Bayesian Maximal Information Coefficient (BMIC) to reason novel trends in large datasets

Wang Shuliang · Tisinee Surapunt

Accepted: 7 December 2021 / Published online: 11 January 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
The Bayesian network (BN) is a probability inference model to describe the explicit relationship between cause and effect, which may be examined in the complex system of rice price with data uncertainty. However, discovering the optimized structure from a super-exponential number of graphs in the search space is an NP-hard problem. In this paper, Bayesian Maximal Information Coefficient (BMIC) is proposed to uncover the causal correlations from a large data set in a random system by integrating probabilistic graphical model (PGM) and maximal information coefficient (MIC) with Bayesian linear regression (BLR). First, MIC is to capture the strong dependence between predictor variables and a target variable to reduce the number of variables for the BN structural learning of PGM. Second, BLR is responsible for assigning orientation in a graph resulting from a posterior probability distribution. It conforms to what BN needs to acquire a conditional probability distribution when given the parents for each node by the Bayes’ Theorem. Third, the Bayesian information criterion (BIC) is treated as an indicator to determine the well-explained model with its data to ensure correctness. The score shows that the proposed BMIC obtains the highest score compared to the two traditional learning algorithms. Finally, the proposed BMIC is applied to discover the causal correlations from the large data set on Thai rice price by identifying the causal changes in the paddy price of Jasmine rice. The results of the experiments show that the proposed BMIC returns directional relationships with clues to identify the cause(s) and effect(s) of paddy price with a better heuristic search.

Keywords Bayesian Maximal Information Coefficient (BMIC) · Causal correlations · Temporal sequential correlation · Bayesian linear regression · Maximal information coefficient · Thai rice price

Mathematics subject classification 62M10 · 62P20 · 91B84

1 Introduction
Uncertain data is a significant issue that arises as data grows in volume, variety, and velocity. Although uncertain data introduces noise into the observed data set and results in a deviation from correctness, it is necessary for computational and statistical analysis in predictive modeling. Indeed, price data has an uncertain value because it is collected in a time-series format that exhibits both data variability and a genuine in comprehension of the data attribute relationships within the system. Price analysis, particularly in predictive analytics models, has to contend with the difficulties inherent in data analysis due to non-stationary and noisy data. However, the price data is highly sought after for economic [1, 2] and agricultural [3, 4] research that focuses on comprehending domestic and international economies’ processes, characteristics, and future development trends. As a result, employing a causal model can aid in the analysis of data with uncertain values. Not only can correlations between data variables be used to represent cause-and-effect relationships, but they are also necessary for data analysis in order to achieve the prediction model. This paper focuses on an agricultural price, specifically the price of Thai rice, because it is intriguing to examine the relevant impacts of price variance in global market rivalry. Many important variables, such as unpredictability in productivity, climatic conditions, competitors (import and export commerce),
and domestic and international events, all contribute to the fluctuation of Thai rice prices [5]. Consequently, using a graphical model will help to describe causal relationships more accurately, which can assist in lowering price barriers and avoiding unanticipated future scenarios.

The Probabilistic Graphical Model (PGM) [6, 7] is a declarative representation of the causality of real-world situations’ relationships based on probabilistic models. It is a graph-based representation that makes use of the joint distribution to assign probabilistic directions to the random variables. Two representations patterns can be used to describe the graphical representations of distributions: Bayesian networks (BN), a directed graph and Markov networks (MN), an undirected graph. The BN is an extraordinary representation that yields a directed acyclic graph (DAG) and a set of conditional probability distributions (CPD) via Bayes’ theorem, as well as machine learning capable of dealing with domain uncertainty. The model can then proceed with inference through the particular structure for the purpose of predictive modeling. Numerous researchers have been interested in the probabilistic forecasting of uncertain data using the BN, including medical analysis [8], ecology system [9, 10], agricultural forecasting [11, 12], and economic application [13]. Unfortunately, when all observed variables are required to construct the DAG, PGM’s structured learning process encounters the NP-hard problem [14]. This implies that the search space contains a super-exponential number of graphical models. The more graphs in the search space, the more time, memory, and computational resources are required to find the optimal network structure.

Consequently, this paper will employ a data-dependent approach to scope the observed variables in the sample space. The maximal information coefficient (MIC) is an interested algorithm and reasonable choice which has been used to determine the degree of dependence [15]. The MIC detects correlation between pairwise variables in a large data set with high precision because of its heuristic properties, generality and equitability. The data correlation coefficient is expressed as an R-square ($R^2$) score in relation to the regression. The generality property allows unrestricted relationship type access, while the equitability maximizes the $R^2$ score to 1 of a noiseless function relationship. Nevertheless, the correlated data will be connected using an undirected graph to represent their relationships. Without a direction between data variables, edges cannot accurately represent causality. As a result of the foregoing, this paper emphasizes the synergy of correlated data on MIC and PGM with the goal of diagnosing the direction of the relationship by statistical inference.

To motivate this study further, Bayesian linear regression (BLR) is selected to be used in this paper for identifying the directional relationship between correlated pairwise variables, $x_i$ and $x_j$, via their dependence on a probability distribution. BLR formulates linear regression using Bayesian inference, implying that variables are dependent on one another in a linear function. It generates a posterior probability for model parameters that can be used to describe the conditional probability of correlated pairwise variables. But the linear relationship can be seen their correlations between pairwise variables in the target ($y$) and predictor ($x$). Instead of estimating a single value, the BLR declares the linear regression model using a probability distribution derived from a normal (Gaussian) distribution. The $x_i$ and $x_j$ will be tested as target and predictor variables, which means that when $x_i$ is a target variable, $x_j$ is a predictor variable, and vice versa. So, switching between being the target and predictor variables produces a different posterior probability. The direction of the relationship between correlated pairwise variables is determined by the highest posterior probability of whether $x_i$ or $x_j$ will be the target or the predictor variable. As a result, a directional relationship and conditional dependency between $x_i$ and $x_j$ can be implied.

## 2 Related methods

PGM and MIC are techniques for acquiring background knowledge.

### 2.1 Probabilistic Graphical Model

Probabilistic Graphical Model (PGM) framework [6, 7] establishes a graph-based model of a probability distribution as a data structure on a conditional basis. The graph representation is a collection of independencies found in the distribution that may satisfy the conditional independence in response to a specific situation. Additionally, it serves as a skeleton of factorization for a high dimension of joint distribution. The joint distribution of conditional independence can be used for inference, which computes the posterior probability of a random variable given the evidence according to a particular algorithm. Utilizing the conditional independencies for inference is more efficient than directly manipulating the joint distribution.

Fortunately, the PGM facilitates model construction using a data-driven approach. Rather than being constructed manually by humans, learning from data enables more efficient model construction. Certain domains, in particular, prefer to learn a probabilistic model based on a distribution that incorporates prior experience (prior). The ability to perform with a set of historical data can be aided by data-driven learning. Thus, the PGM enables the manipulation of the uncertainty inherent in real-world situations via a reasonable probability model-based construction. These three benefits of representation, inference, and learning can all be interpreted as indicating the critical nature of PGM utility.
The PGM can be visualized as a Bayesian network (BN) or a Markov network (MN) to illustrate relationships between observable variables using a fundamental concept of probabilistic and graph theories. Both models can be viewed as a collection of conditional independencies and encode a factorization compactly, but they differ in terms of induce. According to graph theory [16], a graph $G = (V, E)$ naturally consists of a collection of nodes and edges, where $V$ denotes vertices or nodes and $E$ denotes edges. A graph is a data structure in which a set of variables, $V = x_1, ..., x_n$, are connected by an edge. Each pair of variables in a data set, $x_i$ and $x_j$, can be connected via either a directed edge $x_i \rightarrow x_j$ or an undirected edge $x_i - x_j$. The directed edge indicates the direction of the relationship, whereas the undirected edge implies ignorance of the direction of the relationship as shown in Fig. 1.

