Comparing Granger Causality and Transfer Entropy

Supervised by professor Russel Steele

Mariel Yacolca Maguiña

Fall 2023

I. Introduction

The research project titled Granger Causality and Transfer Entropy Tests took place during the fall 2023 semester under the supervision of Professor Russell Steele. The goal of the project was to evaluate the performance of Granger Causality and Transfer Entropy to detect causality between time series as shown in the paper NlinTS: An R Package For Causality Detection in Time Series by Youssef Hmamouche.

Part II. Literature Review provides the theoretical background necessary to understand time series, time series models, and causality tests. Part III. Methodology, explains the methodology of the author to construct the Granger Causality and Transfer Entropy test. Part IV. Simulations explains the tests and results conducted. Part V. Discussion provides comments on the results and possible caveats. Finally, Part VI. Appendix provides all the figures and graphs that I will be referring to in parts IV. and V.

The R code for this project can be found in the GitHub repository.

II. Literature Review

The Basics.

A time-series model explains the value of a variable at time t based on the value of the variable at $t-1, t-2, ..., t-k$. The main reason researchers use time-series models is forecasting. We assume that the variable in question is a random variable and that the point forecast is given by $E[x_t]$ and that the prediction interval is a set of values that x_t can take with high probability.

Some notation: If we forecast y_t using information I, then we write $y_t|I$. The set of values y_t can take is the forecast distribution $Y_t|I$. The estimate of y_t is $\hat{y_t}$, meaning the average possible values of y_t. We write $y_{T+n|T}$ meaning the forecast of Y_{T+n} given information from $t = 0$ to $t = T$.

Features of time series data.

A trend is a long-term increase or decrease in the data. Seasonality refers to the effect of the time of the year or the day of the week in a time series. A cycle refers to data exhibiting rises and falls that are not of fixed frequency.

Since time-series data points depend on each other, we are able to calculate the auto-correlation coefficient

$$r_k = \frac{\sum(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum(y_t - \bar{y})^2} \tag{1}$$

where T is the length of the time series and k is the number of lags. When there is a trend, the auto-correlation for small lags is large and positive, decreasing over time. When data is seasonal, the auto-correlation will be larger for the seasonal lags. It is possible to see a combination of both given a trend and seasonality.

A moving average of m order is given by $\hat{T}_t = \frac{1}{m}\sum y_{t+j}$. A common use of moving averages is to separate the trend cycle from seasonal data. The first step is to compute the trend component $\hat{T}_t$. The second step is to calculate the de-trended series $y_t - \hat{T}_t$. The third step is to estimate the average of each month (day or hour) for the de-trended series and adjust it to ensure that it adds up to 0. We call this $\hat{s}_t$. Finally, the remainder component is given by $\hat{R}_t = y_t - \hat{T}_t - \hat{s}_t$. To perform a multiplicative decomposition we divide instead of subtract in the set of instructions above.

Benchmark Models.

To conduct forecasts, there are a variety of methods available:

1. Mean method: $\hat{y}_{t+n|T} = \bar{y}$

2. Naive method: $\hat{y}_{t+n|T} = y_T$
3. Seasonal naive method: The next value is equal to the last value of the same season.
4. Drift method: Consider the trend from the first to the last observation.

These methods are mostly used for benchmarks to which we will compare more complex models. We will only consider other methods if they are better than these benchmark methods. Regarding residuals $e_t = y_t - \hat{y_t}$, a good model will have uncorrelated residuals with mean 0. It is useful if the residuals have constant variance (homoskedasticity) and are normally distributed. We can produce a prediction interval using the formula $\hat{y}_{T+n|T} \pm \text{c}\hat{\sigma_n}$ where c depends on the probability of Type I error. For a 95% prediction interval, $c = 1.96$.

Forecast evaluation.

The accuracy of a forecast can only be determined by considering how well a model performs on new data that was not used when fitting the model. We separate the data into training and test data. Because the test data is not used in determining the forecasts, it provides a reliable indication of how well the model is likely to forecast in new data. The test data is usually 20% of the total sample and it should usually be as large as the maximum forecast horizon required.

