Bayesian Maximal Information Coefficient (BMIC) to Reason Novel Trends in Large Datasets

Shuliang Wang

Beijing Institute of Technology School of Materials Science and Engineering

Tisinee Surapunt (tsurapunt@gmail.com)

Beijing Institute of Technology School of Computer Science and Technology https://orcid.org/0000-0003-3132-5636

Research Article

Keywords: Bayesian maximal information coefficient (BMIC), Causal correlations, Temporal sequential correlation, Bayesian linear regression, Maximal information coefficient, Thai rice price

Posted Date: April 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-343006/v3

Noname manuscript No.
(will be inserted by the editor)

Bayesian Maximal Information Coefficient (BMIC) to Reason Novel Trends in Large Datasets

Shuliang Wang · Tisinee Surapunt

Received: date / Accepted: date

Abstract Bayesian network (BN) is a probability inference model to describe the explicit relationship of cause and effect, which may examine 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 dataset 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 to assign orientation in a graph resulting by 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, 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 method obtains the highest score compared to the two traditional learning algorithms. Finally, the BMIC is applied to discover the causal correlations from the large dataset on Thai rice price by identifying causality change in the paddy price of Jasmine rice. The experimented results show the proposed BMIC returns the directional relationships with clue to identify the cause(s) and effect(s) on paddy price with 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 (2010) 62M10 · 62P20 · 91B84

Shuliang Wang
School of Computer Science and Technology, Beijing Institute of Technology
E-mail: slwang2011@bit.edu.cn

Tisinee Surapunt*
School of Computer Science and Technology, Beijing Institute of Technology
E-mail: 3820160053@bit.edu.cn
*corresponding author

1 Introduction

With data uncertainty, one important issue is found when data keeps growing in volume, variety, and velocity. The uncertainty resembles noise, contaminating the observed dataset, leading to deviation in correctness. However, the observed data is a valid part of the computational and statistical analysis necessary to obtain knowledge and utilize it further in a prediction. Price is regarded as a key element that holds a high possibility to drive the economy through domestic and international trading. An analysis of price is wildly in demand in economic (Baharom et al., 2009; Ghoshray, 2008) and agricultural (Sujjaviriyasup, 2018; Shao and Dai, 2018) topics, requested to know the mechanism, characteristics, and future trends. The price prediction model is a popular study, yet the data analysis process is quite tricky due to non-stationary and noise data. To achieve the prediction model, the relationship among data variables inform of causality becomes crucial for analyzing data. The price domain is affected by many uncontrollable attributes that gather in time-series data. Thus, the observed dataset simultaneously

presents data variability and genuine incomprehension of data attribute relationships in the system. In this paper, agricultural price is selected for study, particularly the price of Thai rice, due to price variation and global trading competition. The price of Thai rice can fluctuate by many relevant factors such as uncertain productivity, climatic conditions, competitors (import and export trading), and domestic and world situations.(Pandey et al., 2010) Understanding causative relationships among factors can reduce obstacles in pricing and avoid unexpected future situations. Therefore, the interpretation of causative relationships is better described via a graphical model.

Probabilistic Graphical Model (PGM) (Marloes et al., 2013) illustrates causal relationships by probability and graph. It (Koller and Friedman, 2009) is a robust framework to handle uncertainty predictions. The correlated and tuple uncertainty and their relationships are demonstrated by a joint probability distribution.(Singh et al., 2008) Bayesian network (BN) presents a directed acyclic graph (DAG) on a joint probability distribution using a conditional probability of variables from the Bayes' theorem along with machine learning. Much research has underlined the probabilistic forecasting of uncertain data with BN in a specific domain, including medical analysis(Helong et al., 2009), ecology system(Liu et al., 2015; Aguilera et al., 2011), agricultural forecasting(Chawla et al., 2016; Nuvaisiyah et al., 2018), and economic application(Alvi, 2018). Unfortunately, PGM's structured learning process encounters the NP-hard problem(Chickering, 1996) when all observed variables are required to construct the DAG. It implies that the super-exponential number of graphical models is in the search space. The more graphs in the search space, the more time, memory, and cost are required to discover the best network structure.

Consequently, this paper aims to use a data-dependent to scope the observed variables in the sample space. The dependence measure is calculated by maximal information coefficient (MIC).(Reshef et al., 2011) The MIC widely detects the relationships between pairwise variables in a large dataset. The coefficient will result in an R-square (R^2) of data related to the regression. An equitability property, which is one of the main properties of MIC, is reasonable for choosing this algorithm. The equitability will return the result of noiseless function relationships that maximize the R^2 to 1. Hence, MIC coefficient will roughly equal R^2 when the function is tested with different samples, noise models, and noise levels. Even so, the correlated data is going to be connected to their relationships as an undirected graph. Edges without a direction between data variables cannot represent the causality. Based on the reasons above, this paper highlights the synergy of correlated data of MIC with PGM, aiming to diagnose the direction of the relationship by applying a statistical inference. The Bayesian linear regression (BLR) formulates the linear regression with the Bayesian inference, implying the variables' dependency in a linear function. Therefore, the BLR in this paper is preferred to identify the directional relationship of correlated pairwise variables through their dependence by using a probability distribution. The correlated pairwise variables can manifest in the linear relationship between the target (y) and predictor (x) variables. The BLR declares the linear regression model using a probability distribution from a normal (Gaussian) distribution rather than a single value estimated.

Moreover, the target result of BLR draws a posterior probability for model parameters that can describe the conditional probability of correlated pairwise variables. The highest posterior probability determines the direction between correlated pairwise variables relationship. The pairwise variables, x_1 and x_2, will be tested as both target and predictor variables, meaning when x_1 is a target variable, x_2 will be a predictor variable and vice versa. Consequently, the alternation of being the target and predictor variables will obtain a different posterior probability. The highest posterior probability selection determines whether x_1 or x_2 will be the target or the predictor variable. The result can then be implied to be a directional relationship and conditional dependency between x_1 and x_2.

2 Related methods

PGM and MIC are the related methods for background knowledge.

2.1 Probabilistic Graphical Model

Probabilistic Graphical Model (PGM) (Marloes et al., 2013) framework conditionally establishes the network model into Bayesian network (BN) and Markov chain (MC). Both models highly support modeling a joint probability distribution into directed and undirected graphs, respectively. BN illustrates directed PGM, presenting the captured conditional dependencies of associated attributes through the Bayes' theorem, beneficially describing a causality relationship. BN is also applicable to making predictive modeling of uncertain data that explains dependent attributes using the occurrence probability of their relationships. On the other hand, MC is designed to be an undirected PGM that

shows only the nodes' association without any ordered causality.

The joint probability distribution theoretically supports the BN construction. The BN structure is conditionally connected the relationship of random variables. In terms of the PGM framework, the value estimation is represented by the probability distribution. Therefore, the conditional probability distribution (CPD) becomes an answer to help PGM interprets the joint distributed random variables with their intuitive models. The BN is used the Bayes' theorem (Ruohonen, 2013) inference in the presence of uncertain situations and handling many relationships among random variables (factors). The Bayes' theorem is usually made use in the financial domain to support the updating risk evaluation.

The key concept of Bayes' theorem is to simply define a result of the conditional distributions in terms of joint distributions. The Bayes' theorem requires the prior probability distributions in order to calculate the posterior probability. The prior probability is the probability of an event before collecting the new data. While the modification probability of occurring events after receiving the new data will return the posterior probability. The formula of Bayes' theorem is stated as

$$P(A \mid B) = P(A \cap B)/P(B) = P(B \mid A)P(A)/P(B) \tag{1}$$

As a result, the Bayes' theorem benefits the revising predictions that can update probabilities with the new information incorporation. In prediction terminology, the information of occurring events relies on the hypothesis and evidence, which imply to the probability of a hypothesis, H, on a given data, E. So, the formula is altered as

$$P(H \mid E) = P(E \mid H)P(H)/P(E) \tag{2}$$

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.(Ruohonen, 2013)

BN is a model representation based on graph and probability theories. A graph theory (Joyce, 2003) composes a set of vertices and edges, $G = (V, E)$. V indicates vertices or nodes, while E indicates edges in a graph. The edge displays the connection between associated vertices. The cardinality of vertices or nodes $|V|$ and edges $|E|$ in a graph refer to the order and the size of a graph. In BN, nodes represent random variables and edges present relationship on conditional independence. The basic BN initially composes of at least two nodes connected with one edge. BN's edge representation draws a directional edge that uses an arrow to give the best understanding of the causal relationship. Therefore, BN is a directed acyclic graph (DAG) avoiding a self-connection and cycle relationship that determines a fundamental cause-and-effect relationship concept. An example in Fig. 1 is an arrow that directly connects node A to node B, implying that node A is a cause of node B or that node B is affected by node A.

