Compression and Causal Relations Around Macroeconomics: an Improved CUTE Model

Ting Hu[a], Zixuan Cao[b], Luofeng Zhou[a], Bo Han[b,*]

[a]*Economics and Management School, Wuhan University, China*
[b]*School of Computer Science, Wuhan University, China*

Abstract

In big data context, reseachers have easy access to economic growth data. Discovering causal relations between macroeconomic indicators is of great significance in economic trends prediction, trading strategies formulation, and improving economic structure. Ttraditional causal model for time series data usually adopts Granger causality test, which relies on the preset lag order and has the demerits of noise sensitivity and inaccuracy. In this paper, based on the latest development of the causality test model - CUTE, we applied multi-value expansion and proposed the improved CUTE model. We analyzed the causal relations between Canadian GDP and other economic indicators from January, 2009 to May, 2018. Improved CUTE model is irrelevant to preset lag orders and experimental results are proved more accurate with long sequence and high noise. Moreover, the improved algorithm has lower time complexity. Empirical analysis shows that West Texas Oil, Brent oil, new housing starts, unemployment rate, and core consumer price index are significant causes for Canadian GDP changes at level $\alpha = 0.05$.

Keywords: GDP, Granger Causality Test, CUTE Model

1. Introduction

Macroeconomic variables are always considered powerful measures of economic dynamics in national or regional scale. Precise forecast of their variations perform significance in research on development of economic trends, related policy making and financial practice such as foreign exchange trading. Effective methods in identifying causal relations between macroeconomic variables, however, are prerequisite and basic. In this paper, variables intuitively and theoretically correlated with GDP are considered and a model based on compression is applied to identify macroeconomic causal structure.

Prior empirical researchers tried their best to find out various economic indicators to forecast GDP changes effectively. With the rapid development of Internet, electronic sensors, cloud computing and big data, information technology provides easy access to massive economic growth data. Identifying relations between GDP and other economic indicators has always been a meaningful but chanllenging topic in economic research. Previously studied variables involve satellite observed visible-near infrared emissions (Elvidge et al., 1997), dynamic distribution of national population (Lozano and Gutierrez, 2008), sales and price variations in e-commercial platform (Lendle et al., 2013), energy data Pao and Tsai (2011), water consumption (Oki and Kanae, 2006), data from tourism (Gunduz* and Hatemi-J, 2005) and so on. Each of them reflects one or more factors concerned in different economic growth models in time. Under the big data environment, empirical economic research should identify relations between economic growth and observable variables from various sources, such as national electricity consumption and GDP (Jumbe, 2004; Oh and Lee, 2004).

When evaluating causal relations in economic growth, previous works suffer from some technical constraints. First, the macroeconomic variables tends to be rather *complex*. The so-called *complex* variables refers to those facing severe endogeneity, having countless influencials and noisy. It is rather hard to draw conclusions close to reality through ordinary regression analysis. Second, since the vector autoregression (VAR) specification indicates linear assumption, nonlinear structure is hard to handle although this could be fixed when VAR is built by formulating the right for-

*Corresponding Author: Prof. Bo Han, Professor, School of Computer Science, Wuhan University, China. Email: bhan@whu.edu.cn.

m of variables (i.e. logarithms, exponentials or polynomials). Third, additional restrictions are brought by the main traditional means of econometric causal model, Granger causality test. Granger causality test is an easy classic statistical test built on vector autoregression model. However, the model needs to meet one of the following two conditions: both variables are stable, or there is a cointegration relationship between the two variables. For multivariate time series that do not meet both of them, measures should be taken to detrend and demean. Also, the test depends on the preset orders of lag operators, and there is no uniform method to determine them. AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are widely used and empirical intuition is also an alternative. Unfortunately, sometimes different criterion gives different *effective* lag orders, which makes the accuracy of the results obtained by Granger causality test controversial.

Time series-based causality analysis draws much attention in the context of where data are abundant (Hyttinen et al., 2017; Gong et al., 2017; Zhang et al., 2017). The CUTE (**CaU**sal inference on even**T** sequ**E**nces) model is one of the latest theoretical breakthrough proposed by German scholar Budhathoki and Vreeken (2018a,b) at the 2018 International Big Data Analysis Summit SIAM International Conference on Data Mining. Different from regression analysis in Granger causality test, the CUTE model conbines the idea of Granger causality and compression theory. Under the assumption that if two time series preserve causal relation, the *cause* series could help the *consequence* series to reach a higher compression ratio, CUTE model can identify statistically significant linear and nonlinear causal relations from time series without any preset regression specification. The original CUTE model only consider causality for binary time series, which differs a lot from economic data.

Our contribution is twofold. First, we form an improved CUTE model which does better than Granger causality test in causal identification around GDP and other concerned macroeconomic factors. The original CUTE model can only perform causality analysis for binary time series, which differs a lot from economic data. In this paper, multi-value expansion on CUTE model is performed and then it can be applied to economic world. Second, the empirical results give a novel insight for policy making and position forming. Some causal relations between Canadian GDP and other macroeconomic indicators may be hard to identify by traditional Granger causality test while our improved CUTE model gives a definate answer.

The remainder of the paper proceeds as follows. Section 2 lists previous empirical works about causality around GDP. Section 3 describes the improved CUTE model. Section 4 tests robustness of the model under simulated data. Section 5 shows the empirical test for Canadian GDP and other related macroeconomic indicators and its results. Section 6 draws conclusions.

2. Literature Review

2.1. multivariate specification

In 20th century, modelling economic growth is the most concerned part of macroeconomics. Since Solow put forward his model, technical change was always considered a pivot factors in economic growth. Then learning-by-doing model and Romer's endogenous growth model were widely accepted. These specifications are theoretical and go through the path from exogenous machanisms to endogenous ones. Recent researchers, on the contrary, focus more on exogenous impacts. Their belief is more realistic. Although there may be a steady state in theory, underlyings and backgroundings change rapidly. To capture the in-time effect from a seemingly unexpected event on macroeconomics may tell more about the future policy for government and possible trading strategies for investers. Bridge model is an early specification to research on monthly macroeconomic data, but it demonstrates only the statistical coexisting relations rather than causality (Baffigi et al., 2004). Then VAR specifications take the dominant place in empirical macroeconomics. After that, there exist some improved ones such as Bayesian VAR specification (Bańbura et al., 2010) .

