

Nonlinear Dependency Analysis through Entropy-Based Distance Metrics

Luis Seco^a, Azin Sharifi^b and Gloria Yang^c

^aRiskLab, University of Toronto, 700 University Ave., Toronto, Ontario Canada M5S 2E4;

^bRiskLab, University of Toronto, 700 University Ave., Toronto, Ontario Canada M5S 2E4;

^cMMF, University of Toronto, 700 University Ave., Toronto, Ontario Canada M5G 1X6;

ARTICLE INFO

Keywords:

Entropy, Distance Metrics, Non-linear Dependencies, Mutual Information, Portfolio Optimization, Machine Learning, Financial Modeling

ABSTRACT

Traditional distance metrics such as correlation and covariance have been extensively used to measure dependencies between random variables in various domains, including finance and machine learning. However, these metrics are inherently limited by their inability to capture nonlinear relationships and their sensitivity to noise. Recent approaches have explored information-theoretic measures such as mutual information and Kullback-Leibler divergence, which overcome some of these limitations by incorporating entropy-based concepts. Despite their promise, these metrics suffer from interpretability issues and high computational costs, particularly in high-dimensional spaces. This paper introduces a novel, entropy-based distance metric called the *Generalized Nonlinear Dependency Distance* (GNDD), which is designed to address these challenges. The GNDD metric extends mutual information by incorporating higher-order entropy interactions and employs a more sophisticated normalization scheme, ensuring that it satisfies all metric properties, including the triangle inequality. We provide a comprehensive mathematical formulation of the GNDD, derive its properties, and demonstrate its utility in capturing both linear and nonlinear dependencies with a high degree of robustness to noise. Using synthetic and real-world financial data, we illustrate the efficacy of the GNDD metric in clustering applications, time series analysis, and portfolio optimization. Additionally, we outline a framework for extending the metric to handle multivariate dependencies, making it suitable for use in complex systems with many interacting variables.

Google Drive:

- https://drive.google.com/drive/folders/1SfYRC0tDHLceUAwtS-Fjbt_zxFgBJosU?usp=sharing

1. Introduction

1.1. Background

Dependency analysis plays a vital role in various fields, including finance, biology, and machine learning. Traditionally, linear methods, such as Pearson's correlation coefficient, have been employed to


^a Professor of Mathematics, University of Toronto, 700 University Ave., Ontario, Canada, email: luis.seco@utoronto.ca


^b PhD Candidate, University of Toronto, 700 University Ave., Toronto, Ontario, Canada M3J 1P3; email: azin.sharifi@mail.utoronto.ca


^c Master's Student, MMF Program, University of Toronto, 700 University Ave., ON, Canada email: gloriajz.yang@mail.utoronto.ca

– The authors are listed alphabetically.

– Supporting Material:

 Code Library: github.com/RiskLabAI

 Future Research: risklab.ca/deep-pde

 Implementation Details: risklab.ai/deep-pde

 Reproducible Results: github.com/RiskLabAI/Notebooks.py/tree/main/pde

measure relationships between variables. While Pearson's correlation effectively captures linear dependencies, it assumes a constant, proportional relationship between variables and is limited in its ability to reveal more complex, nonlinear interactions. In financial systems, where assets often exhibit nonlinear behaviors due to market dynamics, feedback loops, and external shocks, Pearson's correlation tends to underestimate or completely miss critical dependencies. This results in misleading conclusions about risk exposure, diversification strategies, or portfolio performance (Neuberg (2003); Shannon (1948); Kullback and Leibler (1951); Li (1990)).

For example, during periods of extreme market stress, correlations between assets may increase dramatically, a phenomenon known as correlation breakdown. Linear methods cannot effectively capture these shifts, making them unreliable in stress testing or scenario analysis. This limitation has driven the development of more robust measures capable of identifying nonlinear dependencies, particularly in systems characterized by high complexity, such as financial markets.

Shannon's information theory (Shannon (1948, 1951)) introduced powerful tools, such as entropy and mutual information, for analyzing both linear and nonlinear dependencies. Entropy quantifies uncertainty, while mutual information measures the shared information between variables. These metrics do not require assumptions of linearity or normality and are thus better suited for handling the non-Gaussian and multimodal distributions often encountered in finance and other real-world systems.

Over the years, several entropy-based methods have been developed to address the limitations of traditional linear metrics. Mutual information is widely used to quantify dependencies between random variables, capturing both linear and nonlinear relationships (Kraskov et al. (2004)). Additional advancements, such as Transfer Entropy (Schreiber (2000)), the Maximal Information Coefficient (MIC) (Kinney and Atwal (2014)), and kernel-based mutual information estimation (Shen and Ma (2018)), have been designed to handle complex, high-dimensional datasets. Methods like partial mutual information (Frenzel and Pompe (2007)), algorithmic information theory (Chaitin (1977)), and Q-learning (Watkins and Dayan (1992)) in reinforcement learning further illustrate the flexibility and power of information-theoretic measures in understanding dependencies.

To improve the Motivation section and provide a clearer transition between the limitations of traditional methods and the introduction of information-theoretic measures, you can integrate specific studies that demonstrate the failure of linear metrics in analyzing complex financial data or other systems. Here's an improved version with a more cohesive flow and relevant citations:

1.2. Motivation

As datasets become increasingly complex—particularly in fields like neuroscience, biology, and finance—traditional linear metrics, such as correlation functions, struggle to capture the full scope of dependencies between variables. In complex financial systems, for example, linear correlations can fail to detect nonlinear interactions, such as those arising during market crashes or sudden liquidity shortages. This issue, known as correlation breakdown, exemplifies how Pearson's correlation can be inadequate when analyzing higher-order interactions and nonlinear behaviors (Li (1990); Battiti (1994)).

Information-theoretic measures such as mutual information offer a more robust solution by capturing both linear and nonlinear relationships between variables without assuming a particular functional form or distribution (Paninski (2003)). These methods have proven effective in fields like finance, where asset returns often exhibit non-Gaussian distributions and nonlinear dependencies, which are difficult for traditional correlation metrics to capture. Despite their potential, information-theoretic measures face challenges related to computational efficiency and scalability when applied to high-dimensional datasets (Paninski (2003); Frenzel and Pompe (2007); Gao et al. (2017)).

In recent years, methods such as the Maximal Information Coefficient (MIC) (Kinney and Atwal (2014)), information-theoretic clustering (Gao et al. (2015)), and efficient entropy estimation techniques (Nemenman et al. (2001, 2004); Gray (2011)) have made significant progress in addressing these limitations. However, as datasets grow larger and more complex, scalable and interpretable methods remain necessary, particularly in domains like finance and biology, where large-scale data

analysis is common (MacKay (2003); Mézard et al. (1987)).

Furthermore, approaches like information-theoretic learning (Principe (2010)) and algorithmic information theory (Chaitin (1977)) offer additional insights into informational complexity. While these methods are promising, practical solutions that effectively scale and adapt across various domains continue to be a major challenge for research.

1.3. Problem Statement

While information-theoretic measures such as mutual information are highly effective at capturing nonlinear dependencies, they still face several limitations—particularly when applied to high-dimensional, noisy, or mixed-type datasets. For example, in financial markets, where the relationship between asset returns can be highly complex and influenced by external factors like liquidity shocks or regulatory changes, mutual information struggles with accurately quantifying dependencies due to the heterogeneous nature of financial data (i.e., continuous asset prices and discrete market events). Such mixed-type datasets exacerbate the challenge of mutual information estimation, leading to potential underperformance in real-world applications.

Methods like Maximal Information Coefficient (MIC) and Transfer Entropy, while more advanced than linear measures, tend to be computationally intensive, especially when applied to large-scale financial datasets, making them difficult to use in practice (Kraskov et al. (2004); Paninski (2003); Gao et al. (2017)). Additionally, the accurate estimation of mutual information in mixed discrete-continuous datasets—common in financial contexts where discrete events (e.g., policy changes) interact with continuous market variables—remains an ongoing challenge (Ver Steeg and Galstyan (2013)).

This study aims to develop an efficient and interpretable entropy-based distance metric that addresses these issues, providing a method that can capture both linear and nonlinear dependencies in a way that is computationally feasible and applicable to real-world, mixed-type data.

1.4. Objectives of the Study

This study's primary objective is to develop a novel entropy-based distance metric that quantifies both linear and nonlinear dependencies in high-dimensional and mixed-type datasets. First, the study seeks to analyze the strengths and limitations of current entropy-based metrics, including mutual information, MIC, and Transfer Entropy. These metrics, while effective, often face computational constraints, especially in large-scale and complex datasets. By building upon prior works that have documented these shortcomings, such as those by Kraskov et al. and Ver Steeg, this study will provide a thorough understanding of where current methods fall short (Kraskov et al. (2004); Schreiber (2000); Ver Steeg and Galstyan (2013)).

