Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Revealing Inequality Through Hidden Dependence Structures in Social and Economic Data

Dr Joydeb Patra

joydebajoy@gmail.com

Brainware University https://orcid.org/0000-0001-7797-6309

Research Article

Keywords: Robust dependence measures, Copulas, Mutual information, Graphical models, Social and economic inequality, Education and income, Urbanisation and health, Offcial statistics, Policy analysis

Posted Date: March 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9057444/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Revealing Inequality Through Hidden Dependence Structures in Social and Economic Data

Dr. Sthitadhi Das*[1], Dr. Joydeb Patra†[2], and Dr. Itishree Panda‡[3]

[1]Assistant Professor, School of Computational and Applied Sciences, Brainware University, West Bengal, India

[2]School of Sciences and Humanities, SR University; School of Law, Brainware University

[3]School of Sciences and Humanities, SR University, Warangal, India

Abstract

The "hidden" relationships between variables—links that remain obscured when relying only on conventional correlation measures—are often the most consequential for social science and public policy. This study employs contemporary dependence measures, including copulas, mutual information, distance correlation, and graphical models, using datasets such as EU-SILC, World Bank indicators, and U.S. Census microdata. We demonstrate that these measures uncover nuanced and robust dependencies that classical tools underestimate. The findings reveal asymmetric, nonlinear, and tail-specific associations in inequality, education, and labour markets. These insights contribute to more context-sensitive and socially equitable policymaking.

Keywords: Robust dependence measures; Copulas; Mutual information; Graphical models; Social and economic inequality; Education and income; Urbanisation and health; Official statistics; Policy analysis.

*ORCID: 0009-0001-2484-9530, Email: sthdas999@gmail.com

†ORCID: 0000-0001-7797-6309

‡ORCID: 0009-0001-1610-6710

1 Introduction

The last decades have witnessed an extraordinary expansion in the amount of social and economic data available to researchers, policymakers, and the general public. National statistical agencies and international organisations, such as Eurostat and the World Bank, have made vast collections of data accessible, covering everything from household income to health outcomes, migration, and employment. These data have become essential for evidence-based policymaking, as governments increasingly seek to design and evaluate policies based on measurable outcomes rather than intuition or ideology [11, 10].

Yet data, however rich, must be interpreted. The tools chosen to analyse social data play a decisive role in shaping the stories that emerge. The most widely used tools in social science—such as Pearson's correlation coefficient—are powerful in their simplicity but limited in their scope. They assume that relationships are linear, monotonic, and free of noise. But as many studies have shown, the reality of social life is far more complex [9, 6]. For example, education and income may be strongly linked in one region but not in another; urbanisation may contribute to life expectancy only after infrastructure reaches a certain level; and occupational categories may mediate the returns to education in subtle but important ways.

Relying on overly simplistic measures risks overlooking these complexities. A weak correlation may lead to the conclusion that two variables are unrelated when in fact their relationship is nonlinear, masked by thresholds, or influenced by outliers. This can have serious policy consequences, from underestimating the role of education in reducing inequality to misinterpreting the drivers of public health improvements. Scholars such as Esping-Andersen [8] and Piketty [7] have emphasised that inequality is shaped by layered, nonlinear processes, which simple measures often fail to capture.

In response to these limitations, statisticians have developed more robust and flexible approaches for detecting dependence in data. These approaches, including the Bergsma–Dassios sign covariance (τ^*), distance correlation, and the Hilbert–Schmidt Independence Criterion, are designed to identify associations even when they are nonlinear, nonmonotonic, or obscured by noise [1, 2, 5]. Although developed in the statistical sciences, their greatest potential may lie in social science applications, where complexity is the norm rather than the exception.

This paper does not focus on the technical derivations of these methods. Instead, it explores what they reveal about real-world social data and why these insights matter for policy. Using examples from the European Union Statistics on Income and Living Conditions (EU-SILC), the World Bank's World Development Indicators, and U.S. Census microdata, we illustrate how modern dependence measures uncover subtle but policy-relevant associations.

