

Variable Selection Methods for Multivariate Time Series Data Using Multivariate Granger Causality

Keita Ohmori

Kyushu Institute of Technology, 680-4 kawazu, Iizuka, Fukuoka, 820-8502, Japan

SUMCO, 1-52 Kubara, Yamashiro, Imari, Saga, 849-4256, Japan

Toshiki Saitoh

Kyushu Institute of Technology, 680-4 kawazu, Iizuka, Fukuoka, 820-8502, Japan

Akiko Fujimoto

Kyushu Institute of Technology, 680-4 kawazu, Iizuka, Fukuoka, 820-8502, Japan

Eiji Miyano

Kyushu Institute of Technology, 680-4 kawazu, Iizuka, Fukuoka, 820-8502, Japan

Email: keita.ohmori468@mail.kyutech.jp, toshikis@ai.kyutech.ac.jp, fujimoto@ai.kyutech.ac.jp, miyano@ai.kyutech.ac.jp

Abstract

We study variable selection methods for multivariate time-series data. Hmamouche et al. proposed a method that first constructs a causal graph based on Granger causality among time-series data, and then selects variables from clusters formed by clustering the vertices corresponding to each variable. However, this method only performs pairwise Granger causality tests, which may not fully capture the interactions among variables. To address this issue, we propose a variable selection method that performs multivariate Granger causality tests on all combinations of explanatory variables with respect to the target variable, selecting the combination with the strongest causality. Our method successfully constructs a predictive model with a higher accuracy compared to the previous method.

Keywords: Granger causality, Variable selection, Multivariate time series data

1. Introduction

In recent years, with the proliferation of the IoT, a vast amount of diverse time-series data has been accumulated through sensors in the manufacturing industry. These data are utilized for optimizing manufacturing processes and improving quality, such as analyzing the causes of equipment failures and quality degradation. Machine learning is primarily employed for these analyses, where variable selection is a crucial element, considering computational cost, recognition accuracy, and model interpretability. Moreover, since time-series data inherently involve past data influencing current values, machine learning models for multivariate time-series data must clarify the extent to which the past data of each variable affects the current and future values of the target variable. One method for evaluating such influence is the Granger causality test [1]. The Granger causality test is a statistical technique that examines how well the past information of one variable can explain the future values of another variable. Previous studies have proposed approaches utilizing the Granger causality test for variable selection. For example, in 2015 Sun et al. [2] proposed a method that performs univariate Granger

causality tests between the target variable and explanatory variables, selecting all explanatory variables that have a causal relationship with the target variable. In 2018, Hmamouche et al. [3] proposed a method that constructs a causal graph based on Granger causality among time-series data and selects variables by clustering the vertices corresponding to each variable. However, these methods have limitations, as it only performs pairwise Granger causality tests, which makes it challenging to adequately account for interactions among multiple variables.

To address this issue, we propose a new variable selection method that performs multivariate Granger causality tests on all possible combinations of explanatory variables in the dataset and selects the combination with the strongest causality with respect to the target variable. This new method successfully constructs a predictive model with higher accuracy than traditional methods by considering the interactions among variables. Consequently, it overcomes the limitations of the previous method and expands the possibilities for variable selection in time-series data analysis.

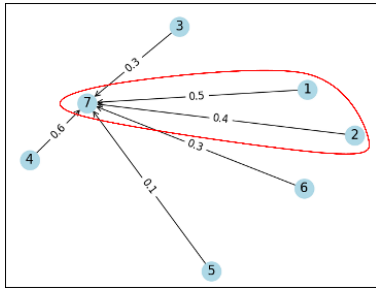


Fig. 1. Causal graph

2. Preliminary

2.1. Granger Causality

Granger causality is a method for identifying causal relationships between time series data, evaluating whether a causal variable contributes to predicting the outcome. Specifically, a variable x is said to causally influence another variable y if a regression model using both the past values of x and y significantly outperforms a model using only the past values of y . This relationship is determined by comparing two regression models using statistical methods such as the F-test. In this study, we use Vector Auto Regression (VAR) model [4] for regression, and the two regression models are expressed by the following equations:

Model 1:

$$y_t = \alpha_0 + \sum_{i=1}^l A_i y_{t-i} + \varepsilon_t, \quad (1)$$

Model 2:

$$y_t = \alpha_0 + \sum_{i=1}^l A_i y_{t-i} + \sum_{i=1}^l B_i x_{t-i} + \varepsilon_t. \quad (2)$$

Where α_0 represents the constant term, for each $i \in \{1, \dots, l\}$, A_i and B_i represent the regression coefficients, l represents the maximum number of lags for x and y , and ε represents the error term (white noise). The accuracy of the two regression models was tested using the F-test. The test statistic is expressed as follows:

$$F = \frac{\frac{RSS_1 - RSS_2}{l}}{\frac{RSS_2}{n - 2l - 1}}. \quad (3)$$

Where RSS_1 and RSS_2 the residual sum of squares for Model 1 and Model 2, respectively, l is the number of lags for x , and n is the sample size. The value F follows an F-distribution with degrees of freedom $(n, n - 2l - 1)$. If the calculated F exceeds significance level, we reject the null hypothesis that x does not cause y and conclude that x causes y . The strength of causality is defined as $1 - p$, where p is the p -value from the Granger causality test.

This value ranges from 0 to 1, with value closer to 1 indicating stronger causality.

2.2. Hmamouche's method

Hmamouche's method [2] first constructs a causal graph based on Granger causality among time-series data and then selects variables from clusters formed by clustering the vertices corresponding to each variable. This method is referred to as the GSM (Granger selection method) algorithm. The outline of the GSM algorithm is as follows:

GSM algorithm

Input: A set of explanatory variables $X = \{x^1, x^2, \dots, x^n\}$, Min-Causality threshold T , the selection size k and a target variable y

Output: the selected variables associated to y

Step 1: Perform Granger causality tests distinct pairs of variables x^i and x^j and construct a causality matrix. Each element a_{ij} of the matrix A is the strength of the causality of x^i and x^j . If the strength of causality is smaller than the threshold T , it is considered as no causality.

Step 2: Partition the input variables X from the causality matrix constructed in Step 1 using the PAM (Partitioning Around Medoids) method. The goal is to group variables by minimizing causalities between clusters and maximizing causalities within clusters.

Step 3: Choose one element from each cluster that has the strongest causality on the target variable.

3. Our Method

3.1. Multivariate Granger causality

In our method, we use a multivariate Granger causality test, which is an extension of the Granger causality test. The multivariate Granger causality test evaluates not only the causality between two variables but also the causality of a combination of multiple variables on another variable. Fig.1 illustrates a causal graph representing the causal relationships among time-series data. The explanatory variables are labeled 1 through 6, and the target variable is labeled as 7. The values annotated on each edge indicate the strength of causality. In Fig. 1, none of the individual explanatory variables exhibit causality with the target variable. However, when variables 1 and 2 are combined, a causal relationship with the target variable is observed. This demonstrates that, while individual explanatory variables may not exhibit causality with the target variable, combinations of multiple variables exhibit causal relationships. The multivariate Granger causality test is capable of capturing such causal relationships. The two models compared in

the multivariable Granger causality test can then be the as follows:

$$\text{Model 1: } y_t = \alpha_0 + \sum_{i=1}^l A_i y_{t-i} + \varepsilon_t, \quad (4)$$

$$\text{Model 2: } y_t = \alpha_0 + \sum_{i=1}^l A_i y_{t-i} + \sum_{j=1}^k \sum_{i=1}^l B_i^j x_{t-i}^j + \varepsilon_t. \quad (5)$$

Where a set of explanatory variables $X = \{x^1, x^2, \dots, x^k\}$, a target variable y , a constant term α_0 , for each $i \in \{1, \dots, l\}$ and $j \in \{1, \dots, k\}$, A_i and B_i^j represents the regression coefficients for y and x , l represents the maximum number of lags for y and x , and ε represents the error term (white noise). The accuracy of the two regression models was tested using the F-test. The test statistic used is the same as in (3).

3.2. Procedure

We perform multivariate Granger causality tests on the target variable using all possible combinations of explanatory variables and select the combination that exhibits the strongest causality with respect to the target variable. The algorithm for our method is presented below:

Our Algorithm

Input: A set of explanatory variables $X = \{x^1, x^2, \dots, x^n\}$, Min-Causality threshold T , the selection size k and a target variable y

Output: The combination of explanatory variables that exhibits the strongest causal relationship with the target variable.