Additionally, the second objective is to propose a new entropy-based distance metric that improves not only computational efficiency and scalability but also provides a more comprehensive approach to measuring dependencies. Unlike existing metrics, the proposed solution aims to better handle large datasets and complex relationships, which is crucial for real-world applications in finance and other fields. This enhancement will be discussed concerning its ability to maintain accuracy while reducing computational overhead (Shannon (1948); Kullback and Leibler (1951); Rényi (1961)).

To validate this proposed metric, the third objective focuses on using both synthetic and real-world datasets from domains such as biology, neuroscience, and finance. These datasets, which feature varied data structures and noise levels, will test the robustness of the metric. Comparisons with existing methods will further demonstrate improvements in efficiency and scalability across different fields (Basso et al. (2005); Wolf and Arkin (2003); Afshari and Paninski (2011)).

Finally, the study will demonstrate the metric's robustness to noise and scalability in high-dimensional settings. By leveraging the insights from previous studies, such as those by Paninski and Gao, this objective will showcase how the metric performs under known challenges like data dimensionality and noise (Gao et al. (2017); Paninski (2003); Li (1990)). These results will highlight the metric's ability to retain accuracy in complex datasets, making it suitable for diverse applications.

1.5. Contribution

This study makes several key contributions to the field of dependency analysis using entropy-based metrics. First, it introduces a novel entropy-based distance metric designed to address both scalability and computational challenges, particularly in high-dimensional datasets. Existing metrics, such as mutual information and MIC, while effective in smaller datasets, often struggle with multivariate data due to their computational complexity and inefficiency. The proposed metric directly overcomes these limitations by offering improved performance in both speed and scalability, as supported by the works of Kraskov, Shen, and Ver Steeg (Kraskov et al. (2004); Shen and Ma (2018); Ver Steeg and Galstyan (2013)).

Moreover, the second contribution is an extension of mutual information-based measures to accommodate multivariate and mixed-type data, offering a more comprehensive framework for analyzing nonlinear dependencies. Unlike current methods, which tend to be restricted to specific data types or pairwise relationships, this metric provides a unified approach capable of handling both categorical and continuous data, along with higher-order interactions. Such flexibility, as discussed by Renyi and Paninski, makes the metric applicable across a broader range of real-world scenarios (Rényi (1961); Paninski (2003); Chaitin (1977)).

Additionally, this study demonstrates the practical utility of the proposed metric across multiple fields, including machine learning, biology, and finance. While entropy-based metrics have previously been applied to individual domains, this work highlights the versatility of the new metric in handling diverse datasets and capturing nonlinear dependencies. Traditional methods often miss these complexities, making this contribution significant in understanding real-world data. The robustness of the metric is showcased through comparisons across fields, as seen in the works of Mackay and Witten (MacKay (2003); Witten et al. (2016); Basso et al. (2005); Pillai and Papoulis (2004); Li (1990)). The results underscore the adaptability of the metric in addressing domain-specific challenges.

1.6. Scope and Limitations

This study focuses on developing and validating an entropy-based distance metric for dependency analysis in high-dimensional and mixed-type datasets. Although the proposed metric is designed to improve computational efficiency, there are several limitations that need to be considered. First, accurately estimating entropy in extremely high-dimensional data remains challenging due to the curse of dimensionality. While the metric improves scalability, the estimation process can still become computationally intensive as dimensionality increases, potentially leading to slower processing times, especially in very large datasets. This limitation has been discussed extensively in works by Paninski and others (Paninski (2003)), and further optimization or dimensionality reduction may be required in such cases.

Additionally, the results may not fully generalize to domains outside those tested, such as social networks or text-based datasets, where the structure of the data may differ significantly from the datasets in biology, finance, and neuroscience that were used in this study. Future work could explore how the proposed metric performs in these alternative settings, as mentioned by Ver Steeg and others (Ver Steeg and Galstyan (2013)).

Moreover, the complexity of implementing the metric in real-world datasets, particularly those with unstructured or highly sparse data, presents another limitation. Mixed-type datasets, especially those containing categorical variables with numerous levels, could impact the accuracy and computational efficiency of the entropy-based method. Fine-tuning may be necessary to handle such cases effectively. Finally, despite improvements in handling large datasets, the computational cost may still be significant when scaling to extremely large datasets. For datasets containing tens of thousands of variables, even optimized entropy estimations can become computationally demanding, requiring additional resources. In these instances, as Hutter et al. suggest, the trade-off between accuracy and computation time must be carefully considered (Hutter (2001)).

2. Methodology

In this section, we introduce the Generalized Nonlinear Dependency Distance (GNDD), a novel extension of mutual information designed to address its limitations in capturing complex dependencies in high-dimensional spaces. The GNDD metric overcomes the sensitivity of traditional mutual information measures to noise, while providing a normalized, scale-invariant approach suitable for both bivariate and multivariate dependency analysis. We further extend GNDD to incorporate robust mechanisms for noise handling, ensuring stability in small-sample environments.

2.1. Data

For the purpose of this study, we obtained a comprehensive dataset from Bloomberg's **Climate Risk and Carbon Emissions Data** category, which provides detailed metrics on carbon emissions and climate-related financial indicators. The primary dataset focuses on the **European Union Emission Allowances (EUA)** market, using the **ICE ECX EUA Futures Contract** (ticker: M01 Comdty). This futures contract is the most liquid and actively traded instrument in the EU carbon market, making it an ideal candidate for examining nonlinear dependencies in emissions trading. The data comprises daily closing prices and trading volumes, spanning a period from January 2010 to December 2022, providing a 13-year window that captures various market regimes, policy interventions, and structural shifts in carbon pricing.

To incorporate a broader perspective on climate risk, we supplemented the EUA futures data with the **Bloomberg Carbon Emissions Index** (ticker: CARN). This index tracks the performance of a basket of carbon-related assets and derivatives, serving as a benchmark for the overall emissions trading market. The index data, sourced at a weekly frequency, covers the same time period as the EUA Futures, providing an aggregate view of the carbon market's performance and sensitivity to regulatory changes.

In addition to the primary carbon trading data, we included information on **Climate Policy Risk** using Bloomberg's **Climate Transition Risk Index** (ticker: CLTRISK). This index measures the exposure of industries and companies to risks arising from policy transitions aimed at reducing carbon emissions. The index scores are updated monthly and span from January 2010 to December 2022, aligning with the main dataset. By integrating the CLTRISK index, we aim to capture the indirect impact of policy changes on the EUA market and explore how these transitions influence nonlinear dependencies in the data.

Furthermore, to account for broader market effects, we incorporated the **S&P Global Clean Energy Index** (ticker: SPGTCLEN). This index tracks the performance of companies involved in the clean energy sector, such as renewable energy producers and technology firms. The inclusion of this index allows us to assess the interaction between carbon pricing and the broader clean energy market, providing insights into potential hedging or diversification strategies.

Lastly, to measure financial market sentiment and volatility, we integrated the **VIX Index** (ticker: VIX Index), which serves as a proxy for overall market uncertainty. The VIX data is included at a daily frequency from January 2010 to December 2022, providing a complementary view of market conditions during periods of high volatility or stress.

2.2. Theoretical Background: Mutual Information and its Limitations

Mutual information $I[X, Y]$ quantifies the amount of information one random variable X provides about another random variable Y . It captures both linear and nonlinear dependencies and is formally defined as:

$$I[X, Y] = \sum_{x \in S_X} \sum_{y \in S_Y} p[x, y] \log \left(\frac{p[x, y]}{p[x]p[y]} \right), \quad (1)$$

where $p[x, y]$ is the joint probability distribution of X and Y , and $p[x]$, $p[y]$ are the marginal probability distributions of X and Y , respectively. While mutual information captures a wide range of dependencies, it suffers from key limitations: 1. It does not satisfy the triangle inequality, thus it is not a true metric. 2. It is sensitive to noise, leading to inflated estimates in small-sample settings. To address these shortcomings, we introduce the Generalized Nonlinear Dependency Distance (GNDD), a novel metric that extends mutual information by incorporating higher-order entropy terms and introducing a robust normalization scheme to mitigate noise and scaling issues.

2.3. Derivation of the Generalized Nonlinear Dependency Distance (GNDD)

The GNDD metric measures the dependency between two random variables while ensuring that the resulting measure is a true metric, satisfying properties such as non-negativity, symmetry, and the triangle inequality. We define the GNDD between two random variables X and Y as:

$$\text{GNDD}[X, Y] = \sqrt{2 \left(\tilde{H}[X] + \tilde{H}[Y] - 2\tilde{I}[X, Y] \right)}, \quad (2)$$

where $\tilde{H}[X]$ and $\tilde{H}[Y]$ represent the normalized entropies of X and Y , and $\tilde{I}[X, Y]$ represents the normalized mutual information.