Fig. 1 A simple graph of a directed relationship

To consider the probability of a BN construction, S corresponds to the structured BN and θ_S corresponds to the associated conditional probabilities which is defined as $<S, \theta_S>$. While $P<S, \theta_S>$ is defined the joint probability distribution of all random variables in a constructed network. According to, the set of vertices V composes of a set of random variables $x_1, x_2, \ldots, x_n$ and edges E provide the conditional independence of two random variables. Therefore, each variable x_i in a set of V has the CPD as $P(x_i, Pa(x_i))$ where $Pa(x_i)$ is a set of parent variables of variable x_i. It can be denoted as

$$P\left\langle S, \theta_B \right\rangle = P\left(x_1, x_2, \ldots, x_n\right) = \prod_{i=1}^{n} P\left(x_i \mid \mathrm{Pa}\left(x_i\right)\right) \tag{3}$$

Consequently, BN draws edges among random variables to represent their dependencies by using structural learning of the observed data sample. There are two learning approaches in the PGM framework, parameter learning and structure learning, emphasizing dependent attributes beyond the conditional probability estimation process. Both learning algorithms aim to construct the DAG, but each model will be done by a different method and target data sample. The BN does not only learn through the pure data sample, but the existing relationships can also be learned to construct the network model.

2.1.1 Parameter learning

The parameter learning approach requires sample data with the existent DAG for capturing the conditional probability of individual variables. The conditional probability is estimated by selecting two components: Maximum likelihood estimation (MLE) and Bayesian estimation. MLE gives the maximized conditional probability of data given the model, $P(Data \mid Model)$. At the

same time, the Bayesian estimation returns the prior conditional distribution.

2.1.2 Structure learning

The structure learning approach captures the maximized probability of variable dependencies without knowing prior knowledge. The network model is obtained from the posterior probability distribution given its data sample, $P(Graph \mid Data)$. According to the Bayesian statistical decision theory, the number of graphs can grow in a super-exponential number of DAGs. Then, only the most optimized structure, which is measured by a network score, will be selected. Although structure learning encounters a limitation of unknown prior knowledge, the solution is to apply the uniform prior, which holds equal probability. It is generally implied that the $P(Graph \mid Data)$ is proportional to $P(Data \mid Graph)$. Hence, two learning algorithms, constraint-based learning and score-based learning, can help determine the network's arcs.

(a) Constraint-based learning
A constraint-based learning approach is structuralized intuitively from conditional independencies that do not overlook the concept of Bayesian Network. The conditional independence test is done by Pearson's χ^2-test, Fisher's Z-test, and t-test instead of testing the probability. This constraint-based learning approach considers three steps to model the network: conditional independence identification, skeleton learning of undirected relationship, and arcs direction learning. In (Alvi, 2018), the Markov blanket was first learned to optimize the number of candidate nodes of DAG. The conditioned edges of the Markov blanket can return the conditional independence of every set of nodes. It benefits skeleton learning by identifying the undirected edges in DAG. Then, the undirected edges will be assigned the direction, which is a complete partial DAG. The results imply the causal relationship between nodes.

(b) Score-based learning
The score-based learning approach emphasizes applying a network score to evaluate the best BN which fits the data. The maximal score returns the highest posterior probability of a graph (nodes and edges) given its observed dataset.

$$P(G \mid D) = P(D \mid G)P(G)/P(D) \tag{4}$$

Since the possible DAGs (BNs) have first seen growth in the search space in the super-exponentially of nodes, $O(n!2^{\binom{n}{2}})$(Bari, 2011) where n is number of nodes in the network , the heuristic search algorithms take responsibility to reduce the number of DAGs, finding the optimal BN structure. The scoring function works as an indicator that can calculate by AIC, BIC (Burnham and Anderson, 2004), K2 (Cooper and Herskovits, 1992) and Bdeu (Heckerman, 1995) for the heuristic search algorithm. The search algorithms also have selections for different purposes such as Hill-climb search, Tabu search, Genetic algorithm, and Greedy equivalent search (GES). In (Beretta et al., 2018), the focus was on the operation efficiency by comparing each scoring method of each heuristic search algorithm. The outcome found that variable types and interest domains can affect learning performance. Only a few heuristic search algorithms operate efficiently with scoring methods. However, the score-based mechanism can demonstrate the whole structural model's dependencies, avoiding the failure of individual conditional independence.(Koller and Friedman, 2009)

As a result, the dependencies among random variables can be described by PGM, which helps to depict relationships by supporting conditional probability. The hard work of identifying the optimized model encounters an NP-hard problem which loads all work to heuristic-search algorithms for comparing all networks' scores in the search space. The model which holds the highest score will be determined as the best model. Although the NP-hard problem in searching for the best network is time-consuming, PGM is still a robust framework that provides functions for illustrating the causal relationship model. Especially, BN can increase the users' comprehension by the intuition of model interpretation on a domain.(Liu et al., 2012)

2.2 Maximal information coefficient

Maximal information coefficient (MIC) (Reshef et al., 2011) statistically measures a dependency between pairwise variables on a large dataset and is described as an exploratory data analysis (EDA) tool. A measure of dependence is assigned by MIC score to determine the strength of the dependent relationship between pairwise variables. MIC score is ranged between 0 to 1 by R-squared (R^2) evaluation normalized from the mutual information. MIC score of 0 means an independent relationship. The upwards magnitude of MIC score, up to 1, depicts the increased strength of dependence which express as the strength type in Table 1. They are usually divided into 5-6 levels (Cai et al., 2019; Liang et al., 2019), including no correlation and perfect correlation levels.

MIC algorithm returns a significant score, when is compared with other algorithms, because of its two

Table 1 MIC score interpretation of correlation coefficient

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

heuristic properties: generality and equitability. The generality property supports unlimited access of relationship types. MIC can detect both linear and non-linear relationships while other algorithms, such as the distance correlation (Székely and Rizzo, 2009) and Pearson's R (Benesty et al., 2009), only perform on the linear function. In addition, the equitability property gives a similar MIC score to different relationship types with a similar noise level. MIC is found on the mutual information but is not the mutual information estimation. It is directly implemented equitable dependence measure which returns a more significant result than the mutual information and others in almost all noise models. The correlation coefficient in Fig. 2 demonstrates the score comparison between MIC and other current functions of noiseless models. It is found that MIC can detect the equal score as 1.00 on different functions of noiseless models except the random function.

Relationship functions	MIC	Pearson	Spearman	Mutual Information		CorGC	Maximal Correlation
				KDE	*Kraskov*	*Principal Curve-Based*	
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal *(Fourier frequency)*	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal *(non-Fourier frequency)*	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal *(varying frequency)*	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

Fig. 2 The correlation score of noiseless models of different relationship functions

In practice of MIC, the utility of maximization and normalization in the definition of MIC (Reshef et al., 2011, 2013) expresses preliminary for measuring a maximal coefficient of variables' dependence. It is found that the MIC suitably works across four basic noise models, an amount of a range of function types and sample size n between 250 to 1000.

Definition 1 (Maximal information coefficient (MIC))
Let $D|_G$ denotes the probability distribution which data D over cells of a grid G, and I denotes a mutual information. Then, let $I^*(D,x,y) = \max_G I\left(D|_G\right)$ where the maximal is from the partitioned x-by-y of a grid G (x and y can possibly be empty rows or columns). Therefore, MIC of pairwise variables on data D with sample size n can be defined as

$$MIC_{x,y}(D) = \max_{xy<B(n)} \{I^*(D,x,y)/\log_2 \min(x,y)\} \quad (5)$$

Where $B(n)$ is a growing function of sample size n on data D which satisfying equals $O(n)$. The default setting of $B(n)$ is $n^{0.6}$ which is from a heuristic suggestion by the founder team.

The equitability of MIC is outstanding the considering in a noise models and range of sampling to evaluate the dependence. The utility of maximization and normalization in the definition of MIC in equation 5 have been proved by omitting the specific features. (Reshef et al., 2013) There are three different variations of MIC which omits the maximization, omits the normalization, and omits both the maximization and normalization.

Definition 2 (Maximal information coefficient (MIC) without the maximization)
Let $E(D,x,y)$ when x-by-y is equipartition of data D. The number of rows and columns should the same with the points of data D. Then, let $I^E(D,x,y) = I\left(D|_{E(D,x,y)}\right)$ which defines the first variation of MIC without the maximization step by

$$MIC_1(D) = \max_{xy<B(n)} \left\{I^E(D,x,y)/\log_2 \min(x,y)\right\} \quad (6)$$

Secondly, the omitting of normalization step is considered on $log_2 \min(x,y)$ which is the upper bound of maximal mutual information on x-by-y of a uniform grid G. Then, the normalization returns the possible value between 0 and 1. Also, the grid with variant resolutions will be able to be compared.