In fact, the essence of VAR specifications is a linear regression and the Granger causality test is a statistical F-test based on the linear system. One main probelm is the linear assumption. This could be partly solved by predetermine the form of regression varibales (i.e. logarithms, exponentials or polynomials). Another problem is brought by the report interval of GDP. GDP is usually reported monthly or quarterly while the causal might occurs within a week (Ghysels et al., 2016; Gong et al., 2017). The estimated data generating process (DGP) may tell a different causal structure skewed by aggregation.

2.2. influencial indicators

A wide range of variables from both economic sector or else have been taken into consideration. Soytas and Sari (2003) used the Granger causality test to verify the causal relation between energy consumption and GDP.

They found a bidirectional causal relation in Argentina. In Italy and South Korea, energy consumption is the cause of changes in GDP, and vice versa in Turkey. Lee (2005) studied the same topic in 18 developing countries from 1975 to 2001 and found that both long-term and short-term Granger test demonstrate that change in energy consumption is the cause of change in GDP, suggesting that the conservative may be detrimental to the economic development of developing countries. Ghosh (2009) found long-term and short-term causal relations from India's real GDP changes and electricity supply to employment while there is no causal relationship between electricity supply and real GDP. Based on their empirical results, they urged that the Indian government could reduce electricity supply to reduce the waste of electricity and not affect the growth of real GDP. Li and Li (2011) analyzed the impact of coal consumption on China and India's GDP, and proposed to reduce carbon emissions and develop cleaner and more efficient energy source to achieve sustainable development. Pao and Tsai (2011) studied the causal relationship between the BRICS countries' CO_2 emissions, energy consumption, FDI (foreign direct investment) and GDP in 1992-2007. They concluded that there was a two-way causality between FDI and GDP. Also, there were also significant two-way causal relations between the GDP-energy consumption ratio and the GDP-pollutant emission ratio. In addition, they proposed that the pollutant emissions have a scale effect and a halo effect. Omri et al. (2014) conducted a similar experiment with the world divided by region and found that there is a two-way causal relation between FDI and carbon dioxide emissions except Europe and northern Asia. Changes in GDP were always caused by carbon dioxide emissions apart from the Middle East and Northern Asia. Under Toda-Yamamoto non-causality test (a method based on the Granger causality test), Amiri and Ventelou (2012) found a two-way causal relation between health care expenditures and GDP in OECD countries. Khan et al. (2017) used the same method to prove that in Malaysia, there was a one-way causal relation from household loan to GDP. The author believed that this study could provide a reference for the Malaysian government's policy of entering high-income countries in 2020. Zhang et al. (2014) used traditional Granger causality test to study the internal mechanism between China's economic growth and energy consumption. Based on the research results, they put forward suggestions for optimizing the industrial structure and vigorously developing the tertiary industry.

Studying the causal relations between GDP and other economic indicators has important reference value for economic policy. The main method adopted in the above research is Granger causality test, which is somewhat cumbersome to deal with hetergeneous data. At the same time, it is difficult to select orders of lag operators. There are different methods and each method has relevant theoretical and intuitive verification. However, the lag order of different methods may vary greatly, leading to different causal relations. Thus, the accuracy of the results is somewhat controversial.

3. The model

The CUTE model provides an alternative to identify causal structure in systems which may not suitable for VAR specification and Granger Causality test (Budhathoki and Vreeken, 2018a,b). Referring to the theory of information compression, the SNML (sequential normalized maximal) method is used to estimate parameters so that it adapts to different joint probability distributions. To test the robustness, the author of CUTE applied the model to the causality analysis of the generated sequence and the river hydrological sequence.

In VAR specification, predetermined form of variables may lead to bias. For example, if x is exponential on y and distributed around zero, we may say that x seems linear on y. When an extreme x comes into being, the model losses explanatory ability. The CUTE model, however, is irrelevant to a preset specification, which shows more compatiblity for nonlinear and opaque systems. When compared to Granger causality test, CUTE model need not to preset orders of lag operators. In addition, CUTE does not require Augmented Dickey-Fuller tests and differences of variables.However, the original CUTE model deals only with binary time series while variables are always continuous in macroeconomics. Thus we provide a multi-value expansion to let the improved model fit well.

3.1. Theoretical Framework

Just like Granger causality test, below are some ordinary but necessary assumptions and the definition of Granger causality.

Assumption 1. *Cause precedes the effect in time.*

Assumption 2. *Cause has unique information about the future values of effect.*

Definition 1. *Let $\mathcal{F}_t$ be filtrations of all information available at time t. We say a time series x_t **Granger-causes** another series y_t if conditional likelihood exhibits $\mathcal{L}(y_{t+1}; \theta \,|\, \mathcal{F}_t) > \mathcal{L}(y_{t+1}; \theta \,|\, \mathcal{F}_t \setminus \{x_t\})$.*

When we associate the likelihood to compression, a sequencial encoded length is considered. We first define ideal encoded length.

Definition 2. *The ideal encoded length of a time series x_t is given by: $len(x_t) = -log\mathcal{L}(x_t \big| \{x_{t-1}\})$*[1]*. This is commonly known as **log loss** in learning theory. Thus the n-period total sequencial encoded length of x_t is given by: $\sum len(x_t) = \sum_{t=1}^{T} -log\mathcal{L}(x_t \big| \{x_{t-1}\})$.*

Nevertheless, another time series y_t may contain some information about x_t. By taking y_t into account when calculating encoded length, we have conditional encoded length.