The BN is a type of directed graph that has the property of being acyclic and having no self-relationship. It is known as a directed acyclic graph (DAG). The model structure is intended to place an emphasis on the usability of causal inference. The BN representation makes use of Bayes’ theorem [17] to determine the posterior probability associated with a particular distribution, which is advantageous for describing causal relationships. Besides, it is applicable to the creation of predictive models for uncertain data that can account for their dependencies via their probability values. The BN structure has been utilized in maximizing the classification accuracy and information [18] and also evaluating the risk quantification with the DEMATEL Model [19]. Initially, the basic graphical model of BN consists of at least two nodes connected by a single edge. BN’s edge representation depicts a direction by using an arrow to convey the most accurate understanding of the causal relationship. In 1(a), an arrow connecting node $x_i$ to node $x_j$, implying that node $x_i$ is the cause of node $x_j$ or that node $x_j$ is affected by node $x_i$. By contrast, MN is intended to be an undirected graph that demonstrates only the variable association without regard for direction. Thus, in 1 (b), nodes $x_i$ and $x_j$ cannot be inferred to have any causal relationship because they must be imagined as either the cause node or the affected node, $x_i \rightleftharpoons x_j$.

In addition, the BN has been constructed naturally by leveraging the distribution’s conditional independence property. It expresses the causative model succinctly and graphically in order to infer the causes and effects between variables. It is not only designed for normal cases but also for uncertainty prediction. In terms of the explicit representation of BN, the objective is to use the conditional distribution to represent the joint probability distribution over a set of random variables. Bayes’ theorem will be used to support the model construction by estimating probability values. It is recommended when causal reasoning or prediction are involved.

The Bayes’ theorem is derived from the conditional probability definition, which is consistent with the joint distribution’s symmetry property. The BN is used to perform Bayesian inference in the presence of uncertainty and to handle numerous relationships between random variables (factors). The Bayes’ theorem is usually used in the financial domain to aid in risk evaluation. The key concept of Bayes’ theorem is that a result of conditional distributions can be simply defined in terms of joint distributions. Prior probability distributions are required to calculate the posterior probability using Bayes’ theorem. Prior probability refers to the probability of an event occurring prior to the collection of new data. While the probability of events occurring after the new data is received will return the posterior probability.

As a result, the Bayes’ theorem benefits revising predictions by allowing for the updating of probabilities in the presence of new information. In prediction terminology, information about occurring events is based on hypotheses and evidence, which refer to the probability of a hypothesis, $H$, based on given evidence, $E$, as defined in definition 1 [17].

**Definition 1** [The Bayes’ theorem on prediction perspective] \[ P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \] (1)

Where $P(H \mid E)$ is a posterior probability. $P(E \mid H)$ is a likelihood probability. $P(H)$ is a prior probability. And $P(E)$ is a marginal probability or model evidence.

To consider the probability of constructing a BN, $S$ denotes the structured BN and $\theta_S$ denotes the associated conditional probabilities. Then, $P < S, \theta_S >$ is defined as the joint probability distribution of all random variables contained within a constructed network. Accordingly, the set of vertices, $V$, is composed of a set of random variables $x_1, x_2, \ldots, x_n$ while the edge, $E$, connects two random variables, representing conditional independence. Thus, each
Bayesian Maximal Information Coefficient (BMIC) to reason novel trends in large datasets

variable \( x_i \) in a set of \( V \) has the CPD as \( P(x_i, \text{Pa}(x_i)) \) where \( \text{Pa}(x_i) \) denotes the set of parent variables of variable \( x_i \). It is a term that can be defined in definition 2.

**Definition 2** [The probability of a BN construction]

\[
P(S, \theta_R) = P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \text{Pa}(x_i)) \tag{2}
\]

BN therefore draws edges among random variables by learning from the observed data sample to represent their dependencies. The PGM framework contains two learning approaches, parameter learning and structure learning, which emphasize the attributes of dependence in addition to the conditional probability estimation process. However, the objective of both learning algorithms is to build the DAG over the observable data set using different methods and target data sets. Not only does the BN gain knowledge from a pure data sample, but it can also learn about existing relationships and use that information to build a network model.

2.1.1 **Parameter learning**

The parameter learning approach necessitates sample data with an existing DAG in order to capture the conditional probability of individual variables. Maximum likelihood estimation (MLE) and Bayesian estimation are used to calculate the conditional probability. MLE returns the maximum-likelihood conditional probability of data given its the model, \( P(\text{Data} \mid \text{Model}) \). Simultaneously, Bayesian estimation returns the conditional prior distribution.

2.1.2 **Structure learning**

The structure learning approach captures the maximum probability of variable dependencies, although it is constrained by the lack of known prior knowledge. The solution is to use the uniform prior, which has equal probability. In this learning approach, the network model or graph, \( G \), is determined by the posterior probability distribution associated with the data sample, \( D \), as \( P(G \mid D) \). According to Bayesian statistical decision theory, the graphs will grow into an exponential number of DAGs. Following that, two different learning algorithms, constraint-based and score-based, will consider only the most optimized and best-performing network structure.

(a) **Constraint-based learning** A constraint-based learning approach is intuitively structured around conditional independencies that do not ignore the Bayesian Network concept. Instead of testing for probability, the conditional independence test is performed using Pearson’s \( \chi^2 \)-test, Fisher’s Z-test, and t-test. This constraint-based learning approach models the network in three stages: conditional independence identification, undirected relationship skeleton learning, and arcs direction learning. In [13], The Markov blanket was first introduced for the purpose of optimizing the number of candidate nodes in a DAG. The Markov blanket’s conditioned edges can be used to determine the conditional independence of any set of nodes. It benefits skeleton learning by identifying the undirected edges in DAG. The direction will then be assigned to the undirected edges, resulting in a complete partial DAG. So, the resulting model implies the existence of a causal relationship between nodes.

(b) **Score-based learning** The score-based learning approach focuses on using a network score to determine the best BN for the data. Given the observed data set, the maximal score returns the graph’s highest posterior probability.

\[
P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)}
\]

Since the number of possible DAGs (BNs) has grown exponentially in the search space, \( O(n!2^n) \) [20], where \( n \) is the number of nodes in the network, heuristic search algorithms are responsible for reducing the number of DAGs in order to find the optimal BN structure. The heuristic search algorithms include options for specific purposes, such as hill-climbing, Tabu, genetic, and greedy equivalent searches (GES) which can be calculated using AIC, BIC [21], K2 [22], and Bdeu [23] as an indicator the scoring function. The emphasis in [24] was on operational efficiency by comparing each scoring method of each heuristic search algorithm. The findings revealed that different types and interest domains can have an impact on learning performance. The score-based mechanism can demonstrate the dependencies of the entire structural model and avoid the failure of individual conditional independence, although only a few heuristic search algorithms work well with scoring methods. As a result, PGM can be used to describe the dependencies between random variables, as it supports conditional probability. The difficult task of identifying the optimal model encounters an NP-hard problem, which subjects all networks in the search space to heuristic-search algorithms for scoring comparison. The model with the highest score will be declared to be the best model. Even if the NP-hard problem requires a time-consuming search for the best network, the PGM framework remains a robust framework for illustrating causal relationship models. BN, in particular, can improve users’ comprehension through the intuitive interpretation of models in a domain [25].
2.2 Maximal information coefficient

The maximal information coefficient (MIC) [15] is a statistical measurement for determining the relationship between pairwise variables in a large data set. It is referred to as an exploratory data analysis (EDA) tool. The MIC score is used to quantify dependence between pairwise variables and to determine the strength of the dependent relationship. To calculate the MIC score, a grid line drawn on the scatter diagram of two variables serves as a data partitioning tool that allows for the exploration of all grids up to the maximum grid resolution.

Given the finite data set $D$ of ordered pairs, $x$ and $y$, which $x$- and $y$-values of $D$ are partitioned to be $x$-by-$y$ on grid $G$. The distribution, $D|G$, then occurs from the points in $D$ on the grid $G$, implying that the probability distribution in each cell is the fraction of points in $D$ in that cell. So, the fixed finite set $D$ with different grids $G$ will result in a different distribution of $D|G$. The defining MIC will be explored in detail by first making the following definition [26].