Definition 3 (Maximal information coefficient (MIC) without the normalization)
Let I^* is the same as in equation 5 (definition 2.2.1). The data D is a set of ordered pairs, x and y, over a grid G. Then, the second variation of MIC without the normalization step is defined by

$$MIC_2(D) = \max_{xy<B(n)} \{I^*(D,x,y)\} \quad (7)$$

Thirdly, both the maximization and normalization steps are omitted

Definition 4 (Maximal information coefficient (MIC) without both the maximization and normalization)

$$MIC_3(D) = \max_{xy<B(n)} \left\{I^E(D,x,y)\right\} \quad (8)$$

From the three behaviors of omitting its features, the MIC with and without both utilities had been tested over the noisy relationship to show the score of coefficients R^2 in Fig. 3. The plots illustrate that the MIC interprets the stronger equitability than three mentioned variations of MIC. That is each data point conforms with an independent of noisy function in a given model. The equitability considers on how tightly of couple points. It is shown that the R^2 scores obtained the small range and does not scatter when comparing to other statistic.

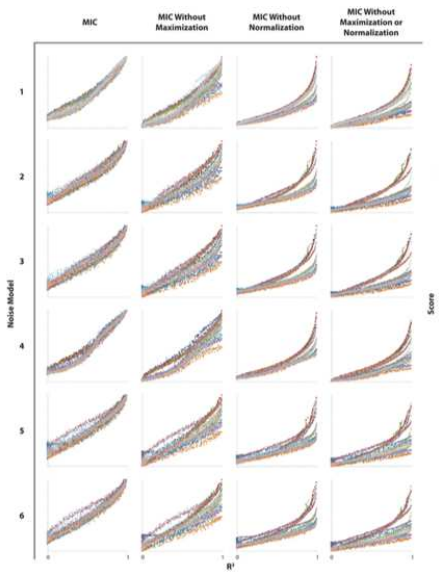

Fig. 3 the scatter plots scores of the MIC and three different characteristic features of MIC variations on the noise model (original picture from (Reshef et al., 2013))

In addition, the symmetric relation of pairwise variables due to the mutual information basis equalizes MIC score, $MIC_(x,y)(D)$ equals to $MIC_(y,x)(D)$. MIC carries powerful benefits on revealing association among variables: the outlier's robustness, no assumptions of variable distribution, and easiness of interpretation. It is widely applied across various research domains such as language recognition, image processing, searching based, spatial data analysis and medical. Wang et.al (Wang et al., 2018) improved MIC methodology offering iMIC, which was intelligent in searching the association relationship of variables in the system. The agriculture domain also selected MIC method to find the impact factors of the rice price variation(Surapunt et al., 2017).

3 Methodological approach

Bayesian maximal information coefficient (BMIC) is proposed to uncover the causal correlations from large dataset in the random system by integrating PGM and MIC with BLR. To identify the cause and effect, it finds out the best causality model on synergy to minimize runtime execution.

3.1 Search the correlated variables

MIC is to narrow down the number of variables in the search space by the generality and equitability which are beneficial in unlocking the accessibility of many functions and noise data manipulation. Since MIC algorithm provides a score for ranking their correlation coefficient, a high MIC score strengthens the dependence between pairwise variables, establishing the structural model. Therefore, the execution time reduces while searching for an optimized model in the structure learning process. Unfortunately, MIC returns an undirected model of variables' dependency, the symmetrical relationship. It implies that MIC score of the relationship $x \rightarrow y$ is equal to $y \rightarrow x$ that cannot represent the causality. The use of MIC requires a directional arrow between any two nodes. As a result, there is interest in studying the orientation of pairwise variables for identifying the cause-and-effect relationship.

3.2 Identify the causal relationship

PGM is the robust framework to provide functions for identifying the causal relationship model. The BN is a representation of the PGM framework, constructing a DAG by capturing the knowledge of the probabilistic numerical information. The conditional probability distribution denotes the variables' dependency, which considers the posterior probability value. They represent the edges in a particular network. The posterior probability is a basis of any inference which requires the integration of prior knowledge and new information. The Bayes' theorem captures the posterior probability by preferring the Bayesian perspective. Therefore, the Bayesian inference is a traditional method to detect random variables' causality. It is an influential statistical theory that can determine the probability of uncertain data.

3.3 Determine the posterior probability

The posterior probability values are determined from the linear regression model under Bayesian view. BLR (Clyde et al., 2007) applies the Bayesian theorem to the linear regression model for posterior probability estimation. BLR is compatible to the linear regression

model that is for understanding the linear relationship between input (feature) variable(s), and an output (target) variable. The linear function is formulated with probability distribution rather than a point estimation. The output parameter y is the response of normal distribution, controlled by a mean and variance as $y \sim N\left(\beta^T x, \sigma^2\right)$. The mean value is a product of parameter β^T and input variable x, while the variance value is a power of standard deviation. When the y variable of the linear regression model is completely computed, the posterior distribution for the model parameter given x and y variables will be detected by the support of Bayes' theorem as in Equation (9). The model parameters have followed the fundamental of Bayesian inference.

$$P(\beta \mid y, x) = P(y \mid \beta, x)P(\beta \mid x)/P(y \mid x) \tag{9}$$

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.

Since continuous values are intractable to parameterize for posterior distribution evaluation, the BLR model uses the Markov Chain Monte Carlo (MCMC) algorithm to sample the posterior distribution to estimate the posterior distribution. The Markov Chain represents the next sample value drawn by the previous sample value. And Monte Carlo means the technique of drawing a random sample. The concept of MCMC emphasizes the more the drawing of posterior distribution samples, the more convergent the posterior distribution between approximated and real value. PyMC3 (Salvatier et al., 2016) is a python library which is used for implementing the Bayesian model by the MCMC method. The Bayesian inference has constructed the Bayesian Linear Models from the formulated linear function. It is provided by the generalized linear models (GLM) module, which can generate the formula from the input variable(s), x, and an output variable, y.

The BLR does not restrict non-informative priors to assign a value for the model parameters. The non-informative prior draws from the normal distribution with the mean and variance of observed data. Another benefit is the ability to quantify the model's uncertainty from the result of the posterior probability distribution. The level of certainty depends on the data quantity; a low level of uncertainty model is due to increasing the quantity of data. Therefore, the BLR has been popular in statistical training that is widely applied in various domains. It is mainly used to examine the relationship between model variables. As in the psychology of (Baldwin and Larson, 2017), the BLR can outperform frequentist statistical methods in examining the relationship of an electroencephalogram (EEG) and anxiety from their clinical data.

The BLR has also been proposed in predictive modelling, corresponding with a predicted value and a confidence (probability) interval (CI).(Kong et al., 2020) The predicted value is the distribution of the target y_i given a set of features x_i : $P\left(y_i \mid x_i\right)$. Moreover, only the posterior probability distribution might not be enough for the Bayesian decision-making. The CI helps to confirm the possible values in a range for the model parameters. The more percentage of CI is, the more the relationship between pairwise variables should be. The study of geology in (Ghosh and Chakraborty, 2020) is based on the demand prediction model of the Seismic fragility that estimates the uncertainty of the fragility curve. The BLR provides more accurate results than other benchmarks.

The CI (Hespanhol et al., 2019) is an uncertainty measurement in frequentist statistics that can also declare a degree of uncertainty. It comprises the lower and upper limit for the result estimation that can be varied by the sample size of observed data and standard deviation (heterogeneity). The width of an interval can interpret the precision of the result estimates. When the data sample is large and has a narrow CI width, a low degree of uncertainty is indicated. However, the heterogeneity is directly proportional to the uncertainty degree as low heterogeneity refers to low uncertainty resulting from the narrow CI. Hence, the narrower the width of an interval is, the more precise the effect estimates are. The CI is preferable to the p-value, especially in the health science area, because of misinterpreting, misusing, and overestimating hypothesis testing. The p-value usually sets a boundary with a significant level of 0.05 (5 percent) for accepting or rejecting the defined null hypothesis. Many researchers have interpreted the p-value as a probability of accepting (rejecting) the null hypothesis as $P\left(H_0 \mid y\right)$ where is the probability of H_0 given the observed data. In fact, the p-value measures the probability of the actual result given the null hypothesis as $P\left(y \mid H_0\right)$. Moreover, researchers also attempt to oversimplify p-value interpretation in practice. They separate the statistically significant and non-statistically significant with the threshold of p-value at 0.05. Some concerns have been ignored the sample size and the variability of results estimation. Thus, the CI is promoted to be an alternative to the p-value, which can describe the variability of estimate and the interval's width. The interval's width indicates the precision of the result, one that is usually set at a 95 percent confidence level.

In the Bayesian approach, the uncertainty estimation of the posterior probability distribution is called

the Bayesian credible interval, which is abbreviated as CrI. The Crl performs under the same principle as CI. The estimating Crl of Bayesian inference evaluates the posterior distribution with two types of Crls: an equal tail Crl and the highest posterior density (HPD) interval. The equal tail Crl is a direct threshold value of posterior probability with an interval of the posterior probability distribution, which calculates easily. For example, the upper bound of Crl is 0.975 means 97.5 percent of the quantile distribution of posterior probability. However, asymmetric posterior probability can affect the yield estimate value with a lower probability inside the interval than outside the interval. The HPD interval also determines the threshold value of posterior probability, indicating an interval with the probability mass around the distribution center. When the HPD interval of -4.0 to -1.0 is 95 percent, it implies that the mean difference mostly emphasizes between -4.0 to -1.0 given the observed data. The possible value with the highest posterior probability would be between -4.0 to -1.0. The symmetry of posterior probability makes the HPD Crl equal to the equal tail Crl, but the HPD Crl is more complicated in a computation interval when comparing with the equivalent tail Crl method.