Definition 3. *The ideal encoded length of a time series x_t conditional on y_t is given by: $len(x_t \big| y_t) = -log\mathcal{L}(x_t \big| \{x_{t-1}\}, \{y_{t-1}\})$. The n-period total sequencial encoded length of x_t conditional on y_t is given by: $\sum len(x_t \big| y_t) = \sum_{t=1}^{T} -log\mathcal{L}(x_t \big| \{x_{t-1}\}, \{y_{t-1}\})$.*

The difference between $len(x_t)$ and $len(x_t \big| y_t)$ is the former shows the predictability by using the past realization of x_t itself and the latter involes the past realization of y_t. Hence, their difference measures the extra predictability of x_t contributed by the past realization of y_t which is not available otherwise.

Definition 4. *The **causal dependence** from y_t to x_t and vice versa are given by*

$$\Delta_{\{y_t\}\to\{x_t\}} = \sum_{t=1}^{T} len(x_t) - \sum_{t=1}^{T} len(x_t \mid y_t),$$

$$\Delta_{\{x_t\}\to\{y_t\}} = \sum_{t=1}^{T} len(y_t) - \sum_{t=1}^{T} len(y_t \mid x_t),$$

Under the two assumptions, the direction with largly dependency is consistent to the causal direction. We form our belif as follow:

- If $\Delta_{\{y_t\}\to\{x_t\}} > \Delta_{\{x_t\}\to\{y_t\}}$, we infer $\{x_t\} \to \{y_t\}$.
- If $\Delta_{\{y_t\}\to\{x_t\}} < \Delta_{\{x_t\}\to\{y_t\}}$, we infer $\{y_t\} \to \{x_t\}$.
- If $\Delta_{\{y_t\}\to\{x_t\}} = \Delta_{\{x_t\}\to\{y_t\}}$, the causal relation is hard to determine.

The problem is how to determine the statistic threshold of the test criteria. The essence is to determine the probability of Type I error. Ryabko and Astola gives the answer for compression. The null hyposis is the unconditional and conditional distributions are the same. Thier framework is described as follows:

[1]All *logs* in this paper refer to log_2.

Theorem 1. *Let α be a given level of significance and $\{x_n\}$ be a sequence over a certain alphabet. Null hypothesis H_0 is defined as the source of x_n has a distribution P, and alternative hypothesis H_1 is that the source distribution of x_n is Q. The probability of Type I error is no larger than α if*

$$logQ(x_n) - logP(x_n) > -log\alpha.$$

In the context of our research, a proposition drawn by this therom is useful.

Proposition 1. *Let α be a given level of significance and $\{x_t\}$, $\{y_t\}$ be two binary time series. Null hypothesis H_0 is defined as the causal direction is vague. Two alternative hypothesis H_1 and H_2 are $\{x_t\} \to \{y_t\}$ and $\{y_t\} \to \{x_t\}$. The probability of Type I error is no larger than α if*

$$\left|\Delta_{\{y_t\}\to\{x_t\}} > \Delta_{\{x_t\}\to\{y_t\}}\right| > -log\alpha.$$

The decision rule is dipicted in Figure 1.

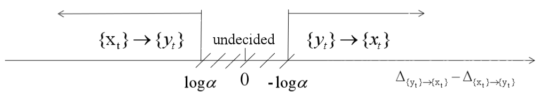


Figure 1: The decision rule

However, the causal identification lies under the assumption that the likelihood functions are already known. In reality, it is definately not the case. So we should characterize distributions of both time series.

3.2. Sequential Normalized Maximum Likelihood

When it comes to specify the likelihood function, parameterized families of distributions are considered. Then the MLE for parameters is done by solving a normal minimax problem. If the true data generating distribution is in the assumed model class, the maximum likelihood strategy which forcasts x_{t+1} by using $P_{\hat{\theta}(x_t)}$ where $\hat{\theta}(x_t)$ denotes the MLE based on x_t. However, the ML strategy is not robust i.e. if the assumption is not the case, the result can be arbitrarily bad (Kotlowski and Grunwald, 2012).

Then sequential normalized maximum likelihood model is proposed as follow in breif.

Theorem 2. *A sequential normalized modification of ML strategy for the minimax problem in parameter estimation can achieve optimality even if the true distribution lies out of the assumed ones. The prediction function of x_t under $\{x_{t-1}\}$ is:*

$$\mathcal{L}^*(x_t; \hat{\theta} \left| \{x_{t-1}\}\right) = \frac{P_{\hat{\theta}(\{x_{t-1}\}, x_t)}}{\sum_x P_{\hat{\theta}(\{x_{t-1}\}, x)}},$$

where $\hat{\theta}(\{x_{t-1}\}, x_t)$ *denotes the MLE based on* $x_1, x_2, ..., x_{t-1}, x_t$ *and* $\hat{\theta}(\{x_{t-1}\}, x)$ *denotes the MLE based on* $x_1, x_2, ..., x_{t-1}, x$.

For binary data structure, a parameterised family of Bernoulli distributions is concerned. Suppose the probability mass function for Bernoulli distribution is given by $P_\theta(X = k) = \theta^k(1-\theta)^{1-k}$, where $\theta \in [0, 1]$ and $k \in \{0, 1\}$.

Proposition 2. *The SNML strategy (likelihood function) for Bernoulli distribution* $P_\theta(X = k) = \theta^k(1-\theta)^{1-k}$ *is specified as:*

$$\mathcal{L}^*(x_t = 1; \hat{\theta} \left| \{x_{t-1}\}\right.) = \frac{(t_1+1)^{t_1+1} t_0^{t_0}}{t_1^{t_1}(t_0+1)^{t_0+1} + (t_1+1)^{t_1+1} t_0^{t_0}},$$

$$\mathcal{L}^*(x_t = 0; \hat{\theta} \left| \{x_{t-1}\}\right.) = 1 - \mathcal{L}^*(x_t = 1 \left| \{x_{t-1}\}\right.),$$

where $t_1 = \sum_{i=1}^{t-1} x_i$ *and* $t_0 = t - 1 - t_1$.

3.3. the improved CUTE model

Original CUTE model is only suitable for binary time series to conduct causality test. We proposed a multi-value expansion on it. For macroeconomic time series $\{x_t\}$ such as monthly GDP, the realization of each period is encoded in two-digit binaries. The encoding specification is described as follows:

- If x_t is greater than x_{t-1}, the output is 10,
- If x_t is smaller than x_{t-1}, the output is 01,
- If x_t equals to x_{t-1}, the output is 00.