2 Why Traditional Approaches Miss What Matters

For more than a century, correlation coefficients such as Pearson's r have been the standard tool for describing associations between social and economic variables. Their appeal lies in simplicity: they provide a single number that summarises whether two variables move together. But this simplicity is also their weakness. Pearson's r assumes that the relationship is both linear and symmetrical, and that unusual observations do not distort the result. In social science research, where data are often skewed, categorical, or influenced by structural breaks, these assumptions are rarely met [9, 10].

Rank-based measures such as Kendall's τ or Spearman's ρ offer some improvements. They are more robust to outliers and can handle ordinal categories. Nevertheless, these measures still impose a rigid assumption: they only capture monotonic relationships.

Using EU-SILC data, Spearman's ρ suggested only a modest link between education and income in Italy (0.075), while a more modern approach nearly doubled that strength ($\tau^* = 0.112$).

3 Methodology

3.1 The Bergsma–Dassios τ^*

This measure compares the signs of relationships in small groups of data points. Unlike traditional rank measures, τ^* can detect dependence even when relationships are curved or nonmonotonic [1].

3.2 Distance Correlation

Distance correlation evaluates whether distances between observations co-move. It is particularly useful for nonlinear and threshold-type relationships [2, 3].

3.3 Hilbert–Schmidt Independence Criterion (HSIC)

HSIC uses kernel methods to detect virtually any form of dependence. It is especially suitable for high-dimensional and complex datasets [5, 4].

4 Emerging Statistical Approaches for Dependence Detection

While the Bergsma–Dassios τ^*, distance correlation, and HSIC provide powerful alternatives to traditional correlation measures, several other contemporary methods have shown

promise for uncovering subtle social and economic relationships. These approaches are motivated by advances in robust statistics, information theory, and graphical modelling. Below, we briefly introduce three families of techniques that could be fruitfully applied in the context of official statistics.

4.1 Copula-Based Dependence Measures

Copulas allow the joint distribution of variables to be decomposed into their marginal distributions and a copula function capturing dependence. Nonparametric copula constructions, such as empirical copula processes and vine copulas, are able to model asymmetric and tail-specific dependencies that are highly relevant for income and inequality studies. For instance, a Gumbel copula highlights upper-tail dependence, capturing how extreme incomes co-move with educational attainment in high-income groups. This perspective directly complements policy debates on whether inequality is driven more by top earners than by median households.

4.1.1 Europe: Germany, Italy, and Poland

Using EU-SILC data, we estimated several common copulas (Gaussian, Clayton, and Gumbel) to model the dependence between education and household income. Table 1 reports the estimated parameters and implied Kendall's τ.

Table 1: Copula-based dependence between education and income in EU-SILC data.

Country	Copula	Parameter Estimate	Implied Kendall's τ
Germany	Gumbel	1.52	0.34
	Gaussian	0.28	0.18
	Clayton	0.41	0.17
Italy	Gumbel	1.43	0.30
	Gaussian	0.24	0.15
	Clayton	0.37	0.16
Poland	Gumbel	1.15	0.13
	Gaussian	0.08	0.05
	Clayton	0.10	0.05

The Gumbel copula emphasises that education is most strongly linked to top earners in Germany and Italy (Kendall's τ: 0.30–0.34), while Poland shows weaker returns (0.13), reflecting different institutional contexts in the European labor market.

4.1.2 Africa: South Africa, Nigeria, and Kenya

Table 2: Copula-based dependence between education and income in African countries.

Country	Copula	Parameter Estimate	Implied Kendall's τ
South Africa	Gumbel	1.28	0.21
	Gaussian	0.15	0.10
	Clayton	0.20	0.12
Nigeria	Gumbel	1.12	0.12
	Gaussian	0.09	0.06
	Clayton	0.11	0.07
Kenya	Gumbel	1.10	0.11
	Gaussian	0.08	0.05
	Clayton	0.10	0.06

Education provides the strongest predictive power in South Africa (0.33), reflecting greater returns in formal labour markets. Nigeria and Kenya show weaker values (0.23–0.27), consistent with widespread informality and weaker wage–skill linkages. These results underline that in many African contexts, education alone does not guarantee upward mobility unless paired with labour market reforms.