**Definition 3** [A finite data set $D \subset \mathbb{R}^2$ and $x$, $y$ are positive integers]

$$I^*(D, x, y) = \max_{x,y \in \mathbb{R}} I(D|G)$$ (3)

where the maximum is on the grid $G$ with $x$ and $y$ axis and $I(D|G)$ is mutual information of $D|G$.

Then, the characteristic matrix and the MIC of $D$ are defined in term of $I^*$. 

**Definition 4** [The characteristic matrix $M(D)$] The characteristic matrix $M(D)$ of the 2 variables, $x$ and $y$, is an infinite matrix with entries

$$M(D)_{x,y} = I^*(D, x, y) / \log \min(x, y)$$ (4)

Therefore, the MIC of a set $D$ of two variables’ data in the sample size $n$ and grid size less than $B(n)$ is denoted in definition 5.

**Definition 5** [Maximal information coefficient (MIC)]

$$MIC_{xy}(D) = \max_{x,y \in \mathbb{R}} I^*(D, x, y) / \log_2 \min(x, y)$$ (5)

where $\omega(1) < B(n) \leq O(n^{1-\epsilon})$, for some $0 < \epsilon < 1$.

As a result, $B(n)$ is the maximum size of the grid. A value for $B(n)$ that is set too high may result in non-zero scores even when testing with random data, as each data point will have its own cell. Otherwise, they concentrate exclusively on a single simple pattern. Therefore, $B(n)$ is set to $n^{0.6}$ by default, based on a heuristic suggestion by the founder team.

| MIC score | Strength type |
|-----------|---------------|
| 1         | Perfect correlation |
| 0.81-0.99 | Strong correlation |
| 0.71-0.80 | Good correlation |
| 0.51-0.70 | Weak correlation |
| 0.01-0.50 | Poor correlation |
| 0         | No correlation |

Three fundamental properties of MIC follow the mutual information properties. To begin, each pair of integers $(x,y)$ can be used to find the largest mutual information using an $x$-by-$y$ grid that is normalized to allow for fair comparisons between grids and generates a score value between 0 and 1. Second, the symmetrical mutual information, $I(x, y) = I(y, x)$, implies that the $x$- and $y$-values of $D$ can be interchanged. Thirdly, the characteristic matrix will remain constant regardless of how the $x$- and $y$-values of $D$ are transformed, as the result of distribution $(D|G)$ will be determined solely by the order of the data.

As mentioned, the MIC score is normalized from mutual information and ranges from 0 to 1. A MIC score of 0 indicates an independent relationship. The increasing magnitude of the MIC score, up to 1, represents the increased strength of dependence between $x$ and $y$ variables, which is expressed in Table 1 as the strength type. They are typically classified into five to six levels [27, 28], which include no correlation and perfect correlation.

When compared to other algorithms, the MIC algorithm produces a significant score due to two of its heuristic properties: generality and equitability. The generality property enables relationship types to have unrestricted access. MIC is capable of detecting both linear and non-linear relationships, whereas other algorithms, such as distance correlation [29] and Pearson’s $R$ [30], are only capable of detecting linear relationships. Furthermore, the equitability property assigns a similar MIC score to different relationship types with similar noise levels. MIC is discovered using mutual information but is not estimated using mutual information. In almost all noise models, it is a directly implemented equitability dependence measure that produces a more significant result than mutual information and others. In Fig. 2, the correlation coefficient illustrates the score comparison between MIC and other current functions in noiseless models. The MIC algorithm is shown to be capable of detecting an equal score of 1.00 on all different noiseless model functions except the random function.

MIC is described as an equitable measure of dependence based on two heuristic properties. The MIC equitability property is exceptional when considering the variable dependence of noise models with a variety of function types and sample sizes ranging from $n = 250$ to $n = 1,000$. 

 Springer
Bayesian Maximal Information Coefficient (BMIC) to reason novel trends in large datasets

There are intuitive reasons for including maximization and normalization in the MIC definition. Both features are necessary for the MIC algorithm to be able to present the effectiveness of dependence measurement [15, 31]. Therefore, [31] proves the omitting of both maximization and normalization of the original MIC in definition 5, which can present effects on the equitable measure over a noisy relationship. There are three distinct MIC variants: maximization is omitted, normalization is omitted, and maximization and normalization are omitted.

The first variant of MIC omits the maximization as $\text{MIC}_1$.

**Definition 6** Let $D$ is a data set of ordered pairs, $x$ and $y$, and let $E(D, x, y)$ is an equipartition of $D$. The number of rows and columns of $E(D, x, y)$ should have the same number of points of $D$. Let $I(\cdot)$ is a mutual information of $E(D, x, y)$, $I^E(D, x, y) = I(D|E(D, x, y))$. Then $\text{MIC}_1$ denotes the variant of MIC which omits the maximization over grid partitions by

$$\text{MIC}_1(D) = \max_{xy \in E(n)} I^E(D, x, y) / \log_2 \min(x, y)$$  \hspace{1cm} (6)

Second, the omission of normalization as $\text{MIC}_2$ considers the $\log_2 \min(x, y)$ which is an upper bound on the maximum mutual information of an $x$-by-$y$ grid, a uniform of all grids. Thus, the normalization returns values between 0 and 1 and compares values between grids of various resolutions.

**Definition 7** Let $D$ is a data set of ordered pairs, $x$ and $y$, and let $I^*$ is the same as in definition 5. Then $\text{MIC}_2$, which is without normalization, is defined by

$$\text{MIC}_2(D) = \max_{xy \in E(n)} I^*(D, x, y)$$  \hspace{1cm} (7)

Thirdly, both maximization and normalization are omitted in the variant of $\text{MIC}_3$.

**Definition 8** Let $D$ is a data set of ordered pairs, $x$ and $y$, and let $I^E$ is the same as in definition 6. Then $\text{MIC}_3$, which is the third variation of MIC omitting both maximization and normalization, is given by

$$\text{MIC}_3(D) = \max_{xy \in E(n)} I^E(D, x, y)$$  \hspace{1cm} (8)

In Fig. 3, the three MIC variants omitting maximization and normalization were tested over the noisy relationship to show the coefficient of $R^2$ score. When compared to all three variants, $\text{MIC}_1$, $\text{MIC}_2$, and $\text{MIC}_3$, the original MIC (first column of plots) clearly exhibits the equitable behavior that maximization and normalization are required. Equitability takes into account the close proximity of couple points that should be in a small range and did not scatter. Eliminating maximization and normalization can result in an unnatural grid arrangement and a more accurately captured relationship on grids with more cells than on simple grids, respectively. Consequently, the outstanding MIC’s heuristic properties are advantageous for estimating the dependency of Thai rice data.

According to MIC utilization, the MIC provides significant benefits for revealing associations between variables: the outlier’s robustness, the absence of variable distribution assumptions, and ease of interpretation. It is extensively used in a variety of research domains, including language recognition, image processing, search-based data analysis, spatial data analysis, and medicine. Wang et al. [32] enhanced the MIC methodology by developing iMIC, which was intelligent in its search for the relationship between variables in the system. Additionally, the agriculture domain [33] chose the MIC method to determine the factors influencing rice price variation.

### 3 Methodological approach

A Bayesian Maximal Information Coefficient (BMIC) is proposed to uncover the causal correlations from large data sets in the random system by integrating PGM and MIC.
Fig. 3  The scatter plots of $R^2$ scores of the MIC behavior and three variants of definition 6 to 8. (original picture from [31])

Springer
Bayesian Maximal Information Coefficient (BMIC) to reason novel trends in large datasets

with Bayesian Linear Regression (BLR). The proposed BMIC determines the best causality model to minimize run-time execution in order to identify the best causality model. In Fig. 4, the methodological approach depicts the processes of this work.