It is easy to see that our encoding series contain three different states for each time series while the original one could explain only two.

To forecast x_t, for each step of time series forecasting, a thourough filtration of information $\mathcal{F}_{t-1}$ could be used. Let $u = \sum_{i=1}^{t-1} x_i \oplus y_i$ denote the number that x_i and y_i are different. Then 2^u binary time series are constructed.

Theorem 3. *The minimum of the likelihood function* $\mathcal{L}^*(x_t = 0; \hat{\theta} \left|\mathcal{F}_{t-1}\right.)$ *is attained by the series which contains the least 1 drawn by u; The minimum of the likelihood function* $\mathcal{L}^*(x_t = 1; \hat{\theta} \left|\mathcal{F}_{t-1}\right.)$ *is attained by the series which contains the least 0 drawn by u.*

Proposition 3. *The conditional SNML strategy (the lower bond of likelihood function) for Bernoulli distribution is specified as:*

$$\mathcal{L}^*(x_t = 1; \hat{\theta} \left|\mathcal{F}_{t-1}\right.) = \frac{(t_1+1)^{t_1+1} t_0^{t_0}}{t_1^{t_1}(t_0+1)^{t_0+1} + (t_1+1)^{t_1+1} t_0^{t_0}},$$

$$\mathcal{L}^*(x_t = 0; \hat{\theta} \left|\mathcal{F}_{t-1}\right.) = \frac{(t_3+1)^{t_3+1} t_0^{t_2}}{t_3^{t_3}(t_0+1)^{t_2+1} + (t_3+1)^{t_3+1} t_0^{t_2}},$$

where $t_1 = max(\sum t_{x_1}, \sum t_{y_1})$, $t_3 = min(\sum t_{x_1}, \sum t_{y_1})$ *and* $t_0 = t - 1 - t_1$, $t_2 = t - 1 - t_3$.

Till now, the causal dependence can be calculated through SNML strategy among macroeconomic indicators.

An interesting thing worth noting is the sum of two conditional SNML strategy is smaller than 1. Another series performs as a chance to squeeze the original series more.

3.4. theoretical evaluation

The improved CUTE model calculates fast. More precisely, $len(x_t)$ is the sum of negative log-likelihood in unconditional SNML specification. For each iteration, a single value t_1 is recorded so the complexity is $O(n)$. $len(x_t \mid y_t)$ is the sum of negative log-likelihood in conditional SNML specification. For each iteration, two t_1s for both series are recorded and their minimum is used to calculate the predicting strategy, whose complexity is also $O(n)$. To conclude, the complexity of improved CUTE model for two-digit binary series is linear to the length of time, leading to a fast speed.

3.5. robustness test through simulations

To test the robustness of the improved CUTE model, we generate time series data of different lengths by simulation and add different proportional noise to them. Traditional Granger causality model and improved CUTE model are used to determine the causality and their accuracies are compared.

Specifically, we first generate time series data of different lengths according to the Bernoulli distribution, in which the success probability θ of the Bernoulli distribution is randomly selected from the interval [0.1, 0.9]. The *cause* time series is randomly advanced by n digits, where n is an integer within the interval [0, 5]. The first n digits of the *consequence* series are filled by 0 or 1 randomly and the following digits are the same as the *cause* series from the begining. Then, according to the preset noise ratio (0%, 10%, 20%, 30%, 40%), some digits from original series are inversed. Sequence lengths are set to be 50, 100, 200, 400 respectively for each experiments.

In traditional Granger causality test, the lag order 1, 2, and 3 are usually chosen. In our experiment, orders of lag are chosen from [1, 5] since the longest time series length in the simulation is 400 and the maximum number of advance in *cause* time series is 5. Significant level α is set 0.05.

Figure 2 gives a visualized example. The *consequence* series is given by the left-shifted *cause* series plus 40% noise.

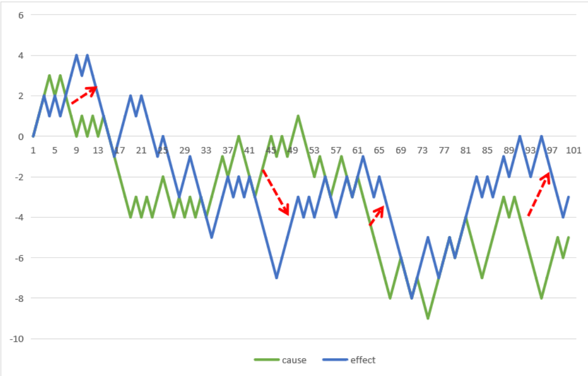


Figure 2: An example with 40% noise

Figure 3 concludes the accuracy of both method and each length of series. Accuracy of Granger causality test decreases sharply when the proportion of noise increases. The improved CUTE model does not exhibit such characteristic. As the sequence grows longer, this merit in robustness of improved CUTE model is more significant.

Next we check the robustness of our encoding method. If the causality drawn by traditional Granger causality before and after encoded remains the same, our improved model could be possibly accurate. The *cause* series is produced by random walk process. The first n numbers of the *consequence* series are filled by 0 or 1 randomly and the following numbers are the same as the *cause* series from the begining. Noise in this test is different from the previous one because of the different essenses. Noise terms are time-independent and i.i.d. white noise with a smaller variance than the random walk. Then, we encode the two series through the encoding process given by the improved CUTE model. Traditional causality test is implemented both before and after the encoding process. The significant level is set $\alpha = 0.01$.