4 Case Study on Thai Rice Price

As a case study, the proposed BMIC is applied to discover the causal correlations from large dataset on Thai rice price by identifying the causality change in the paddy price of Jasmine rice. It properly determines the causality of changes in the price of Thai rice due to its uncertainty and related impacts.

4.1 Thai rice background

Thai rice price is an interesting topic to study the effects of price change due to being a well-known global trading product. Many countries are willing to import Thai rice despite its high price. Thailand is being the top rice exporter.(Fahmy, 2019) Thus, Thai rice is an imperative agricultural product to expand the country's revenue. Rice is not in a monopoly market because many countries have the ability to export, such as China, India, Thailand, Vietnam, Pakistan, Australia, and the USA.(Office of Agricultural Economics, 2017) However, other competitors are fortunate to adjust their rice price following the FOB of Thai rice price criteria. The FOB (Fahmy, 2019; Paisabazaar, 2019) stands for Free on Board which usually implies when shipping the product. According to the incoterms (International Commercial Terms), a trading contract mentions on an FOB that the costs and risks while transporting the cargo to the export port of the origin will be responsible by the exporter or shipper, which includes in the export price of the particular product. While the costs and risk from that point deliver to the destination will be conducted by importer or consignee. Therefore, the export price that is estimated under FOB will not include insurance, loading and unloading, freight, custom, vat, import duty and transportation cost (from the exporter's port to destination).

Unfortunately, Thai rice encounters problems, fluctuating the price of Thai rice, from global and domestic situations. Many agriculturists decide to plant alternative crops to gain more profit. It consequently causes the rice yield to drop. The food crisis in 2008 (Shah, 2008) was a severe situation that caused a spike of 50-100 percent in raw materials and food prices. The demand upsurge, stocks and farming area decrease, lack of the agricultural infrastructure and investment, oil price, and exchange rate influence the rising price, impinging the demand and supply in global markets.(Kha and Trinh, 2017)

Additionally, Thailand faced domestic problems with the Paddy Price Pledging scheme in 2011. The Paddy Price Pledging is the political machinery that supports agriculturists' income. The government allowed pledging the unlimited quota of a paddy and guaranteeing an obtainment of 50 percent higher than the market price. This policy affected to gradually increase the stock of Thai rice while the real market kept decreasing. It made the price of Thai rice suddenly more expensive than other competitors who lost the opportunity to trade with partners. In the study of (Wasawong, 2018), a linear model was developed to reveal the impact of Thailand's rice-pledging policy. The global trading rice price is an indirect cause setting the price of various types of Thai rice in a country. Because rice exporters are the first people to know the price direction from the global market, they attempt to negotiate with millers setting the rice price. Domestic rice price is forced to set a similar price with the export price.

During the same period of launching the Paddy Price Pledging policy, Thailand encountered an emergency disaster, the devastating flooding in 2011 which moved the first rank of being a global rice exporter from Thailand to India because of the rising Thai rice price. The high price attracted farmers to grow more rice. Both the real market and stock held a massive quantity of Thai rice. Unluckily, the revenue of Thai rice decreased, as partners had the advantage of a bargain, knowing that Thai rice needed to be sold. Thailand was under pressure to sell in 2014. The Thai rice price dropped to 385.91 US dollars per ton in 2015 (Thai Rice Exporters

Association, 2019), and all activity finally paused in 2016. The Thai rice price was then adjusted to the normal state compared to the price of Vietnamese rice.(Hoang and Meyers, 2015) Therefore, Thai rice price variation depends not only on change of time but also on impact factors.(BBC News, 2017) This paper aims to reveal the impact factors based on the probabilistic model, determining the causes of variation in the price of Thai rice. The relevant situations, such as the quantity of Thai rice for export, the Thai rice price, exchange rate, and other countries' rice prices, should be considered in detecting the variations in rice price.(Maneejuk et al., 2016) These results contribute to forecasting the Thai rice price for all stakeholders.

4.2 Data sample

Thus far, we have implicitly taken BN to demonstrate the causality of variation in the price of Thai rice due to many surrounding impacts. Every directed edge in a BN means the causal association between parent and child nodes. BN can further be used as a decision-making tool to deal with upcoming situations.

BN requires relevant data for Thai rice price model extraction. Thai paddy price, particularly the Thai Jasmine (Hom Mali) species, is a target to observe the causality of price fluctuation. There are two cultivation seasons that grow different yields. Eighty-five percent of rice grows between May and October and is harvested from August to April of the next year. This season is termed major rice which yields jasmine rice (Hom Mali rice) and parboiled rice. Another fifteen percent of rice grows during January to not later than April. It is called second rice which yields glutinous rice, non-glutinous rice, and indigenous species.(Office of Agricultural Economics, 2017) Accordingly, this paper selects the variables related to major rice to study the paddy price of jasmine rice.

This study's sample collects 11-years of data from 2008 to 2018, containing all relevant variables. As the price of Thai rice has become an interesting topic in economic and data science fields, variable selection has been taken from studying criteria in the following other literature. In economic studies, researchers focused on the relationships among rice prices or the competitive nature of rice in the market. The usability of various models (Baharom et al., 2009; Ghoshray, 2008) presented the asymmetric volatility of rice price, considering case studies in Thailand and Vietnam. They (Kha and Trinh, 2017) studied the surrounding issues of the Thai rice price between 2003-2013. In the data science area, many studies have evaluated some statistical techniques.(Sujjaviriyasup, 2018; Shao and Dai, 2018; Maneejuk et al., 2016; Co and Boosarawongse, 2007) The ARIMA model has been a popular tool to present a prediction result focusing on the export quantity of Thai rice, Thai rice price, exchange rate, and price of rice in other countries.(Maneejuk et al., 2016)

Based on the reasons above, there are 14 variables (including paddy price) with continuous values collected without missing the values of variables of interest. They are the plant area of major rice, the yield of major rice, minimum income of Thai citizen, paddy price of Jasmine rice, Jasmine rice price, export quantity of Jasmine rice, the export price of Jasmine rice, the export price of 5 percent white rice in Thailand, the export price of 5 percent white rice in Vietnam, the export price of 5 percent white rice in Pakistan, gold price, domestic oil price, US exchange rate with the Thai currency and Thailand's GDP of agricultural product. Since we learned that the paddy price could fluctuate due to domestic and global impacts, the selected 14 variables will be collected as relevant attributes. Domestic factors determine the relevant attributes in Thailand's economic mechanism, while global factors highlight competitors' prices on the same category of rice. Table 2 shows for each variable a longer definition (Meaning) and abbreviations (Variable names).

4.3 Exploratory Analysis with MIC

In this paper, we assume a probability model, BN, to understand the cause-and-effect relationship of factors that make Thai paddy price changes. The observed data, including target and predictor variables, is required for the learning structure process to bring out the complete BN. The structure model is learned by measuring the probability of variables' dependency without prior knowledge. The highest posterior probability distribution, given its data, is returned to confirm their relationships. However, there is required learning of many random variables to construct the model, which can cause a super-exponential growth of DAGs in the search space $\sim$ an NP-hard problem.

Our sample contains 14 variables prepared for constructing the model. The constraint-based and score-based algorithms that belong to the structure learning are executed on the observed data to identify edges between variables. The optimized graph is judged by holding the highest posterior probability distribution. We found that the run time spent a long time finding the optimized model due to the NP-hard problem. Therefore, we first apply MIC algorithm to narrow down the variable space, selecting only the highest strength relationship to the paddy price of Jasmine rice. In exploratory data analysis, the process to rank the coef-

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)

ficient is done by measuring pairwise variables' dependence. The correlation coefficient is returned as a MIC score in Table 3. The results are taken from the first ten-ordered pairwise variables because the perfect and strong relationship (as in Table 1) has a high impact on the target variable.

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

x variable	y variable	MIC score
mjYield	mjPlantArea	1
jasminePrice	paddyPrice	1
jasmineEp	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

Unfortunately, MIC results cannot directly imply the cause-and-effect relationship because of their symmetric relationship. When the x variable connects to the y variable, MIC score is identical to when the y variable is connected to the x variable. MIC results diagram can be used preliminarily to connect relevant variables with undirected edges in Fig. 4.