3.1 Search the correlated variables

The purpose of MIC is to reduce the number of variables in the search space through generality and equitability, which are advantageous for enabling the accessibility of numerous functions and noise data manipulation. Since the MIC algorithm assigns a score for ranking their correlation coefficient, a high MIC score indicates a strong dependent relationship between pairwise variables for the purpose of constructing the structural model. The execution time required to find an optimized model during the structure learning process is reduced. Unfortunately, the relationship $x \rightarrow y$ has a MIC score equal to $y \rightarrow x$, which cannot represent causality. A directional arrow between any two nodes is additionally required when using MIC. Consequently, MIC is interesting in studying the orientation of pairwise variables for identifying the cause-and-effect relationship as a result of an undirected model of the variables’ dependency, the symmetrical relationship.
3.2 Identify the causal relationship

PGM is a robust framework that enables the identification of causal relationship models. The BN is a representation of the PGM framework that builds a DAG by encapsulating knowledge about probabilistic numerical information. The conditional probability distribution, which takes the posterior probability value into account, denotes the variables’ dependence. Moreover, posterior probability serves as the foundation for any inference that involves the integration of prior knowledge and new information. They are used to represent the edges in a network. By favoring the Bayesian perspective, Bayes’ theorem captures the posterior probability. It is a powerful statistical theory that can determine the probability of uncertain data. Thus, Bayesian inference is a well-established technique for establishing the causal relationship between random variables.

3.3 Determine the posterior probability

The Bayesian view is used to derive the posterior probability values for the linear regression model. BLR [34] estimates the posterior probability of a linear regression model using the Bayesian theorem. Instead of using point estimation, the linear function is expressed as a probability distribution. The output parameter, \( y \), represents the normal distribution’s response in terms of mean and variance.

\[
y \sim N(\beta^T x, \sigma^2)
\]

The mean value is obtained by multiplying the parameter \( \beta^T \) by the input variable \( x \), whereas the variance value is a power of standard deviation. Once the \( y \) variable of the linear regression model is completely computed, the posterior distribution for the model parameter given the \( x \) and \( y \) variables can be determined using Bayes’ theorem support. The model parameters were chosen in accordance with Bayesian inference principles.

\[
P(\beta \mid y, x) = \frac{P(y \mid \beta, x)P(\beta \mid x)}{P(y \mid x)}
\]

where, \( P(\beta \mid y, x) \) is a posterior probability distribution, \( P(y \mid \beta, x) \) is a likelihood of data, \( P(\beta \mid x) \) is a prior probability, and \( P(y \mid x) \) is a normalization.

Due to the intractable nature of continuous values for parameterizing posterior distribution evaluation, the BLR model samples the posterior distribution using the Markov Chain Monte Carlo (MCMC) algorithm to estimate the posterior distribution. The Markov Chain is a mathematical representation of the next sample value generated by the previous sample value. And Monte Carlo is a term that refers to the technique of selecting a random sample. So, The MCMC concept emphasizes that the more posterior distribution samples are drawn, the more convergent the posterior distribution between approximated and real values will be.

PyMC3 [35] is a python library used to develop the Bayesian model by the MCMC method. Bayesian inference has been used to construct Bayesian Linear Models from the linear function that was formulated. It is provided by the generalized linear models (GLM) module, which can generate the formula from the input variable(s), \( x \), and a target output variable, \( y \).

The BLR does not impose any restrictions on the use of non-informative priors when assigning values to model parameters. The non-informative prior is considered by a normal distribution with the mean and variance of observed data. Another advantage is that the posterior probability distribution can be used to quantify the model’s uncertainty. The degree of certainty relates to the quantity of data; a model with a low level of uncertainty results from increasing the quantity of data. As a result, the BLR has gained popularity in statistical training and is widely used in a variety of domains. It is most frequently used to investigate the relationship between model variables. As with [36], the BLR outperforms frequentist statistical methods when examining the relationship between an electroencephalogram (EEG) and anxiety using clinical data.

The BLR has been proposed for predictive modeling, where it is used to represent a predicted value and a confidence (probability) interval (CI) [37]. The predicted value denotes the distribution of the target \( y \) given a set of features \( x \).

\[
P(y_j \mid x_j)
\]

Additionally, relying solely on the posterior probability distribution for Bayesian decision-making may be insufficient. The confidence interval (CI) assists in validating the model parameters’ possible values within a range. The greater the percentage of CI, the stronger the relationship should be between pairwise variables. The geology study in [38] is based on a demand prediction model for seismic fragility that estimates the fragility curve’s uncertainty. The BLR yields more precise results than alternative benchmarks.

In frequentist statistics, the CI [39] is a measure of uncertainty that can also be used to denote a degree of uncertainty. It consists of the lower and upper bounds for estimating the result, which can be varied according to the sample size and standard deviation of observed data (heterogeneity). The width of an interval determines the precision of the estimated values. A low degree of uncertainty is indicated when the data sample is large and the CI width is narrow. However, heterogeneity is directly proportional to the degree of uncertainty, as low heterogeneity indicates low uncertainty resulting from the narrow confidence interval. Thus, the narrower an interval’s width, the more precise the effect estimates.
The CI is preferred over the p-value, particularly in the health sciences, due to the potential for misinterpretation, misuse, and overestimation of hypothesis testing. Typically, the p-value establishes a threshold for accepting or rejecting the defined null hypothesis at a significance level of 0.05 (5 percent). Numerous researchers have interpreted the p-value as the probability of accepting or rejecting the null hypothesis, denoted by \( P(H_0 \mid y) \), where \( y \) is the probability of \( H_0 \) given the observed data. Indeed, the p-value quantifies the probability of the actual result given the null hypothesis as \( P(y \mid H_0) \). Furthermore, researchers frequently attempt to oversimplify the interpretation of p-values in practice. They distinguish between statistically significant and non-significant data using a p-value threshold of 0.05. Certain concerns, such as sample size and result estimation variability, have been overlooked. As a result, the CI is promoted as an alternative to the p-value, which can describe the variability of the estimate and the width of the interval. The width of the interval indicates the precision of the result, which is typically set at a 95 percent confidence level.

Furthermore, the Bayesian credible interval, abbreviated as CrI, is used to estimate the posterior probability distribution's uncertainty in the Bayesian approach. The CrI operates in the same manner as the CI. Bayesian inference uses two types of CrIs to evaluate the posterior distribution: an equal tail CrI and the highest posterior density (HPD) interval. The equal tail CrI is a direct posterior probability threshold value associated with an interval of the posterior probability distribution, which is easily calculable. For instance, the upper bound of CrI is 0.975, which means that it encompasses 97.5 percent of the posterior probability quantile distribution. Asymmetric posterior probability, on the other hand, has a lower probability of affecting the yield estimate value within the interval than outside the interval. In addition, the HPD interval defines the posterior probability threshold, indicating an interval with the probability mass surrounding the distribution center. Considering the data, a 95 percent confidence interval for the HPD interval of \(-4.0\) to \(-1.0\) implies that the mean difference primarily emphasizes between \(-4.0\) and \(-1.0\) given the observed data. The maximum posterior probability value would be between \(-4.0\) and \(-1.0\). Although the HPD CrI is equivalent to the equal tail CrI due to posterior probability symmetry, the HPD CrI is more complicated in a computation interval when compared to the equivalent tail CrI method.

**4 Casestudy on Thai rice price**

As a case study, the proposed BMIC is used to identify the causal correlations from a large data set on Thai rice prices by identifying the change in the paddy price of Jasmine rice. The causality of changes in the price of Jasmine rice can be accurately determined by its uncertain data and associated impacts.