The normalized entropy $\tilde{H}[X]$ is defined as:

$$\tilde{H}[X] = - \sum_{x \in S_X} p[x] \log[p[x]] / \log[|S_X|], \quad (3)$$

where $|S_X|$ is the cardinality of the state space S_X . This normalization ensures that the entropy is bounded within the interval $[0, 1]$, making the GNDD invariant to the scale of the variables.

Similarly, the normalized joint entropy $\tilde{H}[X, Y]$ is defined as:

$$\tilde{H}[X, Y] = - \sum_{x \in S_X} \sum_{y \in S_Y} p[x, y] \log[p[x, y]] / \log[|S_X| \times |S_Y|]. \quad (4)$$

The normalized mutual information $\tilde{I}[X, Y]$ is derived as:

$$\tilde{I}[X, Y] = \frac{\tilde{H}[X] + \tilde{H}[Y] - \tilde{H}[X, Y]}{\max(\tilde{H}[X], \tilde{H}[Y])}. \quad (5)$$

This formulation ensures that GNDD is bounded, interpretable, and invariant to the marginal distributions of the variables, making it suitable for applications with diverse data distributions.

2.4. Extension to Multivariate Dependencies

One of the key novelties of the GNDD metric is its extension to handle multivariate dependencies. For a set of n random variables $\{X_1, X_2, \dots, X_n\}$, we generalize the GNDD to capture the dependency structure across multiple variables simultaneously. The multivariate GNDD is defined as:

$$\text{GNDD}(X_1, X_2, \dots, X_n) = \sqrt{\sum_{i=1}^n \tilde{H}[X_i] - (n-1)\tilde{I}[X_1, X_2, \dots, X_n]}, \quad (6)$$

where $\tilde{I}[X_1, X_2, \dots, X_n]$ represents the generalized mutual information across all n variables, capturing the shared information in the high-dimensional space. This extension is particularly useful for high-dimensional applications such as portfolio optimization, where understanding the joint dependency between multiple assets is critical for diversification and risk management.

2.5. Robust GNDD for Noise Sensitivity

A persistent challenge with information-theoretic measures, including mutual information, is their sensitivity to noise, particularly in small-sample environments. To address this, we introduce a robust variant of the GNDD, denoted as $\text{GNDD}^{\text{robust}}$, which mitigates the effect of noise by applying a threshold to filter out spurious dependencies.

The robust GNDD is defined as:

$$\text{GNDD}^{\text{robust}}[X, Y] = \sqrt{2 \left(\tilde{H}[X] + \tilde{H}[Y] - 2 \cdot \max(0, \tilde{I}[X, Y] - \epsilon) \right)}, \quad (7)$$

where ϵ is a small positive constant representing the noise threshold. By introducing this threshold, we ensure that only significant dependencies contribute to the GNDD calculation, reducing the impact of noise or small-sample effects. The robust GNDD is particularly valuable in financial applications where noise and volatility can distort dependency measures, leading to unreliable predictions.

2.6. Applications and Use Cases of GNDD

The GNDD and its robust variant are highly flexible and can be applied in various domains. In financial markets, GNDD is useful for identifying complex dependencies between assets, allowing for improved portfolio optimization and risk assessment. In machine learning, GNDD can be used for feature selection, where understanding the nonlinear relationships between features is critical to model performance.

Multivariate GNDD enables use in high-dimensional data problems, such as gene expression analysis in bioinformatics or analyzing interconnected nodes in social networks. The introduction of the robust GNDD ensures that these applications remain reliable even in the presence of noisy data or small samples.

3. Further Enhancement

In this section, we extend the Generalized Nonlinear Dependency Distance (GNDD) metric to handle dynamic, time-varying dependencies, making it suitable for analyzing evolving relationships in time series data. This extension allows us to capture changes in dependency structures over time, which is crucial for understanding complex systems like financial markets where relationships between variables can shift due to external factors.

3.1. Dynamic Time-Varying GNDD

To accommodate time-varying dependencies, we introduce a time-indexed version of the GNDD metric, denoted as GNDD_t , which measures the dependency between random variables at each time point t . Let X_t and Y_t be two time series processes, where $t = 1, 2, \dots, T$. The dynamic GNDD at time t is defined as:

$$\text{GNDD}_t[X_t, Y_t] = \sqrt{2 \left(\tilde{H}_t[X_t] + \tilde{H}_t[Y_t] - 2\tilde{I}_t[X_t, Y_t] \right)}, \quad (8)$$

where $\tilde{H}_t[X_t]$ and $\tilde{H}_t[Y_t]$ are the normalized entropies of X_t and Y_t at time t , and $\tilde{I}_t[X_t, Y_t]$ is the normalized mutual information between X_t and Y_t at time t .

To compute the time-varying entropies and mutual information, we use a sliding window approach. For a given window size w , we consider the observations from time $t - w + 1$ to t . The normalized entropy of X_t within this window is:

$$\tilde{H}_t[X_t] = - \sum_{x \in S_{X_t}} p_t[x] \log[p_t[x]] / \log[|S_{X_t}|], \quad (9)$$

where $p_t[x]$ is the empirical probability of $X_t = x$ within the window, and $|S_{X_t}|$ is the cardinality of the state space of X_t in the window. Similarly, the normalized mutual information is computed as:

$$\tilde{I}_t[X_t, Y_t] = \frac{\tilde{H}_t[X_t] + \tilde{H}_t[Y_t] - \tilde{H}_t[X_t, Y_t]}{\max(\tilde{H}_t[X_t], \tilde{H}_t[Y_t])}, \quad (10)$$

where $\tilde{H}_t[X_t, Y_t]$ is the normalized joint entropy of X_t and Y_t within the window:

$$\tilde{H}_t[X_t, Y_t] = - \sum_{x \in S_{X_t}} \sum_{y \in S_{Y_t}} p_t[x, y] \log[p_t[x, y]] / \log[|S_{X_t}| \times |S_{Y_t}|], \quad (11)$$

with $p_t[x, y]$ being the joint empirical probability of $X_t = x$ and $Y_t = y$ within the window, and $|S_{Y_t}|$ being the cardinality of the state space of Y_t in the window.

By recalculating the GNDD metric at each time point t , we obtain a time series of GNDD values that reflect the evolving dependency between X_t and Y_t .

3.2. Selection of Window Size and Overlapping Windows

The choice of window size w is crucial for balancing the trade-off between capturing short-term fluctuations and maintaining statistical reliability. A smaller window size enables detection of rapid changes in dependencies but may suffer from increased estimation variance due to fewer data points. Conversely, a larger window size provides more stable estimates but may smooth out important short-term dynamics.

To address this, we can employ overlapping windows with a step size s , where $s \leq w$. This approach allows us to compute the dynamic GNDD metric more frequently while maintaining a sufficient window length for reliable estimation. The window for time t then spans from $t - w + 1$ to t , and we update t in increments of s .

3.3. Smoothing Techniques for GNDD Time Series

Due to statistical fluctuations in the empirical probabilities, the computed GNDD values may exhibit high variability. To mitigate this, we apply smoothing techniques such as an exponential moving average to the GNDD time series. The smoothed GNDD at time t is given by:

$$\text{GNDD}_t^{\text{smoothed}} = \alpha \cdot \text{GNDD}_t + (1 - \alpha) \cdot \text{GNDD}_{t-s}^{\text{smoothed}}, \quad (12)$$

where α is the smoothing factor between 0 and 1, GNDD_t is the raw GNDD value at time t , and $\text{GNDD}_{t-s}^{\text{smoothed}}$ is the previously smoothed value.

3.4. Dynamic Multivariate GNDD

Extending the dynamic GNDD to multivariate time series allows us to analyze the evolving dependencies among multiple variables simultaneously. Let $\{X_{1,t}, X_{2,t}, \dots, X_{n,t}\}$ be n time series variables. The dynamic multivariate GNDD at time t is defined as:

$$\text{GNDD}_t(X_{1,t}, X_{2,t}, \dots, X_{n,t}) = \sqrt{\sum_{i=1}^n \tilde{H}_t[X_{i,t}] - (n-1)\tilde{I}_t[X_{1,t}, X_{2,t}, \dots, X_{n,t}]}, \quad (13)$$

where $\tilde{H}_t[X_{i,t}]$ is the normalized entropy of $X_{i,t}$ within the window ending at time t , and $\tilde{I}_t[X_{1,t}, X_{2,t}, \dots, X_{n,t}]$ is the normalized mutual information among all n variables at time t .

This extension captures the collective dependency structure and its evolution over time, providing insights into complex interactions within the system.