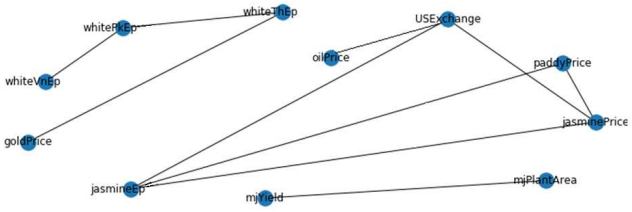

Fig. 4 The relationship diagram of random variables by MIC score

According to the results in Table 3, MIC score shows the perfect correlation of selected factors with the paddy price of Jasmine rice, the Jasmine rice price, and the export price of Jasmine rice. This means that the paddy price of Jasmine rice directly relates to both mentioned factors. The rest of MIC scores are not impractical. They can specify an indirect relationship to the paddy price of Jasmine rice. However, the perfect MIC score is found in the relationship of plant area and yield of major rice despite being unrelated to the paddy price of Jasmine rice. Although the export price of 5 percent white rice in Thailand, Vietnam, and Pakistan obtain a high MIC score, they are obviously irrelevant to the

paddy price of Jasmine rice. These preliminary results of MIC have demonstrated the connections with the paddy price of Jasmine rice conclusively. It seems beneficial to identify a variable's scope and notice only the real relationship to the target domain.

Since MIC score cannot detect the cause-and-effect relationship, it is compulsory to assign orientations to all edges in the diagram. BN representation has been chosen to depict the paddy price of Jasmine rice variation in a graphical model. The correlated data is a parameterization of BN used for parameter and structure learning(Bae et al., 2016). Zhang et al. (Zhang et al., 2013) improved the heuristic search of structure learning using MIC. They also found the triangular loops relationship problem, needing to eliminate the loop by the d-separation rule. Then, the conditional independence test further achieved the assigning orientation in their work. Therefore, our work is an idea to improve and contribute a new method to MIC algorithm, assigning direction to the undirected edges determining the cause-and-effect relationship.

5 Results and Discussion

5.1 Proposed BMIC (method of PGM and MIC with BLR)

Assigning the orientation of edges to MIC diagram in Fig. 3 follows BN construction rule. Since the analysis causality of paddy price of Jasmine rice variation operates with continuous variables, both target and predictor variables occupy the time series data. We attempt to learn the cause-and-effect relationship between nodes from MIC result. In contrast, MIC results' symmetry makes both directions of pairwise variables imply the relevant association.

For analysis, we use the open-source python package for Bayesian statistic, PyMC3, with models and probabilistic machine learning using gradient-based MCMC algorithms. The BLR better supports the Bayesian inference, and it is mostly used in summarizing the posterior distribution by a central tendency (mean) and an uncertainty estimate (variance). The response, y, of a linear regression model, is generated from a normal distribution of the predictor variables mentioned above. The posterior distribution is used to determine the Bayes' theorem's conditional probability from the linear regression parameter. The linear model formula will be created from MIC results, which assumes as $x \sim y$ and $y \sim x$. After that, the MCMC responses on model simulation perform the parameter estimation. The results are returned in a normal distribution (details of mean and variance) of the formula's intercept and predictor variables. Besides, an uncertainty measurement of the effect estimate requires the Cl or Crl. This experiment uses the HPD interval of 3 percent and 97 percent to determine the most probable parameter values. The interval's width represents the degree of uncertainty of the estimated model. The most precise linear model reveals the narrowest HPD interval.

All ten relationships are used to compute the posterior probability from the linear model. The MCMC algorithm is responsible for simulating the Bayesian model, which parameterizes from the identified linear formula. The observed data of parameters that belong to the linear function are compulsory to rely on the normal distribution. Each pairwise relationship's direction has been oriented by choosing the minimum HPD Crl according to the results in Table 4. The HDP Crls of every pairwise do not conform due to their highest posterior distribution.

Table 4 HPD Crls interval of MIC results in both directions of pairwise variables relationship

x variable	y variable	HPD Crl interval of $\mathbf{x} \rightarrow \mathbf{y}$	HPD Crl interval of $\mathbf{y} \rightarrow \mathbf{x}$
mjYield	mjPlantArea	0.245	0.413
jasminePrice	paddyPrice	0.14	0.129
jasmineEp	paddyPrice	0.127	0.133
jasminePrice	jasmineEp	0.138	0.129
whiteThEp	goldPrice	0.282	0.389
whiteThEp	whitePkEp	0.213	0.199
USExchange	jasmineEp	0.229	0.206
USExchange	oilPrice	0.197	0.203
USExchange	jasminePrice	0.246	0.23
whiteVnEp	whitePkEp	0.23	0.27

BN in Fig. 5 has been constructed to illustrate the direct and indirect effects on the paddy price of Jasmine rice following the HPD Crls of Table 4. We found four connections, divided into two groups of factors that are not relevant to the target. The first group is plant area and yield of major rice that return a perfect relationship. And the second group is the gold price and the export price of 5 percent white rice of Thailand, Vietnam, and Pakistan. However, we also found that the paddy price of Jasmine rice has clarified an effect by the export price of Jasmine rice. Also, the paddy price and export price of Jasmine rice impact the domestic jasmine rice price. Simultaneously, the US exchange rate with Thai currency and domestic oil price indirectly impact the paddy price of Jasmine rice.

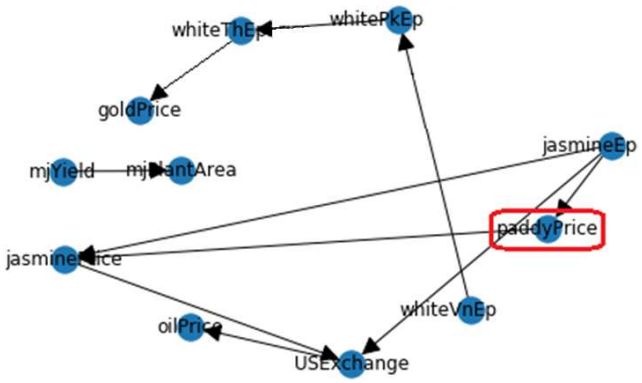

Fig. 5 Proposed BMIC (method of PGM and MIC with BLR)

5.2 The traditional structure learning experiment

BN, provided by either traditional or applied methods, constructs the model. The observed dataset is a crucial attribute realized through PGM framework's learning process, returning the model from the data. The connected nodes are captured by the conditional probability of Bayes' theorem, determining the dependency values. However, there is an issue of time complexity, which is an NP-hard problem. BN candidates grow super-exponentially in the search space, needing more time to find the optimized model. Searching for the best model will be done using heuristic algorithms attempting to find a model with an optimal score. This study selects the Hill Climbing search, a well-known and most straightforward algorithm for comparing networks with the baseline. The Hill Climbing search has been denoted as a fast-searching algorithm of the maximum n variables, $n(n-1)/2$ for the possible edges and $2(n(n-1)/2)$ for a subset of edges in a sample space.

The traditional experiment is tested on both constraint-based and score-based learnings. All variables are taken the consideration in the learning structure. Firstly, the constraint-based algorithm illustrates a BN, as in Fig. 6. The conditional independence is done by correlation analysis, which uses the Chi-square dependency test for estimating the model skeleton. The orientation should then be assigned to each connection between nodes according to the information of separating sets: a set that contains the conditional independence of each pair of indirectly connected nodes. The skeleton model and separate set information can estimate the partial DAG (PDAG), in which the relationships might have both-way direction, $\mathbf{x} \rightarrow \mathbf{y}$, and $\mathbf{y} \rightarrow \mathbf{x}$. Since the fundamental of BN avoids the occurrence of a v-structure and cyclic graph, BN is intuitively constructed as a DAG from PDAG. BN of the selected case shows that Jasmine rice's export price and domestic price control the paddy price of Jasmine rice. In-depth, the domestic price of Jasmine rice is impacted by the export price of Jasmine rice before affecting the paddy price of Jasmine rice. It can be noticed that there is no other effect that connects to the export price of Jasmine rice. Thus, we can state that the export price of Jasmine rice is an initiator of the paddy price of the Jasmine rice variation.

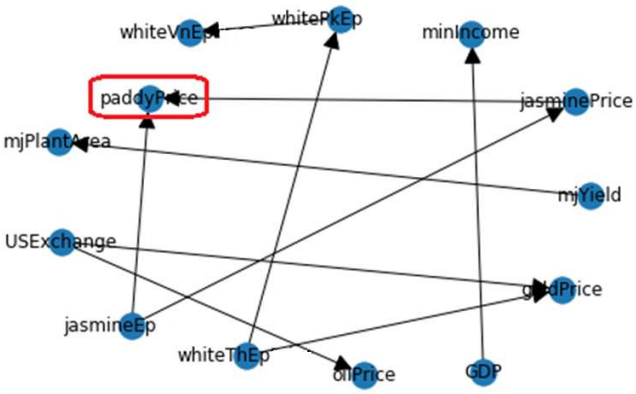

Fig. 6 BN of the constraint-based learning algorithm

Secondly, in Fig. 7, the score-based algorithm shows a structure. The Bayesian estimator will estimate the best model of each learning process to return the optimal score. BN considers the Bayesian Information Criterion (BIC) measurement. It is a log-likelihood score with Dirichlet priors for manifesting how well the given observed data describe a model. The method adds a penalty for network complexity to avoid overfitting. The score-based structure describes that the paddy price of Jasmine rice can be changed from the export price of 5 percent white rice in Pakistan. The dependent relationship is a chain following the export price of 5 percent white rice in Pakistan, domestic oil price, US exchange rate with Thai currency, domestic price, and ending at the export price of Jasmine rice. Therefore, both learning algorithms of traditional PGM return different models on the change of paddy price of Jasmine rice.

