2(f_i) \mid h_3(f_\Phi)).$$

Proof. Strictly monotone transformations preserve rank orderings. Gaussian–copula MI and CMI depend only on the copula correlation matrices of the rank-Gaussianized variables. Since the empirical copula is invariant under strictly monotone transforms, the estimated mutual information terms remain unchanged. Because MCIR is a ratio of MI and CMI terms, the score is likewise unchanged. □

Proposition 10 (Ranking Consistency with PCIR in Weak-Dependence Regimes)**.** *Suppose the conditioning neighbourhood Φ is such that $I(Y'; f_i \mid f_\Phi) \approx I(Y'; f_i)$ and $I(Y'; f_{\Phi\cup\{i\}}) \approx I(Y'; f_i) + I(Y'; f_\Phi)$ for all i up to $o(1)$. Then MCIR induces the same feature ranking as PCIR:*

$$C_i > C_j \quad \Longleftrightarrow \quad I(Y'; f_i) > I(Y'; f_j) \quad \Longleftrightarrow \quad \eta_{f_i} > \eta_{f_j}.$$

Proof. Under the assumptions,

$$C_i = \frac{I(Y'; f_i \mid f_\Phi)}{I(Y'; f_i \mid f_\Phi) + I(Y'; f_{\Phi\cup\{i\}})} \approx \frac{I(Y'; f_i)}{2I(Y'; f_i) + I(Y'; f_\Phi)}.$$

The denominator shares the same additive constant $I(Y'; f_\Phi)$ for all i, and the remaining terms preserve monotonicity in $I(Y'; f_i)$. Since PCIR is also monotone in any scalar association measure between f_i and Y', the two scorings induce identical rankings up to $o(1)$ discrepancies. □

Theorem 20 (Redundancy Collapse under Exact Functional Dependence)**.** *If $f_i = g(f_j)$ almost surely for some measurable g and $j \in \Phi$, then*

$$I(Y'; f_i \mid f_\Phi) = 0 \quad \textit{and hence} \quad C_i = 0.$$

Property	MCIR	PCIR	MI Ranking	SHAP
Locality	Local (m_Φ)	Global	Global	Local to prediction
Conditional dependence	Yes	No	No	Yes (model-based)
Captures unique info	Yes	No	No	Sometimes
Model dependence	None	None	None	Strong
Scales to $k \gg n$	Yes	Yes	Yes	No (kernel SHAP)
Computational cost	Near-linear	Linear	Linear	Exponential/approximate
Interpretable numerator	CMI $(Y; f_i \mid \Phi)$	Correlation	MI	Shapley payoff
Redundancy collapse	Guaranteed	Not guaranteed	Not guaranteed	Not guaranteed

Table 17: Comparison of MCIR with related attribution and dependence measures.

Proof. If $f_i = g(f_j)$ and $j \in \Phi$, then f_i is measurable with respect to $\sigma(f_\Phi)$. By definition of conditional independence,

$$p(Y' \mid f_i, f_\Phi) = p(Y' \mid f_\Phi).$$

Thus $D_{\mathrm{KL}}\big(p(Y' \mid f_i, f_\Phi) \,\|\, p(Y' \mid f_\Phi)\big) = 0$ implying $I(Y'; f_i \mid f_\Phi) = 0$. Plugging this into the MCIR formula yields $C_i = 0$. □

Lemma 3 (MCIR Vanishes Under Conditional Independence)**.** *If $Y' \perp f_i \mid f_\Phi$, then*

$$I(Y'; f_i \mid f_\Phi) = 0 \quad \textit{and hence} \quad C_i = 0.$$

Proof. Conditional independence implies $p(Y' \mid f_i, f_\Phi) = p(Y' \mid f_\Phi)$ almost surely. The conditional mutual information is the Kullback–Leibler divergence between these two conditional densities, which is zero under equality. Substituting into the MCIR formula yields $C_i = 0$. □

F.1 Comparison Against Other Attribution Methods

Table 17 summarises how MCIR differs from PCIR, mutual-information ranking, and SHAP-based explainers. This comparison highlights that MCIR occupies a unique middle ground between statistical association measures and fully model-based attribution methods.