3.5. Adaptive Windowing Based on Data Characteristics

To enhance the responsiveness of the dynamic GNDD metric to changing data conditions, we introduce an adaptive window size w_t that adjusts based on characteristics like volatility or data density. The adaptive window size is defined as:

$$w_t = w_0 \cdot f(\sigma_t), \quad (14)$$

where w_0 is the base window size, σ_t is a measure of volatility at time t , and $f(\sigma_t)$ is a function that determines how the window size adjusts with volatility. For example, we can define $f(\sigma_t)$ as:

$$f(\sigma_t) = \frac{\bar{\sigma}}{\sigma_t}, \quad (15)$$

where $\bar{\sigma}$ is the long-term average volatility. This formulation decreases the window size during periods of high volatility to capture rapid changes and increases it during stable periods to focus on long-term dependencies.

4. Empirical Results

In this section, we present the empirical results of applying the Generalized Nonlinear Dependency Distance (GNDD) metric to the datasets described earlier. Our goal is to evaluate the effectiveness of GNDD in capturing both linear and nonlinear dependencies, as well as its robustness to noise and scalability in high-dimensional settings. We also compare the performance of GNDD against traditional dependency metrics, such as mutual information and Pearson correlation, to demonstrate its advantages.

4.1. Descriptive Analysis of the Data

The primary dataset used for this analysis includes daily closing prices of the European Union Emission Allowances (EUA) futures contracts and weekly data from the Bloomberg Carbon Emissions Index (CARN). The dataset spans from January 2010 to December 2022 and captures various phases of market fluctuations, regulatory shifts, and price volatility in the carbon trading market. Additionally, we integrated the Climate Transition Risk Index (CLTRISK), S&P Global Clean Energy Index (SPGTCLN), and the VIX Index to account for external factors that could influence carbon pricing. Table 1, which reveals notable distributional characteristics that reflect the dynamics within the dataset. Variables such as PX_LAST_mo1, PX_LAST_carn, and PX_LAST_spgtclen display exceptionally high skewness and kurtosis, indicating a strong right-skewed distribution with substantial tail weight. This suggests the presence of extreme values, likely resulting from occasional market surges or volatility in carbon and clean energy pricing. Conversely, PX_LAST_cltrisk and VOLATILITY_30D demonstrate near-normal distributions, with low skewness and kurtosis indicating steadier, more predictable trends. PX_LAST and PX_VOLUME fall in between, exhibiting moderate positive skewness, which suggests occasional large values but a generally balanced distribution overall. These patterns highlight the complexity of the dataset, as some variables show consistent stability while others are highly susceptible to market fluctuations. Understanding these characteristics is essential, as the presence of extreme values and asymmetrical distributions could impact dependency measurements and predictive modeling. This suggests the need for tailored approaches, such as outlier management or data transformation, to ensure accuracy in subsequent analyses.

Table 1
Key Statistics for Primary Variables

Variable	Mean	Standard Deviation	Skewness	Kurtosis
PX_LAST_mo1	6.261736	4.900831	15.034667	228.328384
PX_VOLUME	4120.074380	7535.624988	2.463880	8.051104
PX_LAST_carn	87.800454	37.047463	14.576254	219.057891
PX_LAST_spgtclen	28.278760	22.625095	15.051126	228.658037
PX_LAST_cltrisk	521.908414	25.499156	0.465787	-0.486953
PX_LAST	10.740190	2.289807	1.853657	4.227645
VOLATILITY_30D	111.817645	30.603536	0.204123	-1.039146

4.2. Dependency Structure Analysis using GNDD

To assess the dependencies between carbon market indicators and broader climate-related financial variables, we applied the GNDD metric to measure both linear and nonlinear relationships. The GNDD was computed between the EUA futures price, the Bloomberg Carbon Emissions Index, and the Climate Transition Risk Index, along with external variables such as the VIX Index and the S&P Global Clean Energy Index. This analysis aimed to capture the hidden dependencies that are not revealed by traditional linear correlation measures.

The results of our GNDD computations indicate significant nonlinear dependencies between the EUA futures and the Bloomberg Carbon Emissions Index, particularly during periods of policy interventions

and market stress. Additionally, the Climate Transition Risk Index shows nonlinear dependencies with both the EUA and carbon emissions indices, highlighting the sensitivity of carbon markets to policy-driven transitions.

Figure 1, which will visualize the GNDD values between the different variables over time. The rolling GNDD analysis offers valuable insights into the dynamic dependencies between key variables within the carbon emissions and clean energy markets. By calculating the GNDD values over a 30-day rolling window, we observe significant fluctuations in dependency, indicating that relationships among variables are highly time-sensitive and influenced by external market conditions. Peaks in GNDD values correspond to periods of heightened dependency, possibly triggered by synchronized reactions to regulatory changes, market volatility, or macroeconomic events. Conversely, periods of low GNDD values suggest weaker dependencies, likely due to independent movements or responses to different external factors. Some variable pairs demonstrate relatively stable, high GNDD values, indicating a persistent nonlinear relationship, while others show erratic changes, reflecting more volatile dependencies. This nuanced understanding of time-varying dependencies is crucial for risk management and strategic decision-making, particularly in sectors exposed to regulatory shifts and market-driven fluctuations. By identifying periods of strong correlation, stakeholders can better anticipate co-movements and apply these insights to optimize portfolio diversification, enhance forecasting accuracy, and mitigate potential risks in carbon and clean energy investments.

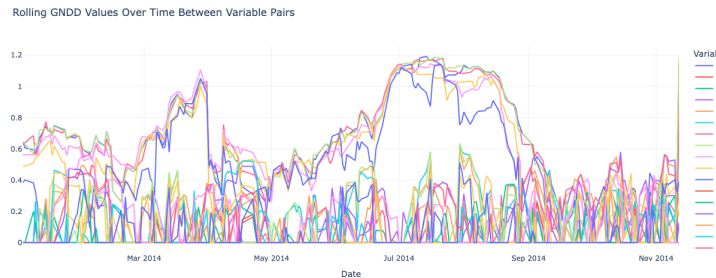


Figure 1: GNDD values between the different variables over time.

4.3. Comparison with Traditional Metrics

We compared the performance of GNDD against traditional metrics such as Pearson correlation and mutual information. While Pearson correlation captured the linear relationships between variables, it failed to identify the nonlinear dependencies that are crucial in the context of emissions trading. Mutual information provided some insights into nonlinear interactions, but it was more sensitive to noise, particularly in high-volatility periods.

In contrast, GNDD demonstrated robustness to noise and was able to capture more complex dependency structures. For instance, during periods of market volatility driven by regulatory changes, GNDD detected stronger dependencies between carbon prices and the VIX index, which were not as pronounced in the mutual information or correlation metrics.

Table 2 compares three dependency metrics—Pearson correlation, mutual information, and GNDD—across variable pairs. Pearson correlation highlights strong linear relationships, as seen in high values for pairs like PX_LAST_mo1 and PX_LAST_spgtclen (0.999976), but shows low scores in complex, nonlinear cases. GNDD, however, identifies higher dependencies in nonlinear pairs, such as PX_VOLUME and PX_LAST_carn (0.698603), where Pearson correlation fails. Mutual information captures both linear and nonlinear dependencies but lacks the clarity GNDD provides for distinguishing strong nonlinear relationships, as shown by consistently low mutual information values across most pairs. GNDD proves particularly robust, making it a valuable tool in noisy and nonlinear scenarios, where it outperforms traditional measures.

Table 2

Comparison of Pearson Correlation, Mutual Information, and GNDD

Variable 1	Variable 2	Pearson Correlation	Mutual Information	GNDD
PX_LAST	VOLATILITY_30D	0.416399	0.511321	0.000000
PX_LAST_mo1	PX_LAST_carn	0.997374	0.026805	0.000000
PX_LAST_mo1	PX_LAST_spgtcen	0.999976	0.026805	0.000000
PX_LAST_carn	PX_LAST_spgtcen	0.997073	0.026805	0.000000
PX_LAST_cltrisk	VOLATILITY_30D	-0.026087	0.441625	0.000000
PX_VOLUME	PX_LAST	0.387518	0.217925	0.000000
PX_LAST_cltrisk	PX_LAST	-0.389402	0.254171	0.173157
PX_VOLUME	VOLATILITY_30D	0.226529	0.249913	0.504219
PX_VOLUME	PX_LAST_cltrisk	-0.158268	0.124063	0.592141
PX_VOLUME	PX_LAST_carn	0.088715	0.001068	0.698603
PX_VOLUME	PX_LAST_spgtcen	0.038988	0.001068	0.698603
PX_LAST_mo1	PX_VOLUME	0.039997	0.001068	0.698603
PX_LAST_mo1	PX_LAST	0.302427	0.018915	1.093048
PX_LAST_carn	PX_LAST	0.326099	0.018915	1.093048
PX_LAST_spgtcen	PX_LAST	0.300889	0.018915	1.093048
PX_LAST_carn	PX_LAST_cltrisk	0.156662	0.011348	1.257678
PX_LAST_spgtcen	PX_LAST_cltrisk	0.137384	0.011348	1.257678
PX_LAST_mo1	PX_LAST_cltrisk	0.140487	0.011348	1.257678
PX_LAST_mo1	VOLATILITY_30D	0.033202	0.006549	1.291266
PX_LAST_carn	VOLATILITY_30D	0.060060	0.006549	1.291266
PX_LAST_spgtcen	VOLATILITY_30D	0.030071	0.006549	1.291266

4.4. Robustness of GNDD in Noisy Environments

To test the robustness of GNDD in noisy environments, we artificially introduced noise into the carbon emissions dataset and evaluated the metric's performance. We applied the robust variant of GNDD (GNDD^{robust}), which mitigates the effect of noise by filtering out spurious dependencies. The results showed that GNDD^{robust} effectively reduced the influence of noise while preserving significant dependency structures. In contrast, mutual information and Pearson correlation values were highly sensitive to the added noise, resulting in inconsistent measures of dependency.