**4.1 Thai rice background**

Due to the fact that Thai rice is a well-known global trading commodity, it is an interesting subject to study the effects of price changes. Despite the high price of Thai rice, many countries are willing to import it. Thailand is the world’s largest exporter of rice [40]. Thai rice is a critical agricultural product for the country’s revenue expansion. Rice is not a monopoly product, as many countries, including China, India, Thailand, Vietnam, Pakistan, Australia, and the United States, have the ability to export it [41]. Resultingly, other competitors benefit from the opportunity to adjust their rice prices in accordance with the FOB of Thai rice price criteria. FOB [40, 42] stands for Free on Board, which is typically associated with the process of product shipment. According to Incoterms® (International Commercial Terms), a trading contract specifies on a FOB basis that the exporter or consignee responsible for the costs and risks associated with transporting the cargo to the export port of origin, which are included in the export price of the particular product. While the importer or consignee will bear the costs and risks associated with the delivery to the destination. Thus, the export price quoted FOB excludes insurance, loading and unloading, freight, customs, value-added tax, import duty, and transportation costs (from the exporter’s port to the destination).

Unfortunately, Thai rice encounters difficulties, causing the price of Thai rice to fluctuate due to global and domestic circumstances. Many agriculturists decide to plant alternative crops in order to increase their profit margins. The 2008 food crisis [43] was a serious event that resulted in a 50–100 percent increase in raw material and food costs. The increase in demand, the decline in stocks and farming area, the lack of agricultural infrastructure and investment, the price of oil, and the exchange rate all have an effect on the rising price, affecting worldwide demand and supply [44].

Furthermore, Thailand encountered domestic difficulties in 2011 as a result of the Paddy Price Pledging scheme. The Paddy Price Pledging scheme is the political mechanism that sustains the income of farmers. The government permitted promising an unrestricted paddy quota and guaranteeing an acquisition price that was 50 percent higher than the market price. This policy resulted in a gradual rise in Thai rice stockpiles while the real market continued to decline. It resulted in a dramatic increase in the price of Thai rice relative to competitors, which caused the loss of access to trading partners. A linear model was created in the study of [45] to determine the impact of Thailand’s rice-pledging
policy. The global trading price of rice act as an indirect factor in determining the price of various varieties of Thai rice inside a country. Because rice exporters are the first to know the price direction from the global market, they attempt to bargain with millers who set the rice price. Accordingly, domestic rice prices are forced to set a similar price to the export price.

During the same period as the Paddy Price Pledging policy was implemented, Thailand faced an emergency disaster, the severe flooding of 2011, which shifted Thailand’s position as the world’s largest rice exporter to be India as a result of the rising Thai rice price. The high price enticed farmers to increase their rice production. Consequently both the actual market and the stockpiles were brimming with Thai rice. Regrettably, Thai rice revenue declined, as partners benefited from a discount, knowing that Thai rice needed to be sold. Thus, Thailand was under pressure to sell rice in 2014. Then, the Thai rice price fell to 385.91 US dollars per ton in 2015 [46], and all activities ceased completely in 2016. The Thai rice price was then restored to its regular level in comparison to the Vietnamese rice price [47]. Therefore, the variance in Thai rice prices is determined not only by time, but also by impact factors [48]. The purpose of this study is to identify the impact factors through the use of a probabilistic model, thereby ascertaining the causes of variance in the price of Thai rice. In identifying fluctuations in rice price, significant situations such as the quantity of Thai rice for export, the Thai rice price, the exchange rate, and the rice prices of other countries should be addressed [49]. These findings aid all stakeholders in forecasting the Thai rice price.

4.2 Data sample

We implicitly used BN to illustrate the causation of volatility in the price of Thai rice as a result of numerous external influences. In a BN, each directed edge represents the causal relationship between parent and child nodes. Additionally, BN can be utilized as a decision-making tool for foreseeable situations.

BN requires pertinent data for the extraction of Thai rice pricing models. Thai paddy prices, particularly those of the Thai Jasmine (Hom Mali) species, are being studied to determine the cause of price fluctuations. There are two cultivation seasons, each of which produces a different crop. 85 percent of the rice is known as the major rice, which yields jasmine rice (Hom Mali rice) and parboiled rice. The major rice is usually grown between May and October and is harvested from August to April the following year. Another 15 percent of rice is classified as second rice. The second rice produces glutinous rice, non-glutinous rice, and indigenous species which is harvested between January and April [41]. As such, this paper examines the paddy price of jasmine rice using characteristics associated with major rice.

The sample for this study spans 11 years, from 2008 to 2018, and includes all the important variables collected from the sources of the Office of Agricultural Economics [50], the Thai Rice Exporter Association [51], Thai Customs [52], and the Thai Meteorological Department [53]. Due to the increasing interest in the price of Thai rice in economic and data science domains, the variable selection was based on criteria studied in the following other publications. Economic studies have concentrated on the links between rice prices or the competitive nature of the rice market. Using case studies from Thailand and Vietnam, the usefulness of several models [1, 2] demonstrated the asymmetric volatility of rice prices. Between 2003 and 2013, they [44] considered the factors influencing the Thai rice price. Numerous studies have been conducted in the field of data science to examine various statistical methodologies [3, 4, 49, 54]. The ARIMA model has been a widely used tool for presenting a prediction result focused on the quantity of Thai rice exported, the price of Thai rice, the exchange rate, and the price of rice in other countries [49].

For the reasons stated previously, 14 variables (including paddy price) were gathered with continuous values in order to avoid missing values for variables of interest. They are the plant area of major rice, the yield of major rice, the minimum income of Thai citizens, the paddy price of Jasmine rice, the price of Jasmine rice, the export quantity of Jasmine rice, the export price of Jasmine rice, the export price of five percent white rice in Thailand, the export price of five percent white rice in Vietnam, the gold price, the domestic oil price, US exchange rate and Thailand’s agricultural GDP. The variables in Table 2 are abbreviated. After determining that paddy prices may fluctuate due to domestic and global factors, we will collect the significant relevant qualities. Domestic elements dictate the relevant characteristics of Thailand’s economic system, whilst global factors highlight competitors’ rice prices in the same category.

4.3 Exploratory Analysis with MIC

In this study, we employ a probability model, BN, to deduce the cause-and-effect relationship between the factors that affect Thai paddy price variations. The observed data, which includes both target and predictor variables, is necessary for the learning structure process to produce the perfect BN. Without prior knowledge, the structure model is discovered by calculating the probability of variable dependency. The highest posterior probability distribution given its data is returned to confirm their relationships. However, learning
Bayesian Maximal Information Coefficient (BMIC) to reason novel trends in large datasets

A large number of random variables is required to develop the model, which might result in a super-exponential expansion of DAGs in the search space, referred as an NP-hard problem.

The structure learning algorithms consist of constraint-based and score-based algorithms that are applied to observed data in order to detect relationships between variables. The optimized graph is determined by its possession of the maximum posterior probability distribution. Due to the NP-hard problem, we discovered that the run time required significant time to determine the optimal model. We begin by narrowing the variable space using the MIC algorithm to select only the highest strength relationship to the paddy price of Jasmine rice. As a result, the first ten-ordered pairwise variables in Table 3 are preferred according to the perfect and strong relationship in Table 1 because of their high impact on the target variable.

Unfortunately, MIC results cannot explicitly infer a cause-and-effect relationship by reason of their symmetric nature. The MIC results diagram can be used as a preliminary result to connect pertinent variables via undirected edges in Fig. 5.

According to Table 3, the MIC score demonstrates a perfect association between target variable of the paddy price of Jasmine rice with the Jasmine rice price, and the Jasmine rice export price. This implies that the paddy price of Jasmine rice is directly related to both aforementioned variables. The remainder of the MIC scores can be used to establish an indirect relationship between the paddy price of Jasmine rice. However, the perfect MIC score is also found in the relationship between plant area and yield of major rice, despite being unrelated to the paddy price of jasmine rice. While the export price of 5 percent white rice in Thailand, Vietnam, and Pakistan has a high MIC score, it is clearly irrelevant to the paddy price of Jasmine rice. These preliminary MIC findings conclusively established the associations to define the scope of a variable and focus on its true relationship to the target domain.