Forecast errors are measured with:

1. Mean Absolute Error: MAE $= \frac{1}{n} \sum |e_t|$
2. Root Mean Squared Error: RMSE $= (\frac{1}{n} \sum |e_t|)^{1/2}$
3. Percentage Error: $p_t = \frac{e_t}{y_t} * 100\%$
4. Mean Absolute Percentage Error: MAPE $= \frac{1}{n} \sum |p_t|$
5. Symmetric Mean Absolute Percentage Error: SMAPE $= \frac{1}{n} \sum \frac{200|y_t - \hat{y_t}|}{y_t + \hat{y_t}}$

Once we have chosen our preferred method of calculating the forecast errors, we can use cross-validation to select the model that minimizes the errors.

Linear Regression.

Besides the benchmark methods, the linear regression $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ is the most commonly used model. In R, we can use the function TLSM() which is very similar to the lm() base R function but with additional time-series functionalities.

To evaluate a regression model, we do the following:

1. ACF plot of residuals: When fitting regression models to time series data, the residuals will likely be auto-correlated. So the assumption of auto-correlation is violated and the forecasts may have large prediction intervals.
2. Histogram of residuals: We must check the distribution of the residuals to check that they are normally distributed.
3. Line graphs: By checking the line graph, we can find out if the variance of the residuals is more or less constant across points in time.
4. Residuals against predictors: We expect the residuals to be randomly scattered without showing any systematic patterns. If we can see any relation between the residuals and the predictor, the model might not be correct.

Some modifications can be done to linear regression to accommodate them to time series data:

1. Trend: We can use time as a the independent variable as in $y_t = \beta_0 + \beta_1 t + \varepsilon_t$
2. Seasonal dummy variables: We need $s - 1$ dummy variables to code s categories. The interpretation of each coefficient is the effect of each season relative to the omitted season.
3. Intervention variables: When some effect lasts one period, we use a dummy variable to indicate that one period (we call this a spike variable). We can also use dummy variables to indicate before and after an event.

Prediction evaluation.

There are four ways of measuring predictive accuracy.

1. R-squared: Measures how well the model will predict the data but does not penalize for the number of predictors.
2. R-squared adjusted: It is calculated in a similar way as R-squared but penalizes for the number of predictors.
3. Cross-validation: To perform cross-validation, remove observation t from the data set and fit the model using the remaining data. Then compute the error $e_t^* = y_t - \hat{y}_t$ for the omitted observation. Repeat this for $t = 1, ..., T$. Then, compute the MSE using $e_1^*, ..., e_T^*$. Follow these steps for every model and choose the best model based on the lower MSE.
4. Akaike's Information Criteria: $AIC = Tlog(\frac{SSE}{T}) + 2(K+2)$ It penalizes the model by the number of predictors.
5. Corrected Akaike's Information Criteria: $AIC_c = AIC + \frac{2(k+2)(k+3)}{T-k-3}$ For small T, the AIC selects too many predictors, so the AIC_c corrects for that.
6. Schwarz' Bayesian Information Criteria (BIC): $BIC = TLog(\frac{SSE}{E}) + (k+2)log(T)$. The BIC is similar to the AIC but the penalty for having more parameters is greater.

Another approach to selecting predictors is step-wise regression. We start with a model that contains all predictors. Then, we remove one predictor at a time and we keep the new model if the measure of accuracy improves. A forward stepwise regression starts with an intercept and adds predictors based on whether it improves accuracy. These procedures are appropriate when the goal is forecasting, but if we wanna study the effect of predictors on the forecast, we must instead look at the p-values.

Two types of forecasting can be done with regressions. Ex-ante forecasts are made using information known in advance. This model requires forecasts of predictors. Ex-post forecasts are made using known values of the predictors. A comparison between these two can help figure out the source of the forecast uncertainty. That is, whether forecast errors come from poor forecasts of predictors or poor forecast models.