From the results of the proposed BMIC and traditional PGM, the three models, Fig. 5 to Fig. 7, indicate the different impact factors that can cause the paddy price of Jasmine rice to change. It is noticeable that pairwise nodes have similar connections but distinct directions. The proposed BMIC can reduce the number of relevant variables by selecting perfect and high strength relationships. The benefits of choosing only relevant variables for learning BN structure are to avoid the NP-hard problem. The result is entirely satisfied due to the conformity with another research. The

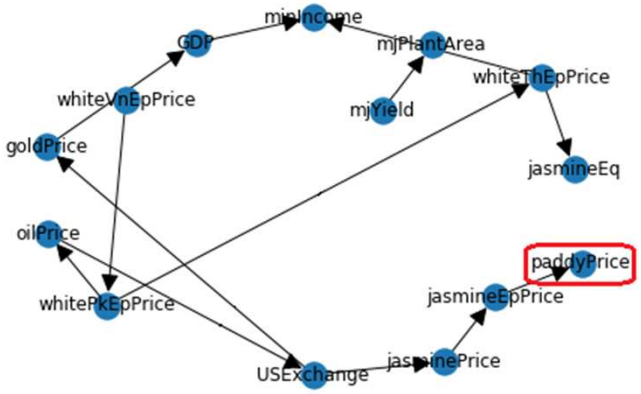

Fig. 7 BN of the score-based learning algorithm

traditional method has then been tested to compare their performance and model accuracy with the proposed BMIC. As in constraint-based learning, ten connections are contained in the best-selected model. The graph is quite similar to the result of MIC (undirected edges). Only one extra relationship: GDP of agricultural product and a minimum income of Thai citizen has been added. After assigning the orientation, the different direction of cause-and-effect relationships implies the opposite effects of Jasmine rice paddy price change. In addition, the score-based learning gives 13 relationships in the constructed BN. The dependency relationships are more complicated than the proposed BMIC and constraint-based learning. The result reveals that the change in the paddy price of Jasmine rice can occur through the forwarding of many effects.

As a result, each algorithm's explicit model reveals a different BN pattern. Therefore, the best BN model should be the most appropriate representation manifesting the cause-and-effect factors when the paddy price of Jasmine rice changes. We select the BIC score to measure how well-described the constructed BN is given its data. The BIC score is a log-likelihood calculation with an additional penalty to prevent overfitting, returning in a negative value. The best model should be assigned with the highest score. The BIC scores in Table 5 prove that the proposed BMIC returns the highest score when comparing with the other two learning algorithms of traditional PGM. Hence, the proposed BMIC obtains the best-fitted BN to learn the structure of the paddy price of Jasmine rice.

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

6 Conclusion

This paper utilizes BN framework's ability to interpret the model. In this study, BN framework, describes the causal model on the Thai rice price situation, focusing on changes in the paddy price of Jasmine rice. Since data uncertainty can be detected in the price of Thai rice and relevant attributes, the use of BN probability inference gives a better explanation. BN fundamentally establishes a model that relies on the Bayes' theorem by structural learning from the observed dataset. The NP-hard problem is usually encountered in the structure learning process because of a super-exponential number of graphs in the search space. We scoped the number of variables using MIC algorithm to capture only high strength relationships in the Thai rice price system. However, MIC result cannot explain any target domain's causality due to undirected relationships in MIC results. Therefore, the edge orientation is determined by the BLR model, emphasizing the return of posterior probability distribution. The BLR model assumes the normal (Gaussian) distribution sample to formulate the linear regression. The model parameters of BLR, calculating beneath the Bayes' theorem, are given the inputs and outputs, of the linear model. The Bayes' theorem handles the probabilistic, non-deterministic model. Since BN demonstrates independent relationships of a given joint probability distribution, each connected node presents a conditional probability that results in the posterior probability. The posterior distribution is computed in both directions on dependent variables, $\mathbf{x} \rightarrow \mathbf{y}$ and $\mathbf{y} \rightarrow \mathbf{x}$, from the undirected MIC results, resulting in the highest posterior probability distribution.

The significant posterior probability distribution is evaluated by the Bayesian credible interval in which the HPD interval is selected to be a criterion. The orientation prefers the direction of dependent variables holding a minimum value of the HPD interval. From the estimated edge direction results, the proposed BMIC reveals the causes and effects of the paddy price of Jasmine rice in the form of BN model. We found that the export price of Jasmine rice can change the paddy price of Jasmine rice. This detail is confirmed in the study of (Wasawong, 2018), which describes exporters' power to foreknow the global rice market direction. Moreover, the model also indicates that the paddy price of Jasmine rice can influence the domestic Jasmine rice price. This becomes supportive information for launching policies and coping with the future situation because we can monitor trends in Jasmine rice's paddy price and its

directed effects. The proposed BMIC is evaluated in reliability by comparing the network scores with a traditional PGM. The network score indicates how well-explained the model is and how it fits the model. The score shows that the proposed BMIC obtains the highest score while the two traditional learning algorithms' scores are lower. Hence, this study improves the optimized model selection with a new BN structure learning using MIC results with the BLR model. When the cause-and-effect relationships of the paddy price of Jasmine rice are finally identified in BN, this provides 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.

7 Ethics declarations

Ethical approval

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

Funding details

This work was supported by National Key R&D Program of China (2020YFC0832600), National Natural Science Fund of China (62076027).

Conflict of interest

The authors declare that they have no conflict of interest.

Informed Consent

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

8 Authorship 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 Tisinee Surapunt; Funding acquisition: Shuliang Wang. All authors read and approved the final manuscript.

References

A. H. Baharom, A. Radam, M. S. Habibullah, and M. T. Hirnissa, The volatility of thai rice price, Munich Personal RePEc Archive, 1-10 (2009)

D. A. Ghoshray, Asymmetric Adjustment of Rice Export Prices: The Case of Thailand and Vietnam, International Journal of Applied Economics, 5, 80-91 (2008)

T. Sujjaviriyasup, 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 (2018)

Y. E. Shao and J.-T. Dai, Integrated Feature Selection of ARIMA with Computational Intelligence Approaches for Food Crop Price Prediction, Complexity, 2018, 1-17 (2018)

S. Pandey, T. Sulser, M. W. Rosegrant, and H. Bhandari, Rice Price Crisis: Causes, Impacts, and Solutions, Asian Journal of Agriculture and Development, 7, 1-15 (2010)

M. Marloes, D. Mathias, L. Steffen, and W. Martin, Handbook of Graphical Models, Springer, New York (2013)

D. Koller and N. Friedman, Probabilistic graphical models: principles and techniques, The MIT Press, Cambridge (2009)

S. Singh, C. Mayfield, S. Mittal, S. Prabhakar, S. Hambrusch, and R. Shah, The Orion Uncertain Data Management System, The 14th International Conference on Management of Data, 273-276 (2008)

Y. Helong, C. Guifen, and L. Dayou, A Simplified Bayesian Network Model Applied in Crop or Animal Disease Diagnosis, Computer and Computing Technologies in Agriculture II, 294, 1001-1009 (2009)

K.-R. Liu, J.-Y. Kuo, K. Yeh, C.-W. Chen, H.-H. Liang, and Y.-H. Sun, 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 (2015)

P. A. Aguilera, A. Fernández, R. Fernández, R. Rumí , and A. Salmerón, Bayesian networks in environmental modelling, Environmental Modelling & Software, 26, 1376-1388 (2011)

V. Chawla, H. Naik, A. Akintayo, D. Hayes P. Schnable, B. Ganapathysubramanian, and S. Sarkar, 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 (2016)

P. Nuvaisiyah, F. Nhita, and D. Saepudin, Price Prediction of Chili Commodities in Bandung Regency Using Bayesian Network, International Journal on Information and Communication Technology (IJoICT), 4, 19-32 (2018)

D. A. Alvi Application of Probabilistic Graphical Models in Forecasting Crude Oil Price, University College London (2018)

D. M. Chickering, Learning Bayesian Networks is NP-Complete, Springer, New York (1996)

D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, and G. McVean, Detecting Novel Associations in Large Data Sets, Science, 334, 1518-1524 (2011)

J. Joyce, Bayes' Theorem, Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, California (2003)

K. Ruohonen, Graph Theory (Translation by Janne Tamminen, Kung-Chung Lee and Robert Piché) (2013)

M. F. Bari, Bayesian Network Structure Learning, 4th Annual Meeting Asian Assoc. Algorithms Comput (AAAC), 1-8 (2011)

K. P. Burnham and D. R. Anderson, Multimodel Inference: Understanding AIC and BIC in Model Selection, Sociological Methods & Research, 33, 261–304 (2004)