Due to the inability of the MIC score to discern cause-and-effect relationships, all edges in the diagram must be assigned orientations. The BN representation has been chosen to illustrate the volatility of the paddy price of Jasmine rice in a graphical model. Correlated data is a parameterization of the Bayesian network that is used for parameter and structure learning [55]. Zhang et al. [56] used MIC to improve the heuristic search for structure learning. In addition, they discovered the triangular loops relationship problem, which must eliminate the loop by using the d-separation rule. Then, the conditional independence test confirmed their work’s assignment orientation. As such, our study is an attempt to develop and provide a new technique for the MIC algorithm that prioritizes assigning directional relationships to support the cause-and-effect relationships.

Table 2 The abbreviations of variables

| Variables of interest                          | Variable names | Meaning                                                                 |
|-----------------------------------------------|----------------|-------------------------------------------------------------------------|
| Plant area of major rice                      | mjPlantArea    | The plant area of major rice (in-season rice) (Thai unit of area = 1,600 square meters) |
| Yield of major rice                           | mjYield        | The yield of major rice (in-season rice) (tons)                         |
| Minimum income of Thai citizen                | minIncome      | The minimum rate income of Thai workers (US-Dollar)                     |
| Paddy price of Jasmine rice                   | paddyPrice     | The paddy price of jasmine rice (US-Dollar)                            |
| Jasmine rice price                            | jasminePrice   | The domestic jasmine rice price (US-Dollar)                            |
| Export quantity of Jasmine rice               | jasmineEq      | The export quantity of jasmine rice (tons)                             |
| Export price of Jasmine rice                  | jasmineEp      | The export price per ton of jasmine rice (US-Dollar)                   |
| Export price of 5 percent white rice in Thailand | whiteThEp     | The export price per ton of 5 percent white rice in Thailand (US-Dollar) |
| Export price of 5 percent white rice in Vietnam | whiteVnEp     | The export price per ton of 5 percent white rice in Vietnam (US-Dollar) |
| Export price of 5 percent white rice in Pakistan | whitePkEp   | The export price per ton of 5 percent white rice in Pakistan (US-Dollar) |
| Gold Price                                    | goldPrice      | The gold price (US-Dollar)                                             |
| Domestic oil price                            | oilPrice       | The oil price (US-Dollar)                                              |
| US exchange rate with Thai currency           | USExchange     | The exchange rate of one US Dollar to Thai Baht (Thai Baht)             |
| Thailand’s GDP of agricultural product         | GDP            | The GDP of agricultural product of Thailand (US-Dollar)                 |

Table 3 MIC score estimate of pairwise variables in the sample space

| x variable       | y variable       | MIC score |
|------------------|------------------|-----------|
| mjYield          | mjPlantArea      | 1         |
| jasminePrice     | paddyPrice       | 1         |
| jasminePrice     | jasmineEp        | 1         |
| whiteThEp        | goldPrice        | 0.8664    |
| whiteThEp        | whitePkEp        | 0.8469    |
| USExchange        | jasmineEp        | 0.8442    |
| USExchange        | oilPrice         | 0.8345    |
| USExchange        | jasminePrice     | 0.8169    |
| whiteVnEp        | whitePkEp        | 0.8041    |
5 Results and discussion

A causality model for understanding the direct and indirect impacts on a target variable, the paddy price of Jasmine rice, is developed using the results of a proposed BMIC and traditional structure learning of PGM. They are investigated using a learning method based on Bayesian inference in order to discover the least complex causal model. Moreover, the proposed BMIC aims to solve the NP-hard problem which can identify the optimal structure from a super-exponential number of graphs in the search space.

5.1 The proposed BMIC (method of PGM and MIC with BLR)

The MIC diagram in Fig. 5 uses the BN construction rule to determine the orientation of the edges. The examination of the causation of the fluctuation in Jasmine rice paddy prices uses continuous variables because both the target and the predictor factor contain the time series data. We attempt to deduce the cause-and-effect relationship between nodes from the MIC result. By contrast, MIC returns the symmetry results, which indicate the relevant association between pairwise variables in both directions.

We conduct the analysis using the open-source Python package PyMC3, which includes models and probabilistic machine learning algorithms based on gradient-based MCMC algorithms. The BLR provides a more robust foundation for Bayesian inference and is typically used to summarize the posterior distribution by a central tendency (mean) and an uncertainty estimate (variance). A linear regression model generates the response, \( y \), from the normal distributions of the predictor variables discussed previously. The posterior distribution is used to derive the conditional probability of the Bayes’ theorem from the linear regression parameter. The linear model formula will be generated from the MIC results, assuming that \( x \sim y \) and \( y \sim x \).

Following that, parameter estimation is carried out using MCMC responses from model simulation. The intercept and predictor variables of the formula are returned in a normal distribution (details of mean and variance). Besides, a measure of the effect estimate’s uncertainty requires the CI or Crl. The most probable parameter values are determined in this experiment using the HPD interval of 3 percent and 97 percent. The width of the interval denotes the uncertainty degree of the estimated model. The most precise linear model reveals the HPD interval with the smallest width.

The posterior probability from the linear model is computed using all ten relationships. The MCMC algorithm is in charge of simulating the Bayesian model, which is parameterized using the identified linear formula. Each parameter’s observed data must conform to the normal distribution. Thus, the direction of each pairwise relationship has been determined by selecting the smallest HPD Crl given the results in Table 4.

The BN in Fig. 6 illustrates the direct and indirect effects on the paddy price of Jasmine rice as a result of the HPD Crls in Table 4. We discovered four connections, which we classified into two categories of factors irrelevant to the target. The first group comprises plant area and yield of major rice, which exhibit a perfect relationship. The second group includes the gold price and the export price of Thailand, Vietnam, and Pakistan’s 5 percent white rice. On the other hand, we discovered that the paddy price of jasmine rice has been clearly affected by the jasmine rice export price. Domestic jasmine rice prices are influenced by the paddy price and export price of jasmine rice. Simultaneously, the US dollar’s exchange rate against the Thai baht and domestic oil price have an indirect effect on the paddy price of Jasmine rice.

5.2 The traditional structure learning experiment

The model is constructed with BN, which can be obtained using traditional or applied approaches. The observed data
The traditional structure learning experiment is validated using both constraint-based and score-based learning. The structure learning process takes all variables into account.

To begin, the constraint-based method depicts a BN, as illustrated in Fig. 7. Correlation analysis is used to determine conditional independence, which use the Chi-square dependency test to estimate the model skeleton. After that, each connection between nodes should be assigned an orientation based on the information contained in the separate sets: a set that contains the conditional independence of each pair of indirectly connected nodes. The skeleton model and separate set information can be used to estimate the partial DAG (PDAG), in which relationships may exist in both directions, \( x \to y \), and \( y \to x \). Due to the fact that the fundamental of BN prevents the occurrence of a v-structure or cyclic graph, BN is intuitively constructed as a DAG from PDAG. It appears the BN analysis of the selected case demonstrates that the export and domestic prices of Jasmine rice exert control over the rice paddy price. In detail, the domestic price of Jasmine rice is influenced by the export price of Jasmine rice, which in turn influences the paddy price of Jasmine rice. As can be seen, no other effect is associated with the export price of Jasmine rice. Therefore, we can assert that the export price of Jasmine rice is an initiatory actor in the variation in the paddy price of Jasmine rice.

Second, Fig. 8 illustrates the structure of the score-based algorithm. The Bayesian estimator will return the optimal score by estimating the best model for each learning process. BN takes into account the Bayesian Information Criterion (BIC), which is a log-likelihood score with Dirichlet priors for expressing how well observed data describes a model. To prevent overfitting, the method incorporates a penalty
for network complexity. It turned out that the paddy price of Jasmine rice can be changed by the export price of 5 percent white rice in Pakistan. The dependent relationship is a chain that begins with the export price of 5 percent white rice in Pakistan, continues with the domestic oil price, the US exchange rate against Thai currency, the domestic price of Jasmine rice and finally with the export price of Jasmine rice.