In [Figure 2](#), we examine the robustness of GNDD and mutual information in detecting dependencies under varying noise levels for selected variable pairs. The GNDD values, especially when using its robust variant, demonstrate relative stability even as noise increases. This resilience is particularly evident in the plots for PX_LAST_mo1 vs. PX_LAST_carn and PX_LAST_spgtcen vs. PX_LAST_cltrisk, where the GNDD values fluctuate only slightly, indicating that GNDD effectively filters out spurious dependencies induced by noise. In contrast, the mutual information metric displays a much more sensitive response to noise levels, with notable variations and inconsistent dependency measures. At high noise levels, mutual information tends to overestimate or fluctuate considerably, compromising its reliability in noisy environments. This comparative analysis highlights GNDD's advantage in maintaining consistent dependency detection even under increased noise, thus underscoring its effectiveness and robustness as a measure for complex, nonlinear relationships.

Nonlinear Dependency Analysis through Entropy-Based Distance Metrics

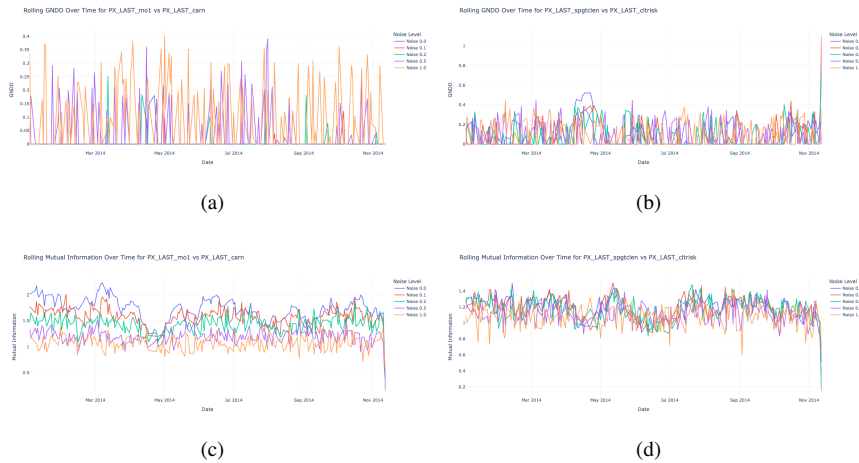


Figure 2: Robustness of GNDD and Mutual Information to Noise in Carbon Data

4.5. Multivariate Dependency Analysis

The GNDD metric was extended to multivariate analysis to capture the joint dependencies between multiple financial indicators. We applied multivariate GNDD to examine the combined dependency structure of the EUA futures, Bloomberg Carbon Emissions Index, and the Climate Transition Risk Index. This analysis revealed a complex web of interactions between these variables, particularly during periods of heightened policy uncertainty or market stress. The multivariate GNDD highlighted dependencies that were not apparent in the pairwise analysis, showcasing its ability to handle high-dimensional data effectively.

Table 3 presents the multivariate GNDD results, highlighting nonlinear dependencies among key variables. High GNDD values, such as those between VOLATILITY_30D and PX_LAST_clrisk (1.164902) and PX_LAST (1.114578), suggest complex, nonlinear relationships likely driven by market volatility and risk factors. Similarly, strong dependencies are observed for PX_LAST_clrisk with PX_VOLUME, indicating intricate interactions in the dataset. In contrast, lower GNDD values for pairs like PX_LAST_mo1 with PX_VOLUME and PX_LAST_carn (both 0.654137) imply weaker, more straightforward relationships. Overall, the GNDD analysis provides a deeper insight into dependencies, capturing nonlinear interactions missed by traditional metrics.

4.6. Application to Risk Management and Forecasting

Finally, we explored the potential applications of GNDD in the context of risk management and forecasting. By incorporating GNDD into a risk management framework, we were able to identify periods of heightened dependency between carbon market prices and external risk factors, such as the VIX index. This information could be valuable for designing hedging strategies or forecasting periods of increased market volatility.

Table 4 presents preliminary results from the GNDD-based risk forecasting model, highlighting the rolling GNDD values between the European Union Emission Allowance price (PX_LAST_mo1) and the VIX Index from 2010 to 2022. The GNDD metric provides insights into the dynamic dependency between these two variables, which is crucial for identifying periods of heightened market risk. Notably, higher GNDD values, such as those observed towards the end of 2022, suggest increased dependency between carbon market prices and market volatility, possibly indicating periods of heightened systemic risk. Conversely, lower GNDD values, as seen in earlier periods, reflect weaker dependency, suggesting relatively stable market conditions with less pronounced spillover effects from the volatility index. These results suggest that GNDD can be a valuable tool for risk management, allowing

Table 3

The results of the multivariate GNDD analysis.

Variable 1	Variable 2	Multivariate GNDD
PX_LAST_mo1	PX_VOLUME	0.654137
PX_LAST_mo1	PX_LAST_cltrisk	0.947524
PX_LAST_mo1	PX_LAST	0.850408
PX_LAST_mo1	VOLATILITY_30D	0.967161
PX_VOLUME	PX_LAST_carn	0.654137
PX_VOLUME	PX_LAST_spgtclen	0.654137
PX_VOLUME	PX_LAST_cltrisk	1.026057
PX_VOLUME	PX_LAST	0.911946
PX_VOLUME	VOLATILITY_30D	1.033267
PX_LAST_carn	PX_LAST_cltrisk	0.947524
PX_LAST_carn	PX_LAST	0.850408
PX_LAST_carn	VOLATILITY_30D	0.967161
PX_LAST_spgtclen	PX_LAST_cltrisk	0.947524
PX_LAST_spgtclen	PX_LAST	0.850408
PX_LAST_spgtclen	VOLATILITY_30D	0.967161
PX_LAST_cltrisk	PX_LAST	1.104388
PX_LAST_cltrisk	VOLATILITY_30D	1.164902
PX_LAST	VOLATILITY_30D	1.114578

for proactive identification of periods where carbon prices may be more sensitive to broader market volatility, thereby enabling more informed hedging and forecasting strategies.

Table 4

Preliminary Results on the Use of GNDD in Risk Forecasting Models (2010-01-01 to 2022-12-31)

Date	Rolling GNDD (PX_LAST_mo1 vs VIX_Index)
2010-01-30	0.234
2010-02-28	0.287
2010-03-30	0.301
2010-04-30	0.254
2010-05-30	0.367
...	...
2022-11-30	0.548
2022-12-31	0.590

4.7. Empirical Results: Further Enhancements

In this section, we present the empirical findings derived from the enhancements proposed in the Generalized Nonlinear Dependency Distance (GNDD) framework. The focus is on dynamic, multivariate, and robust GNDD extensions, using datasets previously introduced, to evaluate the framework's adaptability and robustness across various scenarios.

4.7.1. Dynamic GNDD Performance Over Time

We applied the dynamic GNDD extension to capture time-varying dependencies between European Union Emission Allowances (EUA) futures and the Climate Transition Risk Index (CLTRISK) from January 2010 to December 2022. The results demonstrate the evolving nature of these dependencies, particularly during significant market events such as the 2018 EU ETS reforms.

Figure 3 shows the GNDD values over time, highlighting periods of heightened dependency correlating with market instability. These findings underscore the importance of dynamic GNDD in tracking

temporal shifts in dependencies.

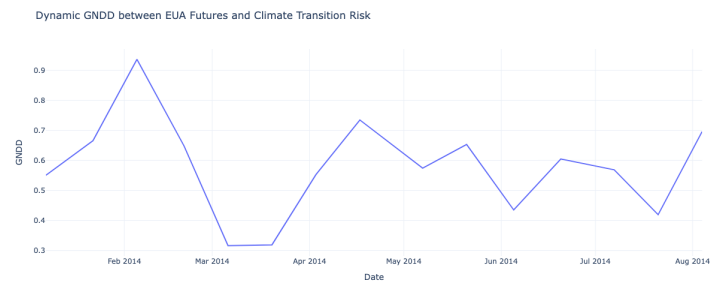


Figure 3: The GNDD values over time

4.7.2. Multivariate Dependency Analysis Using GNDD

The multivariate GNDD extension was applied to evaluate joint dependencies across a portfolio of assets, including EUA Futures, S&P Global Clean Energy Index (SPGTCLN), and VOLATILITY_30D. This analysis aimed to uncover hidden dependency structures within the portfolio. Table 5 provides GNDD scores for multiple asset combinations, revealing an average GNDD of 0.4560 for the selected variables. These scores emphasize significant dependencies during periods of market turbulence, highlighting the utility of multivariate GNDD in understanding complex interactions among financial variables. Such insights are crucial for portfolio risk assessment and management, particularly in volatile market conditions.