G. F. Cooper and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, Machine Learning, 9, 309–347 (1992)

D. Heckerman, Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning, 197–243, (1995)

S. Beretta, M. Castelli, I. Gonçalves, R. Henriques, and D. Ramazzotti, Learning the Structure of Bayesian Networks: A Quantitative Assessment of the Effect of Different Algorithmic Schemes, Complexity, 2018, 1-12 (2018)

Z. Liu, B. Malone, and C. Yuan, Empirical evaluation of scoring functions for Bayesian network model selection, BMC Bioinformatics, 13, 1-31 (2012)

Y. Cai, X. Luo, Z. Liu, Y. Qin, W. Chang, and Y. Sun, Product and process fingerprint for nanosecond pulsed laser ablated superhydrophobic surface, Micromachines, 10, 1–15 (2019)

Y. Liang, D. Abbott, N. Howard, K. Lim, R. Ward, and M. Elgendi, 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 (2019)

G. J. Székely and M. L. Rizzo, Brownian distance covariance, The Annals of Applied Statistics, 3, 1236–1265 (2009)

J. Benesty, J. Chen, Y. Huang, and I. Cohen, Pearson Correlation Coefficient, Noise Reduction in Speech Processing, 2, 1-4 (2009)

D. Reshef, Y. Reshef, M. Mitzenmacher, and P. Sabeti, Equitability Analysis of the Maximal Information Coefficient, with Comparisons, 1-22 (2013)

S. Wang, Y. Zhao, Y. Shu, H. Yuan, J. Geng, and S. Wang, Fast search local extremum for maximal information coefficient (MIC), Journal of Computational and Applied Mathematics, 327, 372–387 (2018)

T. Surapunt, C. Liu, and S. Wang, MIC for Analyzing Attributes Associated with Thai Agricultural Products, International Conference on Geo-Spatial Knowledge and Intelligence, 848, 40-47 (2017)

M. Clyde, M. Cetinkaya-Rundel, C. Rundel, D. Banks, C. Chai, and L. Huang, An Introduction to Bayesian Thinking, Springer, New York (2007)

J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, Probabilistic programming in Python using PyMC3, PeerJ Computer Science, 2, 1-20 (2016)

S. A. Baldwin and M. J. Larson, An introduction to using Bayesian linear regression with clinical data, Behaviour Research and Therapy, 98, 58-75 (2017)

D. Kong, J. Zhu, C. Duan, L. Lu, and D. Chen, Bayesian linear regression for surface roughness prediction, Mechanical Systems and Signal Processing, 142, 106770-106791 (2020)

S. Ghosh and S. Chakraborty, Seismic fragility analysis of structures based on Bayesian linear regression demand models, Probabilistic Engineering Mechanics, 61, 1-12 (2020)

L. Hespanhol, C. S. Vallio, L. M. Costa, and B. T. Saragiotto, Understanding and interpreting confidence and credible intervals around effect estimates, Brazilian Journal of Physical Therapy, 23, 290–301 (2019)

H. Fahmy, Classifying and modeling nonlinearity in commodity prices using Incoterms, The Journal of International Trade and Economic Development, 28, 1019–1046 (2019)

Office of Agricultural Economics, The 2017 agricultural economy statistical information of each commodity. `www.oae.go.th`, Accessed 20 April 2019 (2017)

Paisabazaar, FOB : FOB Price (Free on Board) : What is FOB Price ?, `https://www.paisabazaar.com/tax/fob-price/`, Accessed 15 May 2020 (2019)

A. Shah, Global Food Crisis 2008, `https://www.globalissues.org/article/758/global-food-crisis-2008`, Accessed 9 March 2020 (2008)

P. N. H. Kha and V. H. Trinh, Issues Surrounding the Rice Price of Thailand from 2003 to 2013, SSRN Journal, 1-32 (2017)

P. Wasawong, The Impact of the Rice Pledging Policy: The Case of Thailand, A Linear Programming Approach, Southeast Asian Journal of Economics, 6, 81-113 (2018)

Thai Rice Exporters Association, Rice Exports Statistics, http://www.thairiceexporters.or.th/List_%20of_statistic.htm, Accessed 15 May 2019

H. K. Hoang and W. H. Meyers, Price stabilization and impacts of trade liberalization in the Southeast Asian rice market, Food Policy, 57, 26–39 (2015)

BBC News, 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 (2017)

P. Maneejuk, P. Pastpipatkul, and S. Sriboonchitta, Analyzing the Effect of Time-Varying Factors for Thai Rice Export, Thai Journal of Mathematics, 201–213 (2016)

H. C. Co and R. Boosarawongse, Forecasting Thailand's rice export: Statistical techniques vs. artificial neural networks, Computers & industrial engineering, 53, 610–627 (2007)

S. Gao and Y. Lei, A new approach for crude oil price prediction based on stream learning, Geoscience Frontiers, 8, 183–187 (2017)

H. Bae, S. Monti, M. Montano, M. H. Steinberg, T. T. Perls, and P. Sebastiani, Learning Bayesian Networks from Correlated Data, Scientific Reports, 6, 1-14 (2016)

Y. Zhang, W. Zhang, and Y. Xie, Improved heuristic equivalent search algorithm based on Maximal Information Coefficient for Bayesian Network Structure Learning, Neurocomputing, 117, 186–195 (2013)

Figures

Figure 1

A simple graph of a directed relationship

Relationship functions	MIC	Pearson	Spearman	Mutual Information		CorGC	Maximal Correlation
				KDE	*Kraskov*	*Principal Curve-Based*	
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal *(Fourier frequency)*	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal *(non-Fourier frequency)*	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal *(varying frequency)*	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

Figure 2

The correlation score of noiseless models of different relationship functions

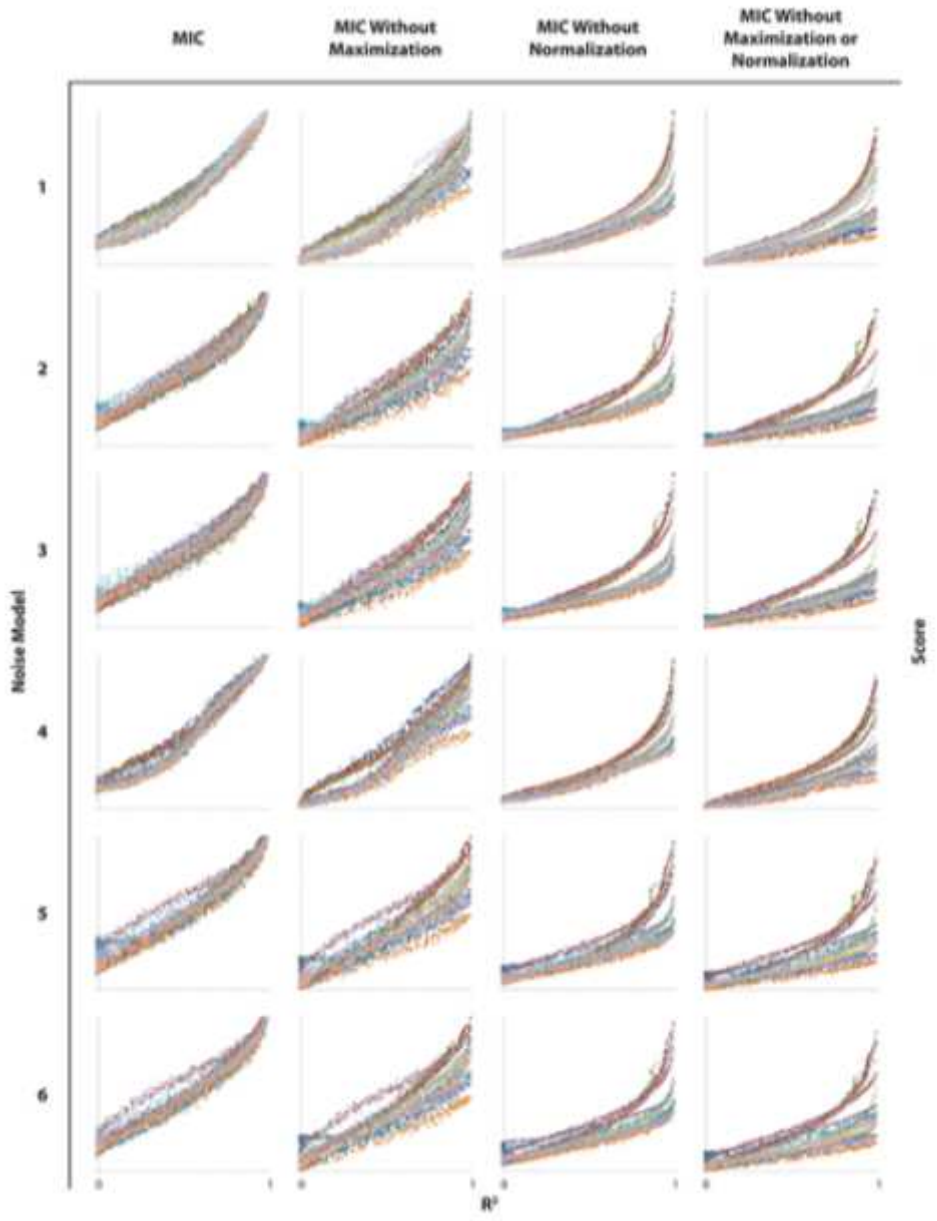

Figure 3

The scatter plots scores of the MIC and three different characteristic features of MIC variations on the noise model (original picture from (Reshef et al., 2013))