The three models, Figs. 6 to 8, show the various impact factors that can cause the paddy price of Jasmine rice to change based on the results of the proposed BMIC model and traditional PGM model. It is noticeable that some pairwise nodes in the different models have similar connections but distinct directions. The proposed BMIC can reduce the number of relevant variables by perfect and high-strength relationships selection. Choosing only relevant variables for BN structure learning has the advantage of avoiding the NP-hard problem. The result is completely satisfactory as it is consistent with the findings of another study. The traditional method will be then compared to the proposed BMIC in terms of performance and model accuracy. Ten connections are contained in the best-selected model in the case of constraint-based learning. The graph quite closely resembles the undirected edges of MIC results. Only one additional relationship has been added: the GDP of agricultural products and the minimum income of Thai citizens. After assigning
Bayesian Maximal Information Coefficient (BMIC) to reason novel trends in large datasets

Table 5 The BIC score for measuring BN complexity of different learning models

| Learning Models                          | BIC score |
|-----------------------------------------|-----------|
| The proposed BMIC (method of PGM and MIC with BLR) | -1072.59  |
| The constraint-based learning algorithm  | -1333.3   |
| The score-based learning algorithm      | -1324.67  |

the orientation, the opposite effects of Jasmine rice paddy price change are implied by the different directions of cause-and-effect relationships. Furthermore, the constructed BN contains 13 relationships as a result of score-based learning. The dependency relationships are more complicated than the proposed BMIC and constraint-based learning approaches proposed previously. The findings of score-based learning indicate that changes in the paddy price of Jasmine rice can occur via the forwarding of numerous effects.

As a result, the explicit model of each algorithm reveals a distinct BN pattern. The optimal BN model should be the one that best reflects the cause-and-effect relationships that exist when the paddy price of Jasmine rice changes. The BIC score is used to assess how well-described the constructed BN is in light of its data. The BIC score is a log-likelihood calculation with an added penalty to avoid overfitting, which results in a negative value. The best model will return the probability with the highest value. The BIC scores in Table 5 demonstrate that the proposed BMIC algorithm outperforms the other two learning algorithms in traditional PGM. Therefore, the proposed BMIC obtains the best-fitting BN in order to learn the structure of Jasmine rice paddy prices.

6 Conclusion

The ability of the BN framework to interpret the model is exploited in this paper. The BN framework is used to describe the causal model underlying the Thai rice price situation, with an emphasis on changes in the paddy price of Jasmine rice. Given the existence of data uncertainty in the price of Thai rice and other relevant attributes, the use of BN probability inference provides a more complete explanation. By structural learning from the observed dataset, BN establishes a model that is fundamentally based on the Bayes’ theorem. The NP-hard problem is frequently encountered during the structure learning process due to the super-exponential number of graphs in search space. We limited the number of variables in the Thai rice price system by using the MIC algorithm in order to capture only the strongest relationships. However, MIC results cannot be used to explain the causality of any target domain because of the undirected nature of MIC results. Therefore, the BLR model determines the edge orientation, with an emphasis on the posterior probability distribution’s return. The BLR model assumes the normal (Gaussian) distribution sample to formulate the linear regression. Under the Bayes’ theorem, the BLR model parameters are given by the inputs and outputs, of the linear model. The Bayes’ theorem is applicable to models that are probabilistic and non-deterministic. Because BN demonstrates the independent relationships of a given joint probability distribution, each connected node presents a conditional probability that results the posterior probability. The posterior distributions for the dependent variables \( x \rightarrow y \) and \( y \rightarrow x \) are computed in both directions from the undirected MIC results, with the highest posterior probability distribution being preferred.

The Crl is used to evaluate the significant posterior probability distribution, with the HPD interval as a criterion. The orientation prefers the direction of dependent variables holding a minimum value of the HPD interval. The proposed BMIC elucidates the causes and effects of the paddy price of Jasmine rice via a BN model based on the estimated edge direction results. We discovered that the export price of Jasmine rice can affect its paddy price. This point is confirmed by [45], which discusses exporters’ ability to forecast the direction of global rice market. Moreover, the model indicates that the paddy price of Jasmine rice can affect the domestic price of Jasmine rice. This information becomes useful for developing policies and coping with future circumstances, as we can track trends in the paddy price of Jasmine rice and its directed effects. The proposed BMIC’s reliability is assessed by comparing network scores to those of a traditional PGMs. The network score indicates how well explained and fitted the model is. The score demonstrates that the proposed BMIC achieves the highest score, while the two conventional learning algorithms achieve lower scores. As a result, this study enhances the optimized model selection process by synergy a new BN structure from MIC results with the BLR model. As a consequence, the BN model establishes the cause-and-effect relationships for the paddy price of Jasmine rice, allowing us to continue developing a predictive model for our future work.

Acknowledgements The authors thank the government of Thailand and the office of Agricultural Economics for providing the historical data of Thai rice.

Author Contributions All authors contributed to the study conception and design a novelty on its application. Conceptualization: Shuliang Wang; Methodology: Shuliang Wang and Tisinee Surapunt; Material preparation, data collection and analysis were performed by Shuliang Wang and Tisinee Surapunt. Writing - original draft preparation: Tisinee Surapunt; Writing - review and editing: Shuliang Wang and...
Conflicts of interest

Participants included in the study.

Ethical approval

This manuscript does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent

Informed consent was obtained from all individual participants included in the study.

The authors declare that they have no conflict of interest.