We can use lagged values as predictors. For instance, $y_{t+h} = \beta_0 + \beta_1 x_{1t} + ... + \beta_1 x_{kt}$. In this model, the predictors are values that happen n periods before y_{t+n}. The prediction interval for this model is $\hat{y} \pm 1.96\sigma\sqrt{1 + \frac{1}{T} + \frac{(x-\bar{x})^2}{(T-1)s_x^2}}$.

It is also possible to run a non-linear regression. For instance, $ln(y_i) = \beta_0 + \beta_1 log(x_i) + \varepsilon_i$ where the coefficient β can be interpreted as the elasticity. Another option is to use the model $ln(y_i + 1) = \beta_0 + \beta_1 log(x_i + 1) + \varepsilon_i$ if there are negative observations in the sample.

An important rule of thumb to remember is that correlation is not causation. It is possible to have confounder variables that affect both the dependent and independent variables. However, correlations are very useful for forecasting. When two or more predictors are highly correlated, it becomes difficult to separate the individual effect. This is called multi-collinearity.

ARIMA Models.

A time series is stationary it it has constant mean and variance. Meaning that it can have cycles, but not seasonality or trends. It is possible to make a series stationary by taking the first or second difference. A way to detect stationarity is that the autocorrelation function drops to 0 quickly. A differenced series is given by $y'_t = y_t - y_{t-1}$. If the series is made up of white noise, we can write $\epsilon_t = y_t - y_{t-1}$. So the random walk model is given by $y_t = y_{t-1} + \epsilon_t$. In particular, random walk models are widely used for non-stationary data like financial and economic data.

An autoregressive model AR(p) has the form $y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t$ where $-1 < \phi_i < 1$. We can only apply an autoregressive model to stationary data and we can do it by using the fable package. A moving average model MA(q) has the form $y_t = c + \epsilon_t + \epsilon_{t-1} + ... + \epsilon_{t-q}$ where $-1 < \theta_i < 1$. We can combine an autoregressive model and a moving average model into an ARIMA(p, q, d) model where d is the degree of differencing. The model has the form $y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t + \epsilon_{t-1} + ... + \epsilon_{t-q} + \varepsilon_t$.

III. Methodology

Linear Granger Causality Test.

The Granger Causality Test (GCT) considers two Vector Auto-Regressive (VAR) models

1. $Y_t = \beta_0 + \sum \beta_i Y_{t-i} + \varepsilon$
2. $Y_t = \beta_0 + \sum \beta_i Y_{t-i} + \sum \alpha_i X_{t-i} + \varepsilon$

where p is the lag.

Given these two models, we conduct a hypothesis test where H_0 is equivalent to X does not cause Y and H_a is equivalent to X causes Y.

H_0: $\forall i \in 1, ..., p, \alpha_i = 0$
H_a: $\exists i \in 1, ..., p, \alpha_i \neq 0$

Non-Linear Granger Causality Test.

Consider a VARNN(p) model; that is, a multi-layered neural network that takes into account the p previous values of the predictor variables and the target variable (Y) to predict future values of Y.
Consider the two VARNN(p) models:

1. $Y_t = \phi(Y_{t-1}, ..., Y_{t-p}) + \varepsilon$
2. $Y_t = \phi(Y_{t-1}, ..., Y_{t-p}, X_{t-1}, ..., X_{t-p}) + \varepsilon$

Then, we compute the Fisher (F) statistic to test the null hypothesis that X does not cause Y.

Transfer Entropy Test.

The concept of Transfer Entropy quantifies the information flow from one time series X to another Y. This measure was created to address a significant limitation of mutual information, which captures shared information between variables. Unlike mutual information, Transfer Entropy focuses on capturing the directional transfer of information from one variable to another. We can express the transfer entropy as follows

$$TE_{(X \to Y)} = H(Y_t | Y_{t-1}, \ldots, Y_{t-q}) - H(Y_t | Y_{t-1}, \ldots, Y_{t-q}, X_{t-1}, \ldots, X_{t-p})$$

This expression represents the difference in entropies between a model that doesn't include X as a predictor and a model that does. A higher transfer entropy indicates a stronger flow of information from X to Y.