4. Empirical evidence

4.1. Variables

National macroeconomic indicators always explain the change of GDP. In national income accounts identity, GDP euqals to consumption, investment, government purchase and net export. In western countries,

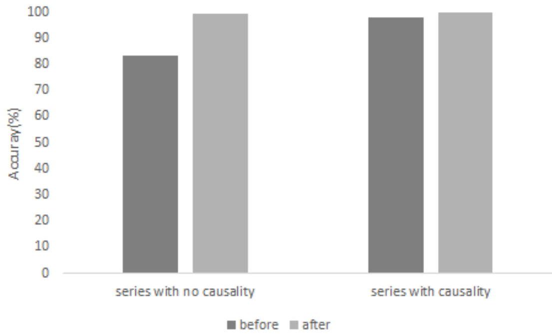
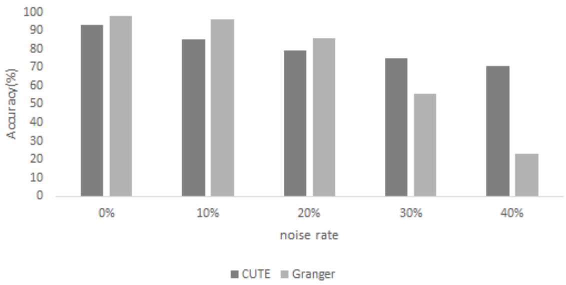


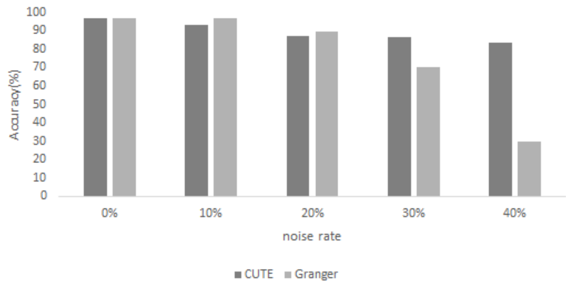
Figure 4: Robustness of encoding process

consumption is considered the leading mechanics of macroeconomics. With the economic growth, increase in wealth per capita directs more consumption. CPI and core CPI reflect the average price level, which directly influence the fisical and monetary policies. Indirect effects are also prominent (i.e. change of marginal propensity to consume). Retail sector explains much of the consumption especially in European countries whose economy is driven by retail sector. Housings preserve duality. They are necessities for citizens while they also act as financial investment goods. To differentiate these two characteristic, the purchase of new house is considered since abundant housing is no longer a necessity. Housing market especially new housing market is sensitive to monetary polices, which bears a powerful spillover effect (Iacoviello and Neri, 2010). The net export reflects the cost and quality of national productivity. Many developing countries take advantage of thier cheap human capital to gain a rapid increase of GDP. Also, the increase of GDP, which indicates the productivity is relatively high, may lead to the increase of current account.

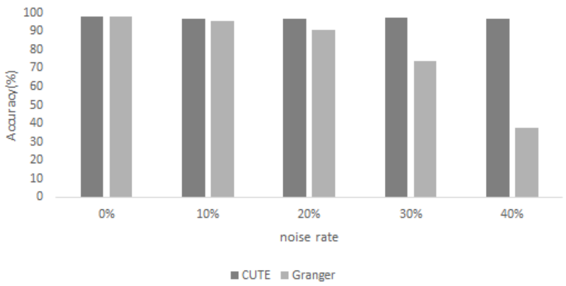
Appart from national indicators, international ones also draw much attention. Carruth et al. (1998) believed past real interest rate, unemployment and crude oil price were predictive for the future unemployment. Hamilton (1996) pointed out crude oil price counld be a good instrument variable for GDP to some degree. Jiménez-Rodríguez* and Sánchez (2005) found the decrease of crude oil price is a bad news to Canadian economy while Korhonen and Ledyaeva (2010) got an opposite conclusion. This might be caused by the changing role Canada plays from an exporter to an importer in crude oil market. Elder and Serletis (2009) researched on this topic from the view of uncertainty. They showed the increase in uncertainty decreased output from manufacture and mining industry and then the real GDP.



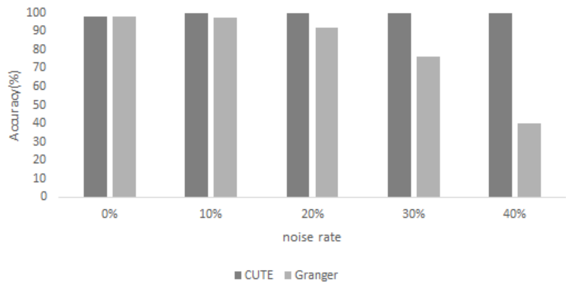
(a) Lenth=50



(b) Lenth=100



(c) Lenth=200



(d) Lenth=400

Figure 3: Accuracy in causal binary series for each model

Taking what have been mentioned above, in this paper, we first calculate the montly change in GDP, CPI, core CPI, new housing price index, sales of new houses, retail sales. These variables are in *difference* form, which shows that we are more interested in their change rather than the base value. For other varibles, the original form is considered. These varibales are: current account, unemployment, average Brent crude oil price and average West Texas crude oil price. Note that the two crude oil price are built since they both act as benchmark price of the world crude oil market. National economic data are collected from Eastmoney and international curde oil price from FRED Economic Data.

Table 1 is the sample description and Figure 5 shows the relations between GDP and other macroeconomic indicators.

Next we encode these time series into two-digit binaries. For variables in *difference* form, we concern the sign of it. For variables in original form, we concern sequencial difference.

4.2. Causality Analysis

Before the Granger causality test, an augmented Dickey-Fuller test is done. Table 2 shows the result. None of the indicators present unit root.

Through method mentioned earlier to determine the order of lag operator, in this test order of 1, 2 and 3 are selected. Table 3 and 4 exhibits the results of traditional Granger causality test from GDP and to GDP respectively. Then the improved CUTE model is implemented on the encoding series. Results are recorded in Table 5. In the test, $\{y_n\}$ is always set to be monthly GDP.