References

1. Baharom AH, Radam A, Habibullah MS, Hirnissa MT (2009) The volatility of thai rice price, Munich Personal RePEc Archive, 1-10
2. Ghoshray DA (2008) Asymmetric Adjustment of Rice Export Prices: The Case of Thailand and Vietnam. International Journal of Applied Economics 5:80–91
3. Sujjiviriyasup T (2018) Predicting prices of agricultural commodities in Thailand using combined approach emphasizing on data pre-processing technique. Songklanakarin Journal of Science and Technology 40:75–78
4. Shao YE, Dai J-T (2018) Integrated Feature Selection of ARIMA models with Computational Intelligence Approaches for Food Crop Price Prediction. Complexity 2018:1–17
5. Pandey S, Sulser T, Rosegrant MW, Bhandari H (2010) Rice Price Crisis: Causes. Impacts, and Solutions, Asian Journal of Agriculture and Development 7:1–15
6. Marloes M, Mathias D, Steffen L, Martin W (2013) Handbook of Graphical Models. Springer, New York
7. Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. The MIT Press, Cambridge
8. Helong Y, Guifen C, Dayou L (2009) A Simplified Bayesian Network Model Applied in Crop or Animal Disease Diagnosis. Computer and Computing Technologies in Agriculture III(294):1001–1009
9. Liu K-R, Kuo J-Y, Yeh K, Chen C-W, Liang H-H, Sun Y-H (2015) Using fuzzy logic to generate conditional probabilities in Bayesian belief networks: a case study of ecological assessment. International journal of environmental science and technology 12:871–884
10. Aguilera PA, Fernández A, Fernández R, Rumi R, Salmerón A (2011) Bayesian networks in environmental modelling. Environmental Modelling & Software 26:1376–1388
11. Chawla V, Naik H, Akintayo AD, Schnable HP, Ganapathysubramanian B, Sarkar S (2016) A Bayesian Network approach to County-Level Corn Yield Prediction using historical data and expert knowledge, The 22nd ACM SIGKDD Workshop on Data Science for Food, Energy and Water, 1–8
12. Nuvaisiyah P, Nitha F, Saepudin D (2018) Price Prediction of Chili Commodities in Bandung Regency Using Bayesian Network. International Journal on Information and Communication Technology (IoICT) 4:19–32
13. Alvi DA (2018) Application of Probabilistic Graphical Models in Forecasting Crude Oil Price, University College London
14. Chickering DM (1996) Learning Bayesian Networks is NP-Complete. Springer, New York
15. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVejan G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC (2011) Detecting Novel Associations in Large Data Sets. Science 334:1518–1524
16. Ruohonen K (2013) Graph Theory (Translation by Janne Tamminen, Kung-Chung Lee and Robert Piché)
17. Joyce J (2019) Bayes’ Theorem, Stanford Encyclopedia of Philosophy (Editor: Zalta EN), Metaphysics Research Lab, Stanford University, California
18. Halbersberg D, Wienreb M, Lerner B (2020) Joint maximization of accuracy and information for learning the structure of a Bayesian network classifier. Machine Learning 109(5):1039–1099
19. Jiao J, Wei M, Yuan Y, Zhao T (2020) Risk quantification and analysis of coupled factors based on the DEMATEL model and a Bayesian network. Applied Sciences 10(1):317
20. Bari MF (2011) Bayesian Network Structure Learning. 4th Annual Meeting Asian Assoc. Algorithms Comput (AAAC), 1–8
21. Burnham KP, Anderson DR (2004) Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research 33:261–304
22. Cooper GF, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9:309–347
23. Heckerman D (1995) Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning, 197-243
24. Beretta S, Castelli M, Gonçalves I, Henriques R, Ramazzotti D (2018) Learning the Structure of Bayesian Networks: A Quantitative Assessment of the Effect of Different Algorithmic Schemes. Complexity 2018:1–12
25. Liu Z, Malone B, Yuan C (2012) Empirical evaluation of scoring functions for Bayesian network model selection. BMC Bioinformatics 13:1–31
26. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVejan G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC (2011) Supporting Online Material for Detecting Novel Associations in Large Data Sets. Science 334:1518–1524
27. Cai Y, Luo X, Liu Z, Qin Y, Chang W, Sun Y (2019) Product and process fingerprint for nanosecond pulsed laser ablated superhydrophobic surface. Micromachines 10:1–15
28. Liang Y, Abbott D, Howard N, Lim K, Ward R, Elgendi M (2019) How Effective Is Pulse Arrival Time for Evaluating Blood Pressure? Challenges and Recommendations from a Study Using the MIMIC Database. Journal of Clinical Medicine 8:1–14
29. Székely GJ, Rizzo ML (2009) Brownian distance covariance. The Annals of Mathematics 334:1283–1293
30. Benesty J, Chen J, Huang Y, Cohen I (2009) Pearson Correlation Coefficient, Noise Reduction in Speech Processing 2:1–4
31. Reshef D, Reshef Y, Mitzenmacher M, Sabeti P (2013) Equitability Analysis of the Maximal Information Coefficient, with Comparisons, 1–22
32. Wang S, Zhao Y, Shu Y, Yuan H, Geng J, Wang S (2018) Fast search local extremum for maximal information coefficient (MIC). Journal of Computational and Applied Mathematics 327:372–387
33. Surapunt T, Liu C, Wang S (2017) MIC for Analyzing Attributes Associated with Thai Agricultural Products. International Conference on Geo-Spatial Knowledge and Intelligence 848:40–47
34. Clyde M, Cetinkaya-Rundel M, Rundel C, Banks D, Chai C, Huang L (2007) An Introduction to Bayesian Thinking. Springer, New York
35. Salvatier J, Wiecki TV, Fonnesbeck C (2016) Probabilistic programming in Python using PyMC3. PeerJ Computer Science 2:1–20
36. Baldwin SA, Larson MJ (2017) An introduction to using Bayesian linear regression with clinical data. Behaviour Research and Therapy 98:58–75
37. Kong D, Zhu J, Duan C, Lu L, Chen D (2020) Bayesian linear regression for surface roughness prediction. Mechanical Systems and Signal Processing 142:106770–106791
38. Ghosh S, Chakraborty S (2020) Seismic fragility analysis of structures based on Bayesian linear regression demand models. Probabilistic Engineering Mechanics 61:1–12
39. Hespanhol L, Vallio CS, Costa LM, Saragiotto BT (2019) Understanding and interpreting confidence and credible intervals around effect estimates. Brazilian Journal of Physical Therapy 23:290–301
40. Fahmy H (2019) Classifying and modeling nonlinearity in commodity prices using Incoterms. The Journal of International Trade and Economic Development 28:1019–1046
41. Office of Agricultural Economics (2017) The 2017 agricultural economy statistical information of each commodity. www.oae.go.th, Accessed 20 April 2019
42. Paisabazaar (2019) FOB : FOB Price (Free on Board): What is FOB Price?. https://www.paisabazaar.com/tax/fob-price/, Accessed 15 May 2020
43. Shah A, Global Food Crisis 2008 (2008) https://www.globalissues.org/article/758/global-food-crisis-2008, Accessed 9 March 2020
44. Kha PNH, Trinh VH (2017) Issues Surrounding the Rice Price of Thailand from 2003 to 2013, SSRN Journal, 1-32
45. Wasawong P (2018) The Impact of the Rice Pledging Policy: The Case of Thailand. A Linear Programming Approach, Southeast Asian Journal of Economics 6:81–113
46. Thai Rice Exporters Association (2019) Rice Exports Statistics. http://www.thairiceexporters.or.th/List_%20of_statistic.htm, Accessed 15 May 2019
47. Hoang HK, Meyers WH (2015) Price stabilization and impacts of trade liberalization in the Southeast Asian rice market. Food Policy 57:26–39
48. BBC News (2017) Trackback the Paddy Price Pledging: Effect the being export leader in global trade market. https://www.bbc.com/thai/thailand-41410157, Accessed 15 May 2019
49. Maneejuk P, Pastpipatkul P, Sriboonchitta S (2016) Analyzing the Effect of Time-Varying Factors for Thai Rice Export, Thai Journal of Mathematics, 201-213
50. The Office of Agricultural Economics. web resource. https://www.oae.go.th Accessed 23 Aug 2020
51. The Thai Rice Exporter Association. web resource. http://www.thairiceexporters.or.th Accessed 23 Aug 2020
52. Thai Customs. web resource. https://www.customs.go.th Accessed 26 Aug 2020
53. The Thai Meteorological Department. web resource. https://www.tmd.go.th Accessed 26 Aug 2020
54. Co HC, Boosarawongse R (2007) Forecasting Thailand’s rice export: Statistical techniques vs. artificial neural networks. Computers & Industrial Engineering 53:610–627
55. Bae H, Monti S, Montano M, Steinberg MH, Perls TT, Sebastiani P (2016) Learning Bayesian Networks from Correlated Data. Scientific Reports 6:1–14
56. Zhang Y, Zhang W, Xie Y (2013) Improved heuristic equivalent search algorithm based on Maximal Information Coefficient for Bayesian Network Structure Learning. Neurocomputing 117:186–195

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Wang Shuliang Professor of Beijing Institute of Technology, Executive Dean of the E-Government Research Institute, is one of National Leading Talents in scientific and technological innovation. He acts as the Vice Chairman of the Chinese Manufacturing Enterprise Double Innovation Development Alliance, Director of the Cognitive and Behavior Professional Committee of the Chinese Command and Control Society, and Chief Scientist of the National Key Research and Development Program. He mainly studies data mining, discovering the mathematical nature of spatial data mining. His monograph “Spatial Data Mining Theory and Application” ranks the first in Springer Nature’s computer-specific monographs, and was praised as “a landmark masterpiece of spatial data mining” by “Science Bulletin”. And his papers were published in Cell Patterns etc. Moreover, his authorized national invention patents are supporting the big data intelligence, and the promulgation of his national standard is standardizing the legal aid. For his achievements, he won the first prize of National Science and Technology Progress Award, National Excellent Doctoral Dissertation, CICC Young Scientist Award, IGI Technology Innovation Award, China Excellent Publication Award Book Award, and so on.

Tisinee Surapunt received the B.S. and M.S. degrees in Software Engineering from Chiang Mai University, Thailand, in 2010, 2015 respectively. She is a Ph.D. student at Beijing Institute of Technology, China. Her research interests include Big data, Data Analytics, Machine Learning and Code optimization.