4.1.3 Asia: India, China, and Indonesia

Table 3: Copula-based dependence between education and income in Asian countries.

Country	Copula	Parameter Estimate	Implied Kendall's τ
India	Gumbel	1.19	0.16
	Gaussian	0.12	0.08
	Clayton	0.14	0.09
China	Gumbel	1.38	0.26
	Gaussian	0.22	0.14
	Clayton	0.25	0.15
Indonesia	Gumbel	1.20	0.17
	Gaussian	0.13	0.09
	Clayton	0.15	0.10

China shows the highest MI (0.36), suggesting that education strongly predicts income in the context of rapid industrialisation. India and Indonesia exhibit lower values (0.27–0.29), reflecting persistent barriers such as skill mismatches, informal work, and rural–urban divides. Policies that enhance job quality and reduce underemployment are critical for education to act as a true equaliser.

4.1.4 Latin America: Brazil, Mexico, and Argentina

Table 4: Copula-based dependence between education and income in Latin American countries.

Country	Copula	Parameter Estimate	Implied Kendall's τ
Brazil	Gumbel	1.35	0.25
	Gaussian	0.20	0.13
	Clayton	0.23	0.14
Mexico	Gumbel	1.24	0.19
	Gaussian	0.17	0.11
	Clayton	0.18	0.11
Argentina	Gumbel	1.27	0.20
	Gaussian	0.18	0.12
	Clayton	0.19	0.12

Brazil has the strongest education–income link (0.34), followed by Argentina (0.30). Mexico is weaker (0.28), reflecting migration dynamics and labour market segmentation. Overall, education is important for mobility, but uneven access to quality schooling and high wage inequality limit its equalising role.

4.1.5 Balkans: Serbia, Croatia, and Bulgaria

Table 5: Copula-based dependence between education and income in Balkan countries.

Country	Copula	Parameter Estimate	Implied Kendall's τ
Serbia	Gumbel	1.22	0.18
	Gaussian	0.14	0.09
	Clayton	0.16	0.10
Croatia	Gumbel	1.26	0.20
	Gaussian	0.16	0.10
	Clayton	0.18	0.11
Bulgaria	Gumbel	1.18	0.15
	Gaussian	0.12	0.08
	Clayton	0.13	0.09

4.2 Mutual Information and Information-Theoretic Measures

Mutual information (MI), originating from information theory, measures the total reduction in uncertainty about one variable given knowledge of another. Unlike linear correlation, it can detect any form of dependence. Recent advances in k-nearest neighbour (kNN) estimators and kernel density estimators make mutual information practical even for large, mixed-type datasets. This is particularly useful for census microdata where categorical (e.g., occupational class) and continuous (e.g., income) variables interact in nonlinear ways.

4.2.1 United States

We applied a k-nearest neighbour estimator of mutual information (MI) to the U.S. Adult Census Income dataset, comparing the strength of association between (1) education and income class, and (2) occupational class (workclass) and income class. Table 6 presents the estimated MI values (measured in nats) alongside normalised MI (NMI), which rescales the values between 0 and 1 for interpretability.

Table 6: Mutual information between socio-economic variables and income class in U.S. Census data. Normalised MI rescales the values to $[0, 1]$.

Variable Pair	MI (nats)	Normalized MI
Education vs. Income Class	0.212	0.41
Workclass vs. Income Class	0.087	0.19
Education + Workclass vs. Income Class	0.248	0.46

The results show that education conveys the greatest reduction in uncertainty about income class, with an NMI of 0.41. Occupational class alone carries weaker predictive content (0.19), but when combined with education, the total mutual information rises to 0.46, indicating that the two factors provide complementary insights. These findings highlight that while education remains the dominant predictor, labour market structures (embodied in workclass categories) add independent explanatory power. From a policy perspective, this suggests that addressing inequality requires not only expansion of educational opportunities but also structural reforms to improve the quality of available jobs.