Step 1: Generate all combinations of k variables from the explanatory variables.

Step 2: For each combination of explanatory variables, perform a multivariable Granger causality test to assess their influence on the target variable.

Step 3: Compute the causality strength s as $1-p$, where p is the p-value for multivariate Granger causality test, and consider it significant if $s \geq T$. Retain the combination of explanatory variables X' and the corresponding causality strength as the result.

Step 4: Identify the combination with the highest causality strength and return it as the best explanatory variable set for predicting y .

4. Experiments

4.1. Evaluation Measure

We evaluate our method by comparing the variables selected using our method with that of Hmamouche's method. For this, prediction models are constructed using a Vector Error Correction Model (VECM) [5] with the selected variables, and the predictive accuracy of these

models are used as the evaluation criterion. For accuracy evaluation, we used the Normalized Root Mean Square Error (NRMSE). VECM model extends the VAR model by considering the non-stationarity of time series and incorporating cointegration equations. If two time series (x_t, y_t) follow a first-order integration process ($I(1)$), VECM can be expressed as:

$$\Delta y_t = \alpha_{0y} - \gamma_y(\beta_0 y_{t-1} - \beta_1 x_{t-1}) + \sum_{i=1}^p v_{iy} \Delta y_{t-j} + \sum_{i=1}^p w_{iy} \Delta x_{t-j} + \varepsilon_t, \quad (6)$$

$$\Delta x_t = \alpha_{0x} - \gamma_x(\beta_0 y_{t-1} - \beta_1 x_{t-1}) + \sum_{i=1}^p v_{ix} \Delta y_{t-j} + \sum_{i=1}^p w_{ix} \Delta x_{t-j} + \varepsilon_t. \quad (7)$$

Where $(\Delta y_t, \Delta x_t)$ are $(y_t - y_{t-1}, x_t - x_{t-1})$, $(\alpha_{0y}, \alpha_{0x})$ represent the constant term, (v_{iy}, v_{ix}) represents the regression coefficients, the coefficients (β_0, β_1) are the cointegrating parameters, and (γ_y, γ_x) are the error correction parameters. NRMSE is calculated using the following formula:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\text{stdev}(y)}. \quad (8)$$

Where $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$ are the forecasts, (y_1, y_2, \dots, y_n) are the real values, and n represents number of data points used for accuracy comparison.

In our method, the p-value threshold for the Granger causality test was set at 10%. The lag parameters for the VECM model and Granger causality tests are determined based on Akaike's Information Criterion (AIC) [6] The training data for the VECM model consisted of the first 90% of the dataset in chronological order, while the last 10% was used as test data.

4.2. Data Sets

A dataset used in this study is obtained from the Machine Learning Repository website [7]. It includes the returns of the Istanbul Stock Exchange along with seven other international indices: SP, DAX, FTSE, NIKKEI, BOVESPA, EU, and EM, covering the period from June 5, 2009, to February 22, 2011.

4.3. Results

Table 1 shows results of variable selection using GSM algorithm. In the GSM algorithm, specifying the number of clusters in the causal graph allows adjustment of the 5 to 7, GSM algorithm outputs BOVESPA, BOVESPA and EU, and BOVESPA, EU, and SP, respectively. The clusters necessarily contain variables with a causality score of 0.9 or higher with respect to the target variable.

Table 1: Variable selection results using the GSM

#Clusters	Selected variables
1	BOVESPA
2~4	BOVESPA, EU
5~7	BOVESPA, EU, SP

Table 2: Variable selection results using our method

# Variables	Selected variables
1	BOVESPA
2	SP, BOVESPA
3	SP, DAX, EU

Table 3: Predictive accuracy for each variable set

Methods	Selected variables	NRMSE
Previous [2], Ours	BOVESPA	1.118
Previous [2]	BOVESPA, EU	1.010
Previous [2]	BOVESPA, EU, SP	1.022
Ours	SP, BOVESPA	1.002
Ours	SP, DAX, EU	0.999

Table 2 shows the combinations of selected variables for each variable count using our method. There is no combinations with four or more variables exceeded a causality score of 0.9 for the target variable. Table 3 summarizes the predictive accuracy (NRMSE) of the models constructed using the selected variables by the methods. These results demonstrate that our method achieves higher predictive accuracy compared to previous studies. Notably, for combinations of three variables, the predictive accuracy improved by 2.3%, i.e., previously 1.022 for {BOVESPA, EU, SP} while newly 0.999 for {SP, DAX, EU}.