Table 2: the Augmented Dickey-Fuller Test

Indicators	AIC	t-value	p-value
Monthly GDP	0	-9.808	0.0000^{***}
Monthly CPI	0	-9.680	0.0000^{***}
Monthly Core CPI	2	-7.492	0.0000^{***}
Retail sales	2	-8.073	0.0000^{***}
Housing Starts	3	-3.245	0.0175^{**}
NHPI	0	-10.660	0.0000^{***}
Current Account	1	-3.800	0.0029^{***}
Unemployment	0	-11.588	0.0000^{***}
Brent oil price	1	-6.228	0.0000^{***}
West Texas oil price	1	-6.222	0.0000^{***}

Traditional Granger causality test shows that retail sales, current account, unemployment, brent oil price and west texas oil price are the causes of GDP and GDP is the cause of housing starts, new house price index and current account. Only the relationship between current account and GDP shows bidirectional causality. Intuitively, the bidirectional causality is due to the existence

Table 1: Sample Description

Indicators	Mean	Std.	Max.	Min.	25^{th} Per.	Median	75^{th} Per.
Monthly GDP(change in ratio)	1.34e-3	2.82e-3	6.00e-3	-7.00e-3	-1.00e-3	2.00e-3	3.00e-3
Monthly CPI(change in ratio)	1.48e-3	3.56e-3	1.15e-2	-7.20e-3	-8.00e-4	1.55e-3	3.48e-3
Monthly Core CPI(change in ratio)	1.64e-3	3.38e-3	1.70e-2	-5.90e-3	0.00e-3	1.70e-3	3.00e-3
Retail sales(change in ratio)	2.10e-3	7.90e-3	2.10e-2	-2.30e-2	-2.00e-3	2.00e-3	7.00e-3
Housing Starts(thousand suites)	1.93e+2	2.40e+1	2.54e+2	1.17e+2	1.82e+2	1.95e+2	2.08e+2
NHPI(change in ratio)	2.53e-3	9.47e-3	1.00e-1	-7.00e-3	1.00e+3	2.00e-3	3.00e-3
Current Account(billion CAD)	-1.09e+0	1.47e+0	2.69e+0	-4.38e+0	-2.13e+0	-8.36e-1	-2.02e-1
Unemployment(change in value)	1.34e-4	1.49e-3	7.00e-3	-4.00e-3	-1.00e-3	0.00e-3	1.00e-3
Brent oil price(change in ratio)	8.30e-3	7.90e-2	2.16e-1	-2.34e-1	-3.99e-2	1.49e-2	6.45e-2
West Texas oil price(change in ratio)	7.98e-3	8.18e-2	2.38e-1	-2.18e-1	-4.46e-2	1.17e-2	5.63e-2