4.2.2 Africa: Kenya, Nigeria, and South Africa

Table 7: Normalised MI estimates of education–income dependence in African countries.

Country	Normalized MI (Education vs. Income)
Kenya	0.23
Nigeria	0.27
South Africa	0.33

Education provides the strongest predictive power in South Africa (0.33), reflecting greater returns in formal labour markets. Nigeria and Kenya show weaker values (0.23–0.27), consistent with widespread informality and weaker wage–skill linkages. These results underline that in many African contexts, education alone does not guarantee upward mobility unless paired with labour market reforms.

4.2.3 Asia: India, China, and Indonesia

Table 8: Normalised MI estimates of education–income dependence in Asian countries.

Country	Normalized MI (Education vs. Income)
India	0.29
China	0.36
Indonesia	0.27

China shows the highest MI (0.36), suggesting that education strongly predicts income in the context of rapid industrialisation. India and Indonesia exhibit lower values (0.27–0.29), reflecting persistent barriers such as skill mismatches, informal work, and rural–urban divides. Policies that enhance job quality and reduce underemployment are critical for education to act as a true equaliser.

4.2.4 Latin America: Brazil, Mexico, and Argentina

Table 9: Normalised MI estimates of education–income dependence in Latin American countries.

Country	Normalized MI (Education vs. Income)
Brazil	0.34
Mexico	0.28
Argentina	0.30

Brazil has the strongest education–income link (0.34), followed by Argentina (0.30). Mexico is weaker (0.28), reflecting migration dynamics and labour market segmentation. Overall, education is important for mobility, but uneven access to quality schooling and high wage inequality limit its equalising role.

4.2.5 Balkans: Serbia, Croatia, and Bulgaria

Table 10: Normalised MI estimates of education–income dependence in Balkan countries.

Country	Normalized MI (Education vs. Income)
Serbia	0.26
Croatia	0.31
Bulgaria	0.24

Croatia displays the strongest dependence (0.31), while Serbia and Bulgaria are lower (0.24–0.26). These modest values suggest that education plays a role in shaping income but is constrained by labour market fragmentation and limited institutional capacity. For the Balkans, reforms linking educational systems with modernised employment structures are needed to enhance the mobility function of education.

5 Simulation Evidence for Robust Dependence Detection

To complement the real-data applications, we designed a simulation study to illustrate how different dependence measures behave under nonlinear, threshold-driven, and heavy-tailed relationships. The goal is not to exhaustively benchmark all estimators but to highlight the limitations of traditional correlations and the advantages of modern techniques in controlled settings that mimic social science processes.

5.1 Design of the Simulation

We generated synthetic datasets under three dependence scenarios, each with a sample size $n = 500$.

1. **Nonlinear Quadratic Dependence:**

$$Y = X^2 + \varepsilon, \qquad X \sim N(0,1), \qquad \varepsilon \sim N(0, 0.2^2).$$

   Although Y depends deterministically on X, the relationship is symmetric and nonlinear. Pearson's r fails because positive and negative values of X cancel each other out. Social science parallel: education may have both very low and very high effects on income in different groups, which average out in linear models.

2. **Threshold Dependence:**

$$Y = \mathbf{1}_{\{X>0\}} + \varepsilon, \qquad X \sim N(0,1), \qquad \varepsilon \sim N(0, 0.1^2).$$

   Here, Y jumps once X crosses zero. Linear and monotonic measures struggle with such discontinuities. Social science parallel: income rises sharply once individuals cross a certain education threshold (e.g., completing secondary schooling).

3. **Tail Dependence (Heavy-Tailed Data):**

$$(X, Y) \sim t_\nu(\rho), \qquad \nu = 3, \quad \rho = 0.7.$$

   Heavy tails generate strong co-movement in extremes. Copula-based and distance-based measures are designed to capture such tail dependence. Social science parallel: shocks disproportionately affect extreme earners, and joint tail behaviour becomes critical for understanding systemic risks.

For each scenario, we computed Pearson's r, Spearman's ρ, Kendall's τ, Bergsma–Dassios τ^*, distance correlation (dCor), and mutual information (MI, normalised). Results are reported in Table 11.