IV. Simulations

The goal of this project was to compare the results of the Granger Causality Test (GCT) and the Transfer Entropy Test (TET). In order to do so, I proceeded with the following steps:

First, I coded the following data-generating processes. Note that $\epsilon \sim \mathcal{N}(0,1)$.

1. Linear data with a strong linear relationship between X and Y: $y_t = y_{t-1} + x_{t-1} + \epsilon$
2. Linear data with a weak linear relationship between X and Y: $y_t = y_{t-1} + 0.2x_{t-1} + \epsilon$
3. Linear Y: $y_t = y_{t-1} + \epsilon$
4. Non-linear data with a strong relationship between X and Y: $y_t = \frac{1}{\exp(y_{t-1})} \times \frac{1}{\exp(x_{t-1})} + \epsilon$
5. Non-linear data with a weak relationship between X and Y: $y_t = \frac{1}{\exp(y_{t-1})} \times \frac{1}{\exp(0.2x_{t-1})} + \epsilon$
6. Non-linear Y: $y_t = \frac{1}{\exp(y_{t-1})} + \epsilon$

Then, I generated 100 vectors of 100 entries each using each formula and performed the GCT and TET on each vector. The graphs and tables in the appendix characterize the distribution of p-values after carrying out each of the 100 tests for each data-generating process. In particular, you will find the following in the appendix:

1. Tables 1 through 6 show the percentage of simulations where both tests reject the null, each test rejects the null, and none of the tests reject the null at the 5% level.
2. Tables 7 through 12 show similar tables but at the 1% level.
3. For each data-generating process, a graph of the distribution of p-values.
4. For each data-generating process, a graph showing the empirical cumulative distribution of p-values.
5. For each data-generating process, a graph showing the correlation of p-values between each test.

V. Discussion

In this section, I will comment on the plots and tables that show the results of the simulation and discuss the performance of each test. The plots can be found in the Appendix.

Strong linear relationship: In the plots and tables, we can observe that the GCT rejects the null for all the simulated data at the 5% level. However, according to Tables 1 and 7, the TET fails to reject 23% of the time at the 5% level and 37% at the 1% level. In this case, the GCT shows better results than the TET. However, it must be noted that both tests coincide 76% of the time.

Weak linear relationship: The GCT rejects the null more often than the TET. In fact, the GCT fails to reject 68% of the time and the TET 83% of the time. Once again, the GCT shows better performance than the TET.

Absent linear relationship: The GCT incorrectly rejects the null more often than the TET, but the difference is equivalent to 5 percentage points (at the 5% level). This shows that, in general, the GCT rejects the null more often but the difference is not very substantial.

Strong non-linear relationship: The GCT rejects the null 100% of the time. The TET rejects the null 95% of the time (at the 5% level). At the 1% level, the GCT still rejects the null 100% of the time, but the TET does it 79% of the time.

Weak non-linear relationship: The GCT rejects the null substantially more often than the TET. At the 5% level, the GCT fails to reject 31% of the time and the TET fails to reject 87% of the time.

Absent non-linear relationship: At the 5% and 1% level, the TET and GCT incorrectly reject the null equally as often. This can be seen in Tables 6 and 12.

In general, the GCT is more likely than the TET to detect causality even when there is none. We can see this in the graph of the distribution of p-values and the empirical cumulative distribution graph. Interestingly, the scatter plot of p-values does not show a strong correlation between the p-values of the GCT and the TET in any case. In the tables, we can see that the TET is able to detect causality at a higher rate in the case of the non-linear relationship than the linear relationship.