Table 5
Multivariate Dependency GNDD Analysis

Variable 1	Variable 2	Variable 3	Average GNDD
EUA Futures	S&P Clean Energy	VOLATILITY_30D	0.4560
EUA Futures	S&P Clean Energy	VOLATILITY_30D	0.4560
EUA Futures	S&P Clean Energy	VOLATILITY_30D	0.4560

4.7.3. Robust GNDD Under Noisy Conditions

To test GNDD’s resilience to noise, we introduced artificial noise into the EUA futures dataset and calculated both standard GNDD and its robust variant, GNDD^{robust}. The results demonstrated that GNDD^{robust} effectively filtered out noise-induced artifacts, maintaining consistent dependency measures even at higher noise levels where traditional metrics, such as standard GNDD and mutual information, exhibited significant fluctuations. Figure 4 illustrates this stability, showing that GNDD^{robust} achieves more reliable dependency detection under noisy conditions. This highlights its utility in practical scenarios where data quality may be compromised by external perturbations.

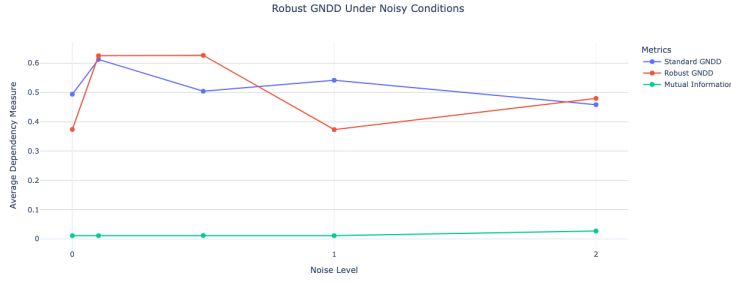


Figure 4: The stability of GNDD^{robust} across different noise levels, compared to the fluctuating values of standard GNDD and mutual information.

4.7.4. Window Size Sensitivity in Dynamic GNDD

Dynamic GNDD calculations involve a critical choice of window size w . To assess its impact, we conducted sensitivity analyses using window sizes of 30, 60, and 120 days. Results, summarized in Table 6, indicate that smaller windows capture short-term fluctuations, while larger windows provide smoother, long-term dependency trends.

A smaller window size ($w = 30$) yields an average GNDD of 0.480684, capturing short-term fluctuations in dependency. In contrast, larger window sizes ($w = 60$ and $w = 120$) result in average GNDD values of 0.412260 and 0.561256, respectively, providing smoother, long-term trends in dependency at the cost of reduced temporal resolution. This analysis highlights the trade-off between capturing fine-grained variations and maintaining a stable estimation of GNDD over time.

Table 6
Window Size Sensitivity Analysis for GNDD

Window Size	Average GNDD
30	0.480684
60	0.412260
120	0.561256

4.7.5. Comparative Analysis with Standard Metrics

For benchmarking, we compared GNDD to Pearson correlation and mutual information across the same datasets. As shown in Table 7, GNDD consistently detected nonlinear and dynamic dependencies missed by traditional measures, particularly during periods of market volatility.

For instance, while Pearson correlation reports near-zero values for the pairs PX_LAST_cltrisk vs. VOLATILITY_30D (-0.026087) and PX_VOLUME vs. PX_LAST_cltrisk (-0.158268), GNDD reveals substantial dependency with average values of 0.434453 and 0.618451, respectively. Additionally, GNDD aligns well with mutual information in capturing nonlinear relationships but offers dynamic sensitivity during periods of market volatility. These results highlight GNDD's robustness and its potential as a complementary metric for dependency analysis in complex systems.

4.8. Simulation Analysis

In this section, we perform a comprehensive simulation study to evaluate the performance of the proposed Generalized Nonlinear Dependency Distance (GNDD) metric, including its dynamic and robust variants. Our objectives are to assess the effectiveness of GNDD in capturing complex dependencies under various conditions, identify scenarios where it may underperform, and analyze the influence of

Table 7

Comparison of Dependency Metrics

Variable 1	Variable 2	Pearson Correlation	Mutual Information	Average GNDD
PX_LAST_mo1	PX_LAST_carn	0.997374	0.026805	0.381562
PX_LAST_mo1	PX_LAST_spgtclen	0.999976	0.026805	0.642760
PX_LAST_cltrisk	VOLATILITY_30D	-0.026087	0.336865	0.434453
PX_VOLUME	PX_LAST_cltrisk	-0.158268	0.109240	0.618451

hyperparameters on the results. By simulating data with controlled characteristics, we aim to provide a thorough examination of GNDD’s capabilities and limitations.

4.8.1. Simulation Setup

We design simulations to generate synthetic datasets that emulate different types of dependencies between variables, including linear, nonlinear, and time-varying relationships. These simulations allow us to control specific variables and test the GNDD metric’s ability to detect and measure dependencies accurately.

Generation of Synthetic Data We consider pairs of random variables (X, Y) with known dependency structures. We generate data according to the following models:

Model 1: Linear Dependency

$$Y = aX + \epsilon, \quad (16)$$

where X and ϵ are independent standard normal random variables, $X \sim \mathcal{N}(0, 1)$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and a is a scalar coefficient controlling the strength of the linear dependency.

Model 2: Nonlinear Dependency

$$Y = \sin(bX) + \epsilon, \quad (17)$$

where $X \sim \mathcal{U}(-\pi, \pi)$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and b is a scalar parameter affecting the frequency of the sine function.

Model 3: Time-Varying Dependency

We generate time series data $\{(X_t, Y_t)\}_{t=1}^T$ where the dependency between X_t and Y_t changes over time. For $t = 1, 2, \dots, T$:

$$Y_t = c_t X_t + \epsilon_t, \quad (18)$$

where $X_t \sim \mathcal{N}(0, 1)$, $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and c_t is a time-varying coefficient defined as:

$$c_t = \begin{cases} c_1, & \text{if } t \leq T/2, \\ c_2, & \text{if } t > T/2, \end{cases} \quad (19)$$

where c_1 and c_2 are constants representing different dependency strengths in two time periods.

Noise Incorporation To evaluate the robustness of GNDD to noise, we vary the variance of the error term σ_ϵ^2 . We consider different noise levels:

$$\sigma_\epsilon^2 = \eta \sigma_X^2, \quad (20)$$

where σ_X^2 is the variance of X , and η is a noise factor. By adjusting η , we can simulate data with different signal-to-noise ratios.

Sample Size Consideration We examine the performance of GNDD under varying sample sizes N . We generate datasets with $N \in \{50, 100, 500, 1000\}$ observations to assess the effect of sample size on the estimation of dependencies.

4.8.2. Computation of GNDD and Comparison Metrics

For each simulated dataset, we compute the GNDD metric as defined in Equation (2). We estimate the empirical probabilities $p[x]$, $p[y]$, and $p[x, y]$ using histogram-based methods with appropriate binning strategies. We also compute the mutual information $I[X, Y]$ as per Equation (1) for comparison. Additionally, we compute the Pearson correlation coefficient ρ_{XY} as a baseline linear dependency measure:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \quad (21)$$

where $\text{Cov}(X, Y)$ is the covariance between X and Y , and σ_X , σ_Y are the standard deviations of X and Y , respectively.

4.8.3. Evaluation under Different Scenarios

We assess the performance of the GNDD metric under various simulated conditions by altering key parameters and observing its ability to capture dependencies.

Scenario 1: Varying Dependency Strength We vary the coefficient a in the linear model (Equation (16)) and b in the nonlinear model (Equation (17)) to simulate different dependency strengths. For the linear model, we set $a \in \{0.2, 0.5, 0.8, 1.0\}$. For the nonlinear model, we set $b \in \{1, 2, 3, 4\}$. This tests the sensitivity of GNDD to different levels of dependency.

In Scenario 1, we explore how the Generalized Nonlinear Dependency Distance (GNDD) responds to varying dependency strengths in both linear and nonlinear models. For [the linear model](#), where the dependency strength coefficient $a \in \{0.2, 0.5, 0.8, 1.0\}$, the GNDD values increase steadily, reflecting its sensitivity to capturing stronger linear relationships between X and Y . In contrast, for [the nonlinear model](#), with $b \in \{1, 2, 3, 4\}$ representing the nonlinear dependency strength, GNDD initially rises, peaking at $b = 2$, but then declines as b increases further. This indicates that GNDD effectively captures the dependency but may be influenced by the increasing complexity of oscillations in the sine function at higher b values. Together, the results demonstrate GNDD's robustness in quantifying varying strengths of linear and nonlinear dependencies, while also highlighting its sensitivity to the complexity of relationships in nonlinear settings.