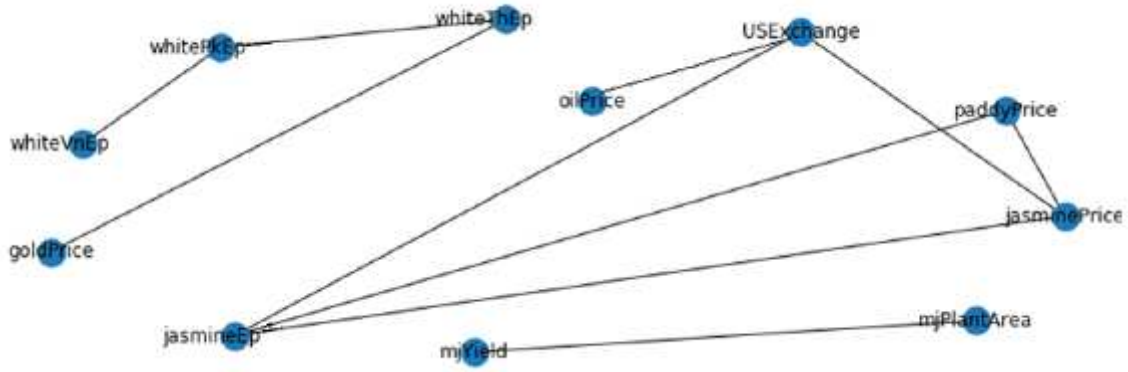

Figure 4

The relationship diagram of random variables by MIC score

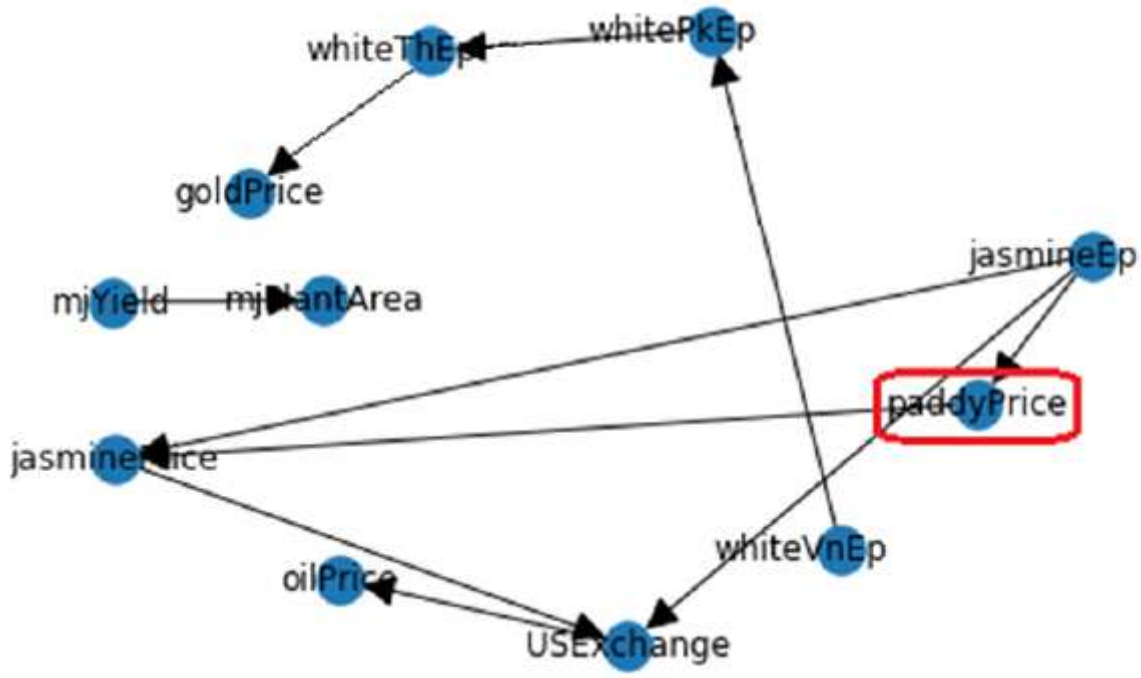

Figure 5

Proposed BMIC (method of PGM and MIC with BLR)

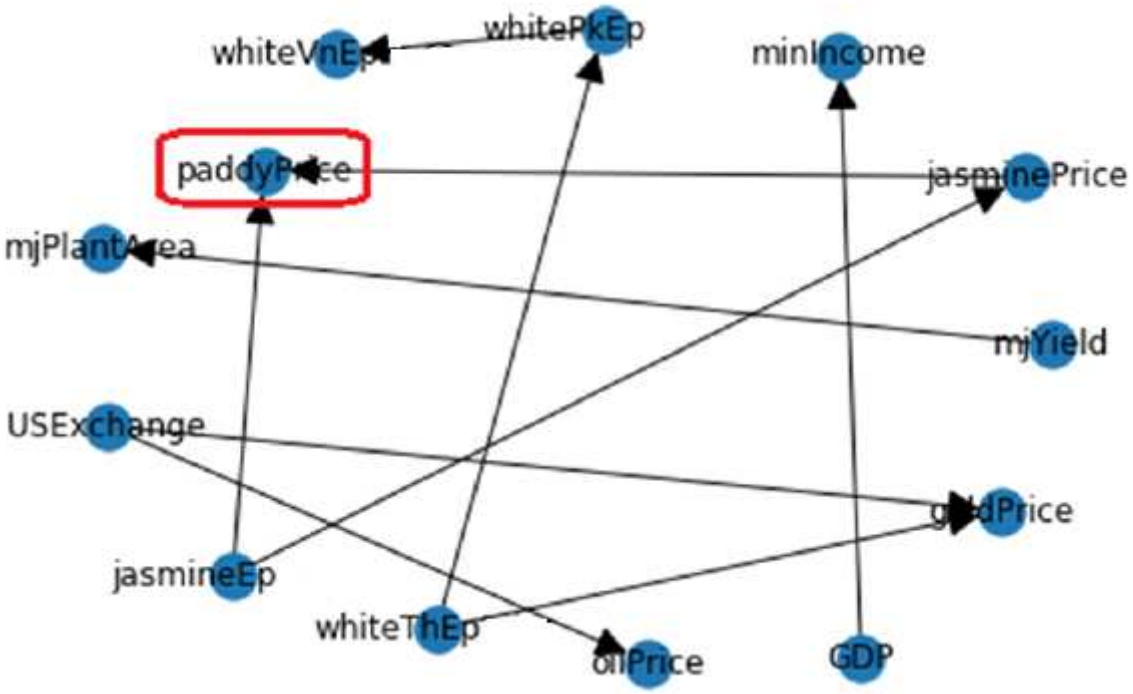

Figure 6

BN of the constraint-based learning algorithm

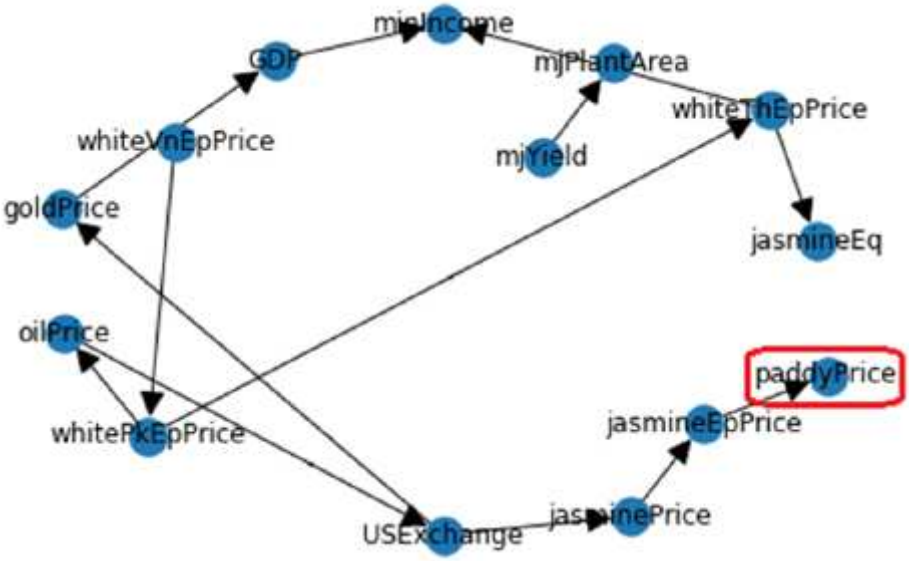

Figure 7

BN of the score-based learning algorithm