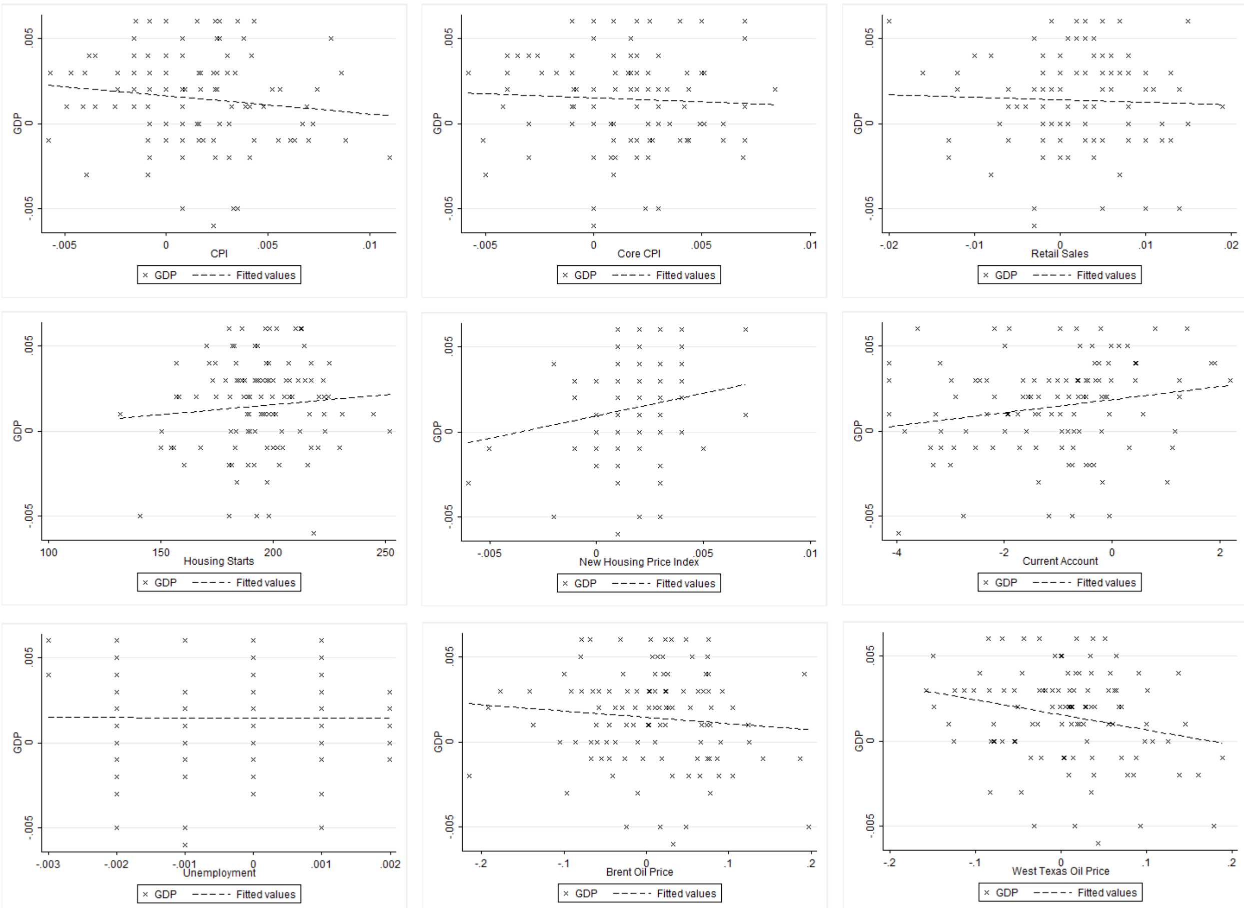


Figure 5: Relations between GDP and other indicators

Table 3: Granger causality test from other indicators to GDP

Indicators	Lag=1		Lag=2		Lag=3		Granger causality
	χ^2 value	p-value	χ^2 value	p-value	χ^2 value	p-value	(α = 0.05)
Monthly CPI	1.4549	0.228	1.8726	0.392	2.9844	0.394	NO
Monthly Core CPI	2.6888	0.101	2.6754	0.262	4.0518	0.256	NO
Retail sales	11.482	0.001***	12.345	0.002***	14.884	0.002***	YES
Housing Starts	1.2457	0.264	0.69362	0.707	0.6727	0.880	NO
NHPI	0.72817	0.393	1.1702	0.557	4.3115	0.230	NO
Current Account	3.9589	0.047**	4.6323	0.099*	15.925	0.001***	YES
Unemployment	4.5077	0.034**	4.6759	0.097*	9.5969	0.022**	YES
Brent oil price	7.0074	0.008***	8.7551	0.013**	8.627	0.035***	YES
West Texas oil price	8.3249	0.004***	12.638	0.002***	11.262	0.010***	YES

Table 4: Granger causality test from GDP to other indicators

Indicators	Lag=1		Lag=2		Lag=3		Granger causality
	χ^2 value	p-value	χ^2 value	p-value	χ^2 value	p-value	(α = 0.05)
Monthly CPI	0.24548	0.620	0.43324	0.805	5.1795	0.159	NO
Monthly Core CPI	0.33657	0.562	0.05562	0.973	1.6283	0.653	NO
Retail sales	0.56293	0.453	2.189	0.335	3.4177	0.332	NO
Housing Starts	0.71918	0.396	12.734	0.002***	10.497	0.015**	YES
NHPI	1.0666	0.302	5.6309	0.060*	9.6966	0.021**	YES
Current Account	1.2141	0.271	2.6933	0.260	9.0075	0.029**	YES
Unemployment	0.28035	0.596	1.8002	0.407	1.1622	0.762	NO
Brent oil price	0.00063	0.980	0.11091	0.946	1.6367	0.651	NO
West Texas oil price	0.39793	0.528	0.55157	0.759	1.5543	0.670	NO

Table 5: the improved CUTE model

$\{x_n\}$	$\{y_n\}$	$\Delta_{\{y_t\}\to\{x_t\}}$	$\Delta_{\{x_t\}\to\{y_t\}}$	$\lvert\Delta_{\{y_t\}\to\{x_t\}} - \Delta_{\{x_t\}\to\{y_t\}}\rvert$	p-value
Monthly CPI	Monthly GDP	66.33	65.80	0.753	0.9334
Monthly Core CPI	Monthly GDP	68.52	70.14	1.620	0.3253
Retail sales	Monthly GDP	70.32	70.49	0.170	0.8888
Housing Starts	Monthly GDP	68.39	91.12	22.73	0.0000***
NHPI	Monthly GDP	36.11	35.36	0.75	0.5946
Current Account	Monthly GDP	75.44	71.71	3.730	0.0754*
Unemployment	Monthly GDP	75.02	93.50	18.48	0.0000***
Brent oil price	Monthly GDP	66.92	96.12	29.20	0.0000***
West Texas oil price	Monthly GDP	70.67	99.97	29.30	0.0000***

of the equilibrium (National Income Identity) and feedback mechanism of macroeconomy. If GDP increases, indicating an increase in productivity, the current account may increase since the gap between domestic and foreign productivity grows. On the other hand, current account is always believed to be one of the sources of economic growth. The export of goods brings influx of currencies and the efficiency in each market also grows, thus leading an increasing GDP. Retail sales is a major part of Canadian economy as mentioned before and it's quite plausible to show a unidirectional Granger causality. The result from unemployment is consistent with prior empirical research and macroeconomic theories. In our experiment period, the role of Canada in international crude oil market is unchanged. The change in oil price thus shows high causality towards GDP. For housing starts and new house price index, when GDP grows, there are more spare capital that could be invested in housing sector.

As for improved CUTE model, the conclusions around unemployment and two measures of international crude oil price are the same as those drawn by traditional Granger causality test, which provides another empirical evidence of our improved CUTE model's validity. For current account, if we choose a significant level to be 5%, i.e. $\alpha = 0.05$, the causality cannot be determined through improved CUTE. This exposes one of a disadvantage for improved CUTE, that is , the series with bidirectional and those with no causal relations at all are mixed together. If improved CUTE tells us that the causal relations between two series are not clear, there are two possible explanations: no causal relations or bidirectional causality. Another different result is for retail sales. Although in traditional Granger causality retail sales has an significant impact on GDP, improved CUTE shows a rather high p-value (0.8888).Plus, the result around housing starts goes to an opposite direction. It is quite difficult to say which test reveals the *true* result, but the difference between two instruments warns us to be more careful about the relations between the change in retail sales (housing starts) and the change in GDP.

5. Conclustions

In the big data environment, we can obtain a variety of economic growth-related data from government websites, professional and academic databases, business platforms and other various related data sources. How to distinguish the causal relationship among macroeconomics variables from these complicated and diverse data is an important research topic for predicting changes in economic development trends. The traditional Granger causality test model has certain limitations. It can only be applied to time series passing ADF test and cointegration test, and it is necessary to determine the lag order in advance, and then for different economic variables, the lag order may vary greatly, leading to some controversy about the accuracy of the Granger causality test results.

In this paper, a novel model named CUTE is implemented, which combine Granger causality to information compression theory, and a multi-value expansion is done to fit the macroeconomics variavbles better. We use the improved CUTE model to determine causality in Canadian macroeconomy from Januray, 2009 to May, 2018. Our findings are basically consistent with traditional Granger causality test but there are also differences. The improved CUTE indicates an opposite causality between GDP and housing starts and undecided causality between retail sales and GDP.