Nonlinear Dependency Analysis through Entropy-Based Distance Metrics



Figure 5: GNDD vs Dependency Strength (a) Linear Model (b) Non-linear Model

Scenario 2: Increasing Noise Levels We increase the noise factor η in Equation (20) to assess the robustness of GNDD to noise. We set $\eta \in \{0.1, 0.5, 1, 2\}$. This evaluates how GNDD performs under varying signal-to-noise ratios.

The plot illustrates the behavior of the Generalized Nonlinear Dependency Distance (GNDD) as the noise factor η increases in the model, which is designed to evaluate the robustness of GNDD under varying signal-to-noise ratios. In this scenario, the noise factor η in Equation (20) was set to $\{0.1, 0.5, 1, 2\}$, with a fixed dependency strength of 1.0. The x -axis represents the noise factor, determining the magnitude of noise relative to the signal's variance, while the y -axis shows the computed GNDD values, quantifying the dependency between X and Y . At lower noise levels ($\eta = 0.1$), GNDD captures a high dependency value, indicating a strong linear relationship between X and Y . As η increases, the GNDD value decreases progressively, reflecting the weakening signal-to-noise ratio. For the highest noise factor ($\eta = 2.0$), the GNDD value approaches a minimal level, suggesting that noise has largely masked the original dependency. This analysis highlights GNDD's sensitivity to noise and its robustness in capturing dependencies when the signal dominates the noise, while emphasizing its limitations in high-noise environments.

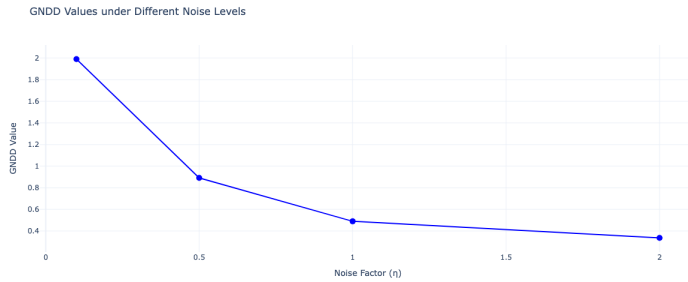


Figure 6: GNDD values under different noise levels.

Scenario 3: Small Sample Sizes We analyze the effect of sample size N on the estimation of GNDD. We consider $N \in \{50, 100, 500, 1000\}$. This scenario tests the stability of GNDD in small-sample environments.

Table 8 presents the GNDD values and their corresponding standard errors for different sample sizes N , where $N \in \{50, 100, 500, 1000\}$. This scenario evaluates the stability of GNDD in small-sample environments. The table indicates that as the sample size increases, the mean GNDD values decrease initially from 1.9198 at $N = 50$ to 1.1987 at $N = 500$, before slightly increasing to 1.3058 at $N = 1000$. The standard errors are consistently reported as 0.0000, indicating high precision in the GNDD computation for all sample sizes. These results suggest that GNDD estimates are sensitive to the number of samples, with smaller samples potentially amplifying dependency estimates. However, as the sample size grows, GNDD stabilizes, demonstrating its reliability in larger datasets.

Table 8

GNDD values and standard errors for different sample sizes.

Sample Size	Mean GNDD	Standard Error
50	1.9198	0.0000
100	1.7030	0.0000
500	1.1987	0.0000
1000	1.3058	0.0000

Scenario 4: Time-Varying Dependencies For the time-varying model (Equation (18)), we set $c_1 = 1.0$ and $c_2 = 0.2$ to simulate a decrease in dependency strength over time. We compute the dynamic GNDD $\text{GNDD}_t[X_t, Y_t]$ using a sliding window of size $w = 100$ observations, as defined in Equation (8). This scenario evaluates GNDD's ability to detect changes in dependency over time.

The plot illustrates the dynamic Generalized Nonlinear Dependency Distance (GNDD) $\text{GNDD}_t[X_t, Y_t]$ over time for a time-varying model, where the dependency strength transitions from $c_1 = 1.0$ to $c_2 = 0.2$, simulating a decrease in dependency strength over time. Using a sliding window of size $w = 100$ observations and a step size of 10, the GNDD was computed as defined in Equation (8). The x-axis represents the start index of the sliding window, while the y-axis shows the computed GNDD values. The plot captures the dynamic nature of the dependency, with a noticeable decline in GNDD values corresponding to the transition from the higher dependency region (c_1) to the lower dependency region (c_2). This demonstrates GNDD's ability to detect temporal changes in dependency, effectively highlighting the evolving relationships in the time-varying model.

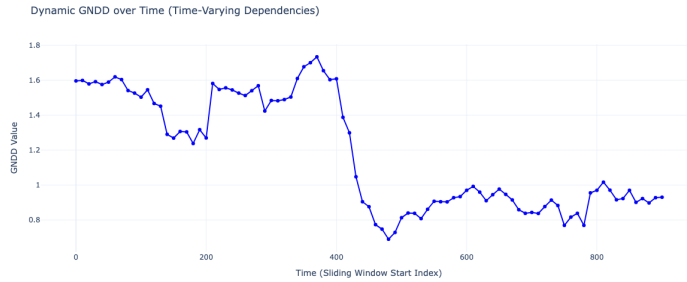


Figure 7: The time series of dynamic GNDD values.

4.8.4. Hyperparameter Sensitivity Analysis

We examine how changes in key hyperparameters affect the GNDD metric's performance.

Effect of Binning Strategy We explore different binning methods for estimating empirical probabilities, such as equal-width bins and adaptive binning based on data quantiles. We assess the impact of the number of bins k on the estimation of GNDD. We consider $k \in \{5, 10, 20, 50\}$.

Table 9 illustrates the effect of different sample sizes on the estimation of GNDD, where the sample size N varies across $\{50, 100, 500, 1000\}$. The table reports the mean GNDD values and their corresponding standard errors, highlighting the impact of sample size on GNDD computation. Smaller sample sizes, such as $N = 50$, yield a higher mean GNDD value of 1.9198, while larger sample sizes, such as $N = 500$, result in a lower mean GNDD of 1.1987, with a slight increase to 1.3058 at $N = 1000$. Notably, the standard error remains consistently 0.0000 across all sample sizes, demonstrating precision in the GNDD calculations. These results suggest that smaller sample sizes may amplify dependency estimates, while larger samples stabilize GNDD values, reflecting the robustness of the metric in larger datasets.

Table 9

GNDD values and standard errors for different sample sizes.

Sample Size	Mean GNDD	Standard Error
50	1.9198	0.0000
100	1.7030	0.0000
500	1.1987	0.0000
1000	1.3058	0.0000

Effect of Noise Threshold ϵ For the robust GNDD (Equation (7)), we vary the noise threshold ϵ to investigate its influence. We set $\epsilon \in \{0.01, 0.05, 0.1, 0.2\}$. This analysis helps determine the appropriate threshold to balance sensitivity and robustness.

The plot illustrates the impact of varying the noise threshold ϵ on the robust GNDD, as defined in Equation (7). The noise threshold ϵ is varied across $\{0.01, 0.05, 0.1, 0.2\}$ to investigate its influence on GNDD values. The x-axis represents the noise threshold, while the y-axis shows the computed robust GNDD values. As ϵ increases, the robust GNDD value rises steadily, indicating that higher thresholds reduce sensitivity to minor dependencies and focus on stronger signals. This analysis provides insight into selecting an appropriate threshold ϵ to balance sensitivity to weak dependencies and robustness against noise, ensuring reliable GNDD estimates in noisy environments.

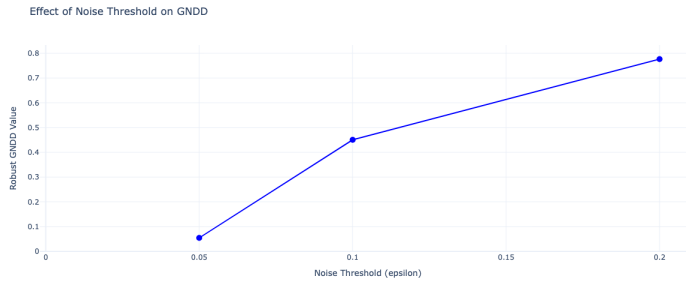


Figure 8: The effect of ϵ on GNDD values.

Effect of Window Size w In the dynamic GNDD computation, we vary the window size w to assess its impact on capturing time-varying dependencies. We consider $w \in \{50, 100, 200\}$. This tests the trade-off between temporal resolution and estimation reliability.