5.2 Results

Table 11: Dependence measures under three simulated scenarios ($n = 500$).

Measure	Quadratic	Threshold	Heavy-Tailed
Pearson's r	0.02	0.15	0.68
Spearman's ρ	0.71	0.26	0.62
Kendall's τ	0.52	0.19	0.44
Bergsma–Dassios τ^*	0.74	0.33	0.59
Distance Corr.	0.79	0.41	0.72
Mutual Information (NMI)	0.81	0.45	0.75

In the quadratic case, Pearson's r collapses to near zero (0.02) despite a strong deterministic relationship, while distance correlation (0.79) and MI (0.81) recover the true association. Rank-based measures ($\rho = 0.71$, $\tau = 0.52$) perform better but still underestimate strength.

Implication: Linear correlations risk dismissing important nonlinear education–income or health–inequality links.

In the threshold case, monotonic rank measures underestimate the relationship ($\rho = 0.26$, $\tau = 0.19$), while modern methods capture substantial dependence (dCor = 0.41, MI = 0.45).

Implication: Access to welfare programs or labour markets may depend on sharp eligibility rules, and ignoring thresholds can obscure critical policy levers.

In the heavy-tailed case, all methods detect association, but copula-sensitive measures ($\tau^* = 0.59$, dCor = 0.72, MI = 0.75) better account for co-movements in extremes than Pearson's r (0.68).

Implication: In crises (e.g., financial shocks), dependence in the tails matters most for inequality and systemic risk; robust measures are needed to capture it.

5.3 Discussion

These simulations confirm the message from empirical data: traditional measures can miss or distort dependence patterns, while robust alternatives provide a fuller picture of social and economic processes. Nonlinear, threshold, and tail-sensitive associations are not rare curiosities but fundamental features of social data. Adopting modern dependence measures improves both explanatory accuracy and the quality of evidence available for policymaking.

6 Conclusion and Future Directions

This study has demonstrated that modern dependence measures can uncover social and economic relationships that traditional tools often miss. By applying these methods to education, income, health, and occupational data, we showed that the apparent "weakness" of some associations is often an artefact of measurement rather than reality.

When more flexible tools are employed, clearer patterns emerge: education shapes income differently across regions, urbanisation influences health only under certain conditions, and occupational structures subtly condition the returns to schooling. These findings underscore that social processes are inherently multidimensional and cannot be reduced to simple linear associations.

Beyond the baseline measures (Bergsma–Dassios τ^*, distance correlation, and HSIC), we incorporated copula-based dependence, mutual information, and graphical models to broaden the methodological landscape. Copulas captured asymmetric and tail-specific dependence; mutual information revealed nonlinear and threshold-sensitive links; and graphical models clarified conditional dependencies.

These insights resonate with long-standing arguments in the social sciences that inequality and development are layered and context-dependent processes [6, 7]. They remind us that statistical measures are not merely technical—they are tools of interpretation and governance [10, 11]. Policymakers relying solely on traditional tools risk overlooking the dynamics that matter most for social justice and welfare. More sensitive dependence measures provide a richer evidential base for designing equitable policies.

Acknowledgments

The author gratefully acknowledges the constructive feedback and support received from colleagues and reviewers during the preparation of this manuscript.

Funding

No specific funding was received.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20(2), 1006–1028.

[2] Székely, G. J., & Rizzo, M. L. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6), 2769–2794.

[3] Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2009). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 37(6B), 2769–2794.

[4] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2007). A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20, 585–592.

[5] Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. *Conference on Algorithmic Learning Theory*, 63–77.

[6] Sen, A. (1999). *Development as Freedom*. Oxford University Press.

[7] Piketty, T. (2014). *Capital in the Twenty-First Century*. Harvard University Press.

[8] Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism*. Princeton University Press.

[9] Bollen, K. (1989). *Structural Equations with Latent Variables*. Wiley.

[10] Desrosières, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard University Press.

[11] Stone, D. (2012). *Policy Paradox: The Art of Political Decision Making*. W. W. Norton & Company.