Although the improved CUTE shows a great number of advantages, such as the robustness towards high noise term and need not to predetermine the order of lag operators and pass ADF and cointergration tests, there are some disadvantages brought by it. First, for encoding process, the causal relations are limited to change-to-change pattern and characteristics which bother the detection of causality such as seasonality are not considered. A better encoding process (to preprocess the disturbance effect) should be drawn. Second, there are only three states in improved CUTE model while four in traditional Granger causality. Bidirectional causality is unable to detect through improved CUTE. Maybe a conbination between the difference of Δs and the absolute of Δs could be combined and draw another more detailed criteria. Third, when the improved CUTE tells a different story from traditional Granger causality, it is hard to say which one is *correct* or even both of them are *wrong*. However, it does give us a good reference on macroeconomic causality from a quite different and interdisciplinary perspective from Granger causality. Conclusions are likely to be intensified when two models offer the same outcomes.

References

Amiri, A., Ventelou, B., 2012. Granger causality between total expenditure on health and gdp in oecd: Evidence from the toda–yamamoto approach. Economics Letters 116, 541–544.

Baffigi, A., Golinelli, R., Parigi, G., 2004. Bridge models to forecast the euro area gdp. International Journal of forecasting 20, 447–460.

Bańbura, M., Giannone, D., Reichlin, L., 2010. Large bayesian vector auto regressions. Journal of Applied Econometrics 25, 71–92.

Budhathoki, K., Vreeken, J., 2018a. Causal inference on event sequences, in: Proceedings of the 2018 SIAM International Conference on Data Mining, SIAM. pp. 55–63.

Budhathoki, K., Vreeken, J., 2018b. Origo: causal inference by compression. Knowledge and Information Systems 56, 285–307.

Carruth, A.A., Hooker, M.A., Oswald, A.J., 1998. Unemployment equilibria and input prices: Theory and evidence from the united states. Review of economics and Statistics 80, 621–628.

Elder, J., Serletis, A., 2009. Oil price uncertainty in canada. Energy Economics 31, 852–856.

Elvidge, C.D., Baugh, K.E., Kihn, E.A., Kroehl, H.W., Davis, E.R., Davis, C.W., 1997. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. International Journal of Remote Sensing 18, 1373–1379.

Ghosh, S., 2009. Electricity supply, employment and real gdp in india: evidence from cointegration and granger-causality tests. Energy Policy 37, 2926–2929.

Ghysels, E., Hill, J.B., Motegi, K., 2016. Testing for granger causality with mixed frequency data. Journal of Econometrics 192, 207–230.

Gong, M., Zhang, K., Schölkopf, B., Glymour, C., Tao, D., 2017. Causal discovery from temporally aggregated time series, in: Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence, NIH Public Access.

Gunduz*, L., Hatemi-J, A., 2005. Is the tourism-led growth hypothesis valid for turkey? Applied Economics Letters 12, 499–504.

Hamilton, J.D., 1996. This is what happened to the oil price-macroeconomy relationship. Journal of Monetary Economics 38, 215–220.

Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F., Danks, D., 2017. A constraint optimization approach to causal discovery from subsampled time series data. International Journal of Approximate Reasoning 90, 208–225.

Iacoviello, M., Neri, S., 2010. Housing market spillovers: evidence from an estimated dsge model. American Economic Journal: Macroeconomics 2, 125–64.

Jiménez-Rodríguez*, R., Sánchez, M., 2005. Oil price shocks and real gdp growth: empirical evidence for some oecd countries. Applied economics 37, 201–228.

Jumbe, C.B., 2004. Cointegration and causality between electricity consumption and gdp: empirical evidence from malawi. Energy economics 26, 61–68.

Khan, H.H.A., Abdullah, H., Samsudin, S., 2017. The relationship between household debt composition and gdp in malaysia. Pertanika Journal of Social Sciences & Humanities .

Korhonen, I., Ledyaeva, S., 2010. Trade linkages and macroeconomic effects of the price of oil. Energy Economics 32, 848–856.

Kotlowski, W., Grunwald, P., 2012. Sequential normalized maximum likelihood in log-loss prediction, in: Information Theory Workshop (ITW), 2012 IEEE, IEEE. pp. 547–551.

Lee, C.C., 2005. Energy consumption and gdp in developing countries: a cointegrated panel analysis. Energy economics 27, 415–427.

Lendle, A., Olarrega, M., Schropp, S., Vézina, P.L., 2013. ebays anatomy. Economics Letters 121, 115–120.

Li, J., Li, Z., 2011. A causality analysis of coal consumption and economic growth for china and india. Natural Resources 2, 54.

Lozano, S., Gutierrez, E., 2008. Non-parametric frontier approach to modelling the relationships among population, gdp, energy consumption and co2 emissions. Ecological Economics 66, 687–699.

Oh, W., Lee, K., 2004. Causal relationship between energy consumption and gdp revisited: the case of korea 1970–1999. Energy economics 26, 51–59.

Oki, T., Kanae, S., 2006. Global hydrological cycles and world water resources. science 313, 1068–1072.

Omri, A., Nguyen, D.K., Rault, C., 2014. Causal interactions between co2 emissions, fdi, and economic growth: Evidence from dynamic simultaneous-equation models. Economic Modelling 42, 382–389.

Pao, H.T., Tsai, C.M., 2011. Multivariate granger causality between co2 emissions, energy consumption, fdi (foreign direct investment) and gdp (gross domestic product): evidence from a panel of bric (brazil, russian federation, india, and china) countries. Energy 36, 685–693.

Ryabko, B., Astola, J., . Application of data compression methods to hypothesis testing for ergodic and stationary processes, in: International Conference on Analysis of Algorithms DMTCS proc. AD, p. 408.

Soytas, U., Sari, R., 2003. Energy consumption and gdp: causality relationship in g-7 countries and emerging markets. Energy economics 25, 33–37.

Zhang, H., Zhou, S., Zhang, K., Guan, J., 2017. Causal discovery using regression-based conditional independence tests., in: AAAI, pp. 1250–1256.

Zhang, Q., Zeng, W., Liu, L., 2014. Research on the inner relationship between chinese economic growth and energy consumptionbased on granger causality test of the vecm model and grey relevance analysis. Contempory Economic Management 36, 30–34.