VI. Appendix

General Title: $\alpha = 0.05$

Table 1: Strong Linear Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.76	0.23
GC $> \alpha$	0.00	0.00

Table 2: Weak Linear Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.05	0.26
GC $> \alpha$	0.11	0.57

Table 3: Absent Linear Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.01	0.06
GC $> \alpha$	0.04	0.88

Table 4: Strong Non-Linear Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.95	0.05
GC $> \alpha$	0.00	0.00

Table 5: Weak Non-Linear Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.07	0.62
GC $> \alpha$	0.06	0.25

Table 6: Absent Non-Linear Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.00	0.06
GC $> \alpha$	0.06	0.88

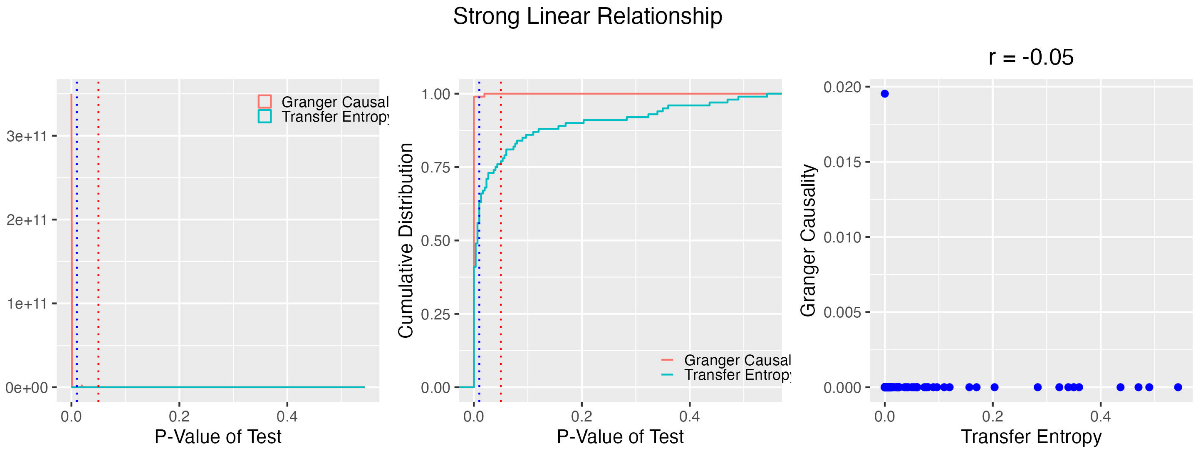

Figure 1: Strong Linear Relationship

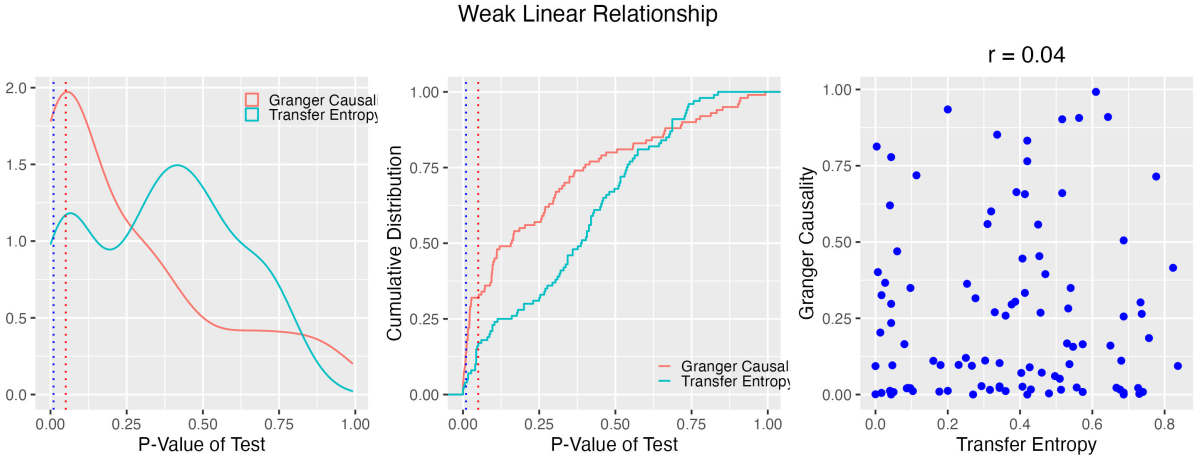

Figure 2: Weak Linear Relationship

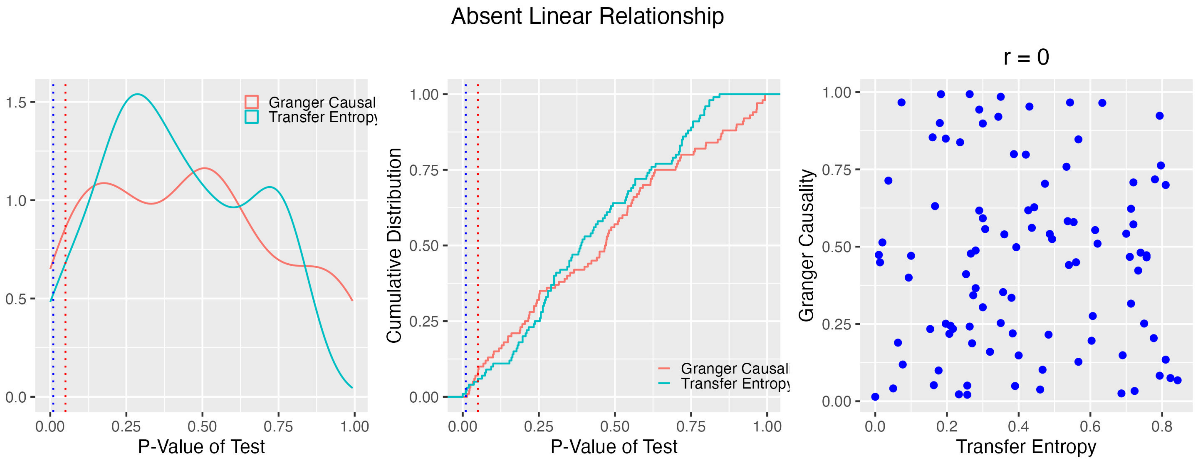

Figure 3: Absent Linear Relationship

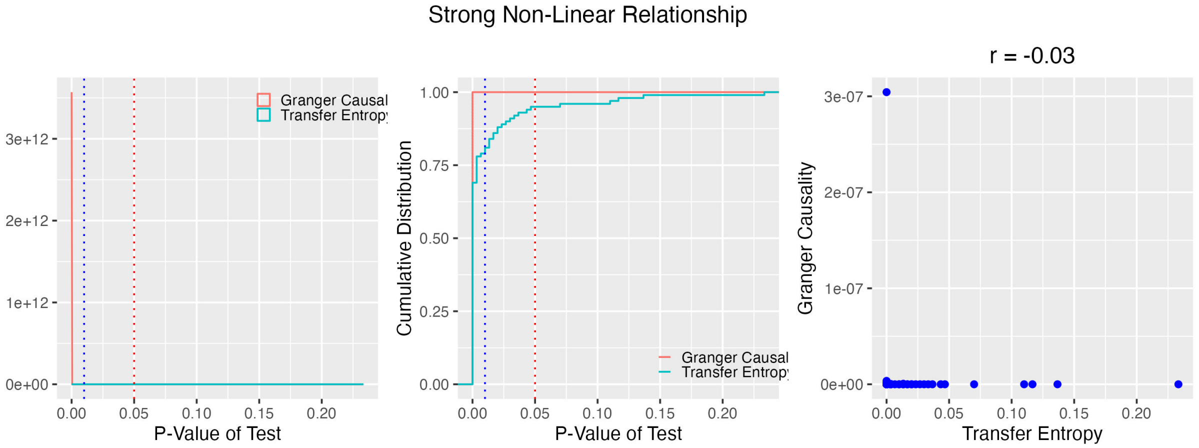

Figure 4: Strong Non-Linear Relationship

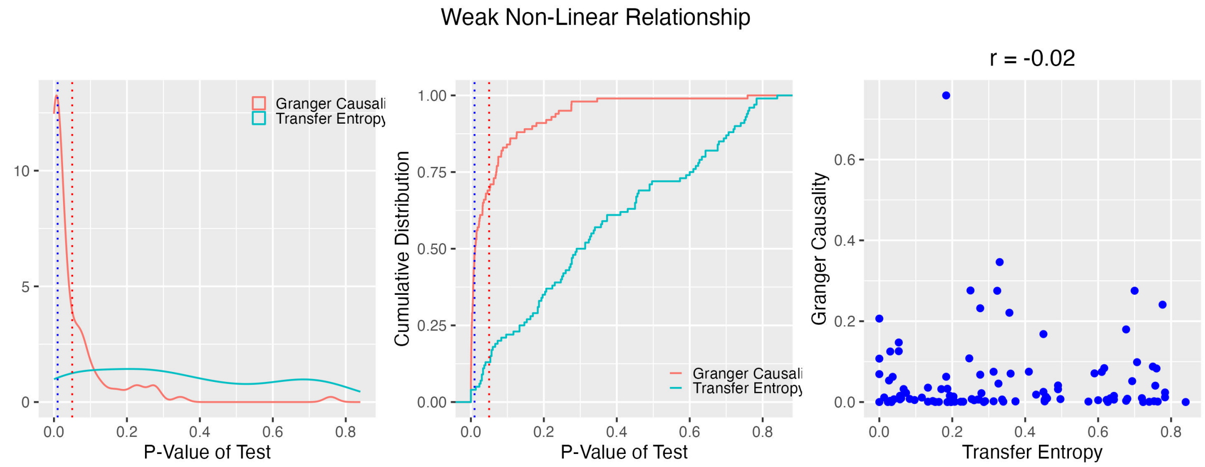

Figure 5: Weak Non-Linear Relationship

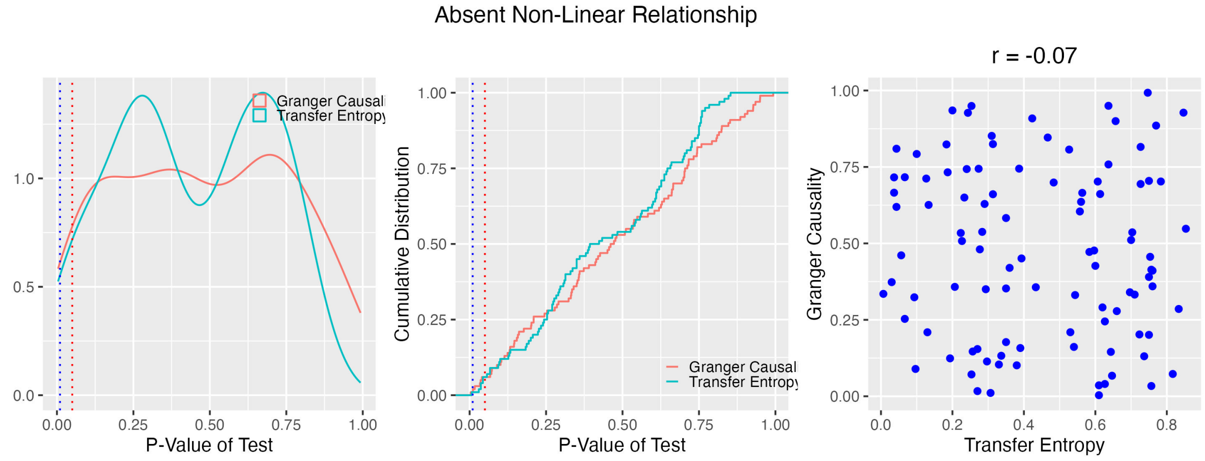

Figure 6: Absent Non-Linear Relationship

General Title: $\alpha = 0.01$

Table 7: Linear Data with Strong Linearity

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.55	0.37
GC $> \alpha$	0.01	0.00

Table 8: Linear Data with Weak Linearity

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.01	0.13
GC $> \alpha$	0.03	0.83

Table 9: Linear Data with No Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.00	0.00
GC $> \alpha$	0.01	0.98

Table 10: Non-Linear Data with Strong Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.79	0.19
GC $> \alpha$	0.00	0.00

Table 11: Non-Linear Data with Weak Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.01	0.47
GC $> \alpha$	0.03	0.49

Table 12: Non-Linear Data with No Relationship

	TE $< \alpha$	TE $> \alpha$
GC $< \alpha$	0.00	0.01
GC $> \alpha$	0.01	0.98