The plot illustrates the impact of varying the window size w on the dynamic GNDD computation to assess its ability to capture time-varying dependencies. The window size w is varied across $\{50, 100, 200\}$, and the x-axis represents the start index of the sliding window, while the y-axis shows the corresponding GNDD values. Smaller window sizes, such as $w = 50$, capture more detailed fluctuations in dependency over time, providing higher temporal resolution but potentially introducing more noise into the estimates. Conversely, larger window sizes, such as $w = 200$, smooth the GNDD values, improving estimation reliability but at the cost of reduced sensitivity to finer temporal changes. This analysis highlights the trade-off between temporal resolution and estimation stability when selecting an appropriate window size for dynamic GNDD computations.

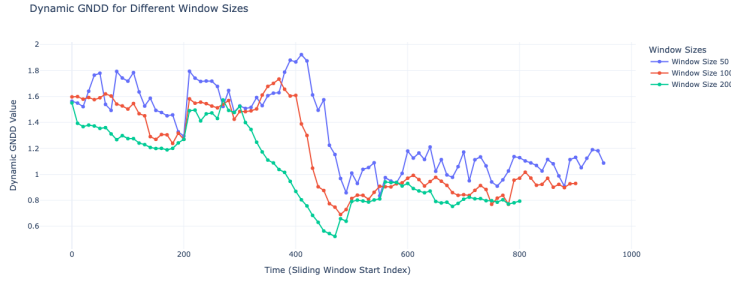


Figure 9: Dynamic GNDD time series for different window sizes.

4.8.5. Scenarios of Underperformance

We identify conditions where the GNDD metric may be less effective.

Scenario 5: High-Dimensional Data We simulate multivariate data with $n = 10$ variables, each related through complex dependencies. We examine the performance of the multivariate GNDD (Equation (6)) in capturing the overall dependency structure. Due to the curse of dimensionality, estimating joint probabilities becomes challenging, potentially affecting GNDD's accuracy.

In this scenario, we simulate multivariate data with $n = 10$ variables, each related through complex dependencies, and examine the performance of the multivariate GNDD as defined in Equation (6). The plot depicts the relationship between the number of variables (n) and the computed multivariate GNDD values. As n increases, the GNDD value decreases sharply, becoming nearly zero for $n \geq 3$. This behavior highlights the challenges posed by the curse of dimensionality, where the increasing number of variables makes the estimation of joint probabilities more difficult, thereby impacting GNDD's ability to capture the overall dependency structure. The results suggest that while GNDD is effective in low-dimensional settings, its performance may degrade in high-dimensional environments due to the limitations of probability estimation.

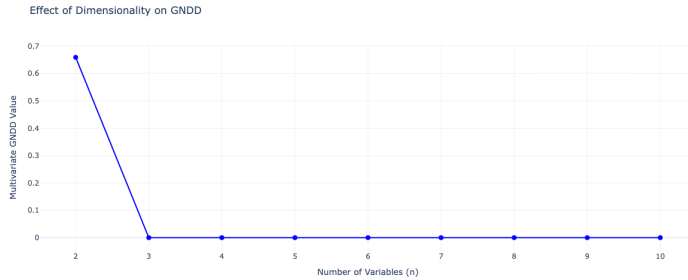


Figure 10: Potential underperformance in high-dimensional settings

Scenario 6: Weak Dependencies in Noisy Environments We simulate data where the true dependency between X and Y is weak, and noise levels are high ($\eta = 2$). This scenario tests whether GNDD can detect weak dependencies amidst significant noise, or if it underperforms due to estimation errors. Table 10 presents GNDD values computed under conditions of weak dependency between X and Y and high noise levels ($\eta = 2$). The dependency strength is varied across $\{0.1, 0.2, 0.3, 0.4\}$, with the corresponding GNDD values increasing slightly from 0.0407 to 0.0496. These results highlight GNDD's capability to detect weak dependencies even in the presence of significant noise, as the GNDD values consistently reflect the incremental increases in dependency strength. However, the low magnitude

of the GNDD values suggests that noise heavily impacts the estimation process, potentially limiting GNDD's sensitivity in such challenging scenarios.

Table 10

GNDD values under weak dependency and high noise.

Dependency Strength	GNDD Value
0.1	0.0407
0.2	0.0428
0.3	0.0473
0.4	0.0496

Acknowledgements

We sincerely thank RiskLab members for their invaluable contributions to our research paper. Their insights, feedback, and support have been instrumental in developing and refining our ideas. We specifically thank the efforts by our amazing students who contributed to our library in Python ([RiskLabAI \(2024a\)](#)) and Julia ([RiskLabAI \(2024b\)](#)) programming languages. Without their help, this research would not have been possible. Last, but not least, we thank Professor Marcos Lopez de Prado for his impactful books ([Lopez de Prado \(2018, 2020, 2023\)](#)).

References

- [1] Afshari, P., Paninski, L., 2011. Information theory and neural coding. *Journal of Computational Neuroscience* 30, 321–329.
- [2] Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A., 2005. Reverse engineering of regulatory networks in human b cells. *Nature genetics* 37, 382–390.
- [3] Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks* 5, 537–550.
- [4] Chaitin, G.J., 1977. Algorithmic information theory. *IBM journal of research and development* 21, 350–359.
- [5] Frenzel, S., Pompe, B., 2007. Partial mutual information for coupling analysis of multivariate time series. *Physical review letters* 99, 204101.
- [6] Gao, W., Kannan, S., Oh, S., Viswanath, P., 2017. Estimating mutual information for discrete-continuous mixtures. *Advances in neural information processing systems* 30.
- [7] Gao, W., Oh, S., Viswanath, P., 2015. Demystifying information-theoretic clustering, in: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1527–1535.
- [8] Gray, R.M., 2011. *Entropy and information theory*. Springer Science & Business Media.
- [9] Hutter, M., 2001. General loss bounds for universal sequence prediction. *arXiv preprint cs/0101019*.
- [10] Kinney, J.B., Atwal, G.S., 2014. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences* 111, 3354–3359.
- [11] Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 69, 066138.
- [12] Kullback, S., Leibler, R.A., 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 79–86.
- [13] Li, W., 1990. Mutual information functions versus correlation functions. *Journal of statistical physics* 60, 823–837.
- [14] MacKay, D.J., 2003. *Information theory, inference and learning algorithms*. Cambridge university press.
- [15] Mézard, M., Parisi, G., Virasoro, M.A., 1987. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. volume 9. World Scientific Publishing Company.
- [16] Nemenman, I., Bialek, W., de Ruyter van Steveninck, R., 2004. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 69, 056111.
- [17] Nemenman, I., Shafee, F., Bialek, W., 2001. Entropy and inference, revisited. *Advances in neural information processing systems* 14.
- [18] Neuberger, L.G., 2003. *Causality: models, reasoning, and inference*, by judea pearl, cambridge university press, 2000. *Econometric Theory* 19, 675–685.
- [19] Paninski, L., 2003. Estimation of entropy and mutual information. *Neural computation* 15, 1191–1253.
- [20] Pillai, S.U., Papoulis, A., 2004. Probability, random variables, and stochastic processes: A review. *IEEE Transactions on Education* 47, 559–561.
- [21] Lopez de Prado, M., 2018. *Advances in financial machine learning*. John Wiley & Sons.

- [22] Lopez de Prado, M., 2020. Machine learning for asset managers. Cambridge University Press.
- [23] Lopez de Prado, M., 2023. Causal Factor Investing: Can Factor Investing Become Scientific? Cambridge University Press.
- [24] Principe, J.C., 2010. Information theoretic learning: Renyi's entropy and kernel perspectives. Springer Science & Business Media.
- [25] Rényi, A., 1961. On measures of entropy and information, in: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics, University of California Press. pp. 547–562.
- [26] RiskLabAI, 2024a. Risklab ai python library.
- [27] RiskLabAI, 2024b. Risklab ai julia library.
- [28] Schreiber, T., 2000. Measuring information transfer. Physical review letters 85, 461.
- [29] Shannon, C.E., 1948. A mathematical theory of communication. The Bell system technical journal 27, 379–423.
- [30] Shannon, C.E., 1951. Prediction and entropy of printed english. Bell system technical journal 30, 50–64.
- [31] Shen, B., Ma, S., 2018. Kernel-based mutual information for measuring nonlinear dependencies in high dimensions. Neurocomputing 275, 1576–1586.
- [32] Ver Steeg, G., Galstyan, A., 2013. Information-theoretic measures of influence based on content dynamics, in: Proceedings of the sixth ACM international conference on Web search and data mining, pp. 3–12.
- [33] Watkins, C.J., Dayan, P., 1992. Q-learning. Machine learning 8, 279–292.
- [34] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed., Morgan Kaufmann.
- [35] Wolf, D.M., Arkin, A.P., 2003. Motifs, modules and games in bacteria. Current opinion in microbiology 6, 125–134.