4.4. Discussions

These results indicates that an advantage of our method is that it enables the selection of variables that improve the accuracy of the predictive model by taking into account the interaction effects among multiple variables.

In contrast, previous methods involve clustering the causal graph to group variables with similar effects on the target variable and then selecting variables with strong causality to the target variable from those groups. However, in this method, the causal graph is constructed based on pairwise causality tests, which are likely insufficient to fully account for the effects of interactions among multiple variables.

5. Conclusion

We proposed a method for variable selection in multivariate time-series forecasting that accounts for variable interactions. Unlike previous methods relying on pairwise Granger causality tests and clustering, our method evaluates all possible combinations of explanatory variables. This enables the identification of variable sets that improve predictive accuracy by considering the combined effects of multiple variables. Experiments demonstrated that our method outperforms traditional approaches, with a 2.3% improvement in NRMSE for three-variable combinations. These results underscore the importance of considering variable interactions in predictive modeling. However, this method requires high computational costs due to testing all possible combinations.

Future efforts will aim to reduce computational cost and extend the method to handle nonlinear interactions. Incorporating nonlinear Granger causality techniques, such as those by Chen et al. [8], could further enhance its effectiveness. In conclusion, our method advances variable selection for time-series forecasting, offering more accurate predictions and deeper insights into variable dependencies. Future developments in efficiency and nonlinearity handling will broaden its applicability across domains.

References

1. C. W. Granger, "Testing for causality: a personal viewpoint", *Journal of Economic Dynamics and control*, vol. 2, 1980, pp. 329–352.
2. Y. Sun, J. Li, J. Liu, C. Chow, B. Sun, and R. Wang, "Using causal discovery for feature selection in multivariate numerical time series", *Machine Learning*, vol. 101, no. 1-3, 2015, pp. 377–395.
3. Y. Hmamouche, A. Casali, and L. Lakhal "A Causality Based Feature Selection Approach for Multivariate Time Series Forecasting", HAL Id: hal-01467523, 2018.
4. M. Quenouille, "The analysis of multiple time-series, ser", Griffin's statistical monographs & courses. Griffin, 1957.
5. S. Johansen, "Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models", *Econometrica: Journal of the Econometric Society*, 1991, pp. 1551–1580.
6. H. Akaike, "A new look at the statistical model identification", *IEEE transactions on automatic control*, vol. 19, no. 6, 1974, pp. 716–723.
7. M. Lichman, "UCI machine learning repository", 2013, [accessed: 2024-11-01].
8. Y. Chen, G. Rangarajan, J. Feng and M. Ding, "Analyzing multiple nonlinear time series with extended Granger causality", *Phys. Lett. A* 324(1), 2004, pp. 26–35

Authors Introduction

Mr. Keita Ohmori



He received Bachelor's degree in Science in 2019 and M.S. degree in Science in 2021 from Tohoku University. Since 2021, he has been working at SUMCO Corporation. In 2024, he enrolled as a Ph.D. student at Kyushu Institute of Technology, Japan.

Dr. Toshiki Saitoh



He received the B.S.E. degree from Shimane University in 2005, and the M.S. and Ph.D. degrees (Information Science) from Japan Advanced Institute of Science and Technology in 2007 and 2010, respectively. He is a professor at Kyushu Institute of Technology.

Dr. Akiko Fujimoto



She received the B.S., M.S., and Ph.D. in Science degrees from Kyushu University in 2005, 2007, and 2010, respectively. She is an Associate Professor of the Department of Artificial Intelligence, Kyushu Institute of Technology.

Dr. Eiji Miyano



He received the B.Eng., M.Eng., and Dr.Eng., degrees in computer science from Kyushu University in 1991, 1993, and 1995, respectively. He is currently a professor of the Department of Artificial Intelligence, Kyushu Institute of Technology.