Multi-Model Variational Bayesian Approaches for Causality Analysis

by

Aswathi Prabhakaran

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Process Control

Department of Chemical and Materials Engineering
University of Alberta

Abstract

Causality analysis using data-driven models helps in the construction of graphical models that illustrate the interaction among the variables of a process system. A majority of industrial processes operate in multiple operating modes and thus the measurements from these processes exhibit multi-modal characteristics. However, the literature for causality analysis is skewed towards analyzing unimodal processes. In this work, we propose an approach for causality analysis in multi-modal systems.

Granger causality analysis is one of the widely popular methods for causality analysis. Classical techniques for multivariate Granger causality analysis rely on significance tests on parameters of vector autoregressive (VAR) models or vector moving average (VMA) models of the actual unimodal processes. In this work, we propose a Granger causality analysis technique with multi-modal VAR models. Our technique relies on variational Bayesian analysis of multi-modal VAR models. It imposes a soft constraint through Normal-Gamma priors on multi-modal VAR model parameters. This soft constraint ensures that the causal graphs extracted from different modes are consistent while allowing the strengths of interaction to vary across modes. Our approach also provides a single metric to assess the significance of each causal interaction in multi-modal systems. We illustrate the proposed algorithm using both simulation and industrial data. Furthermore, Bayesian network based approach for Granger causality analysis in multi-mode systems can handle data with outliers. The performance of the robust method is also tested using simulation and industrial process data.

Acknowledgements

First and foremost, I would like to express my sincere gratitude and acknowledge my indebtedness to my supervisior Prof Biao Huang for the inspiring guidance and support he has shown towards me throughout this thesis work. I truly believe that his constructive comments and feedbacks have immensely improved the way I conduct my research and approach a problem. I would also like to thank him for the support and empathy he has shown me during my maternity phase. He has been really understanding and provided me the flexibility to work remotely which has made the timely completion of this thesis possible. I am extremely grateful for what he has offered me.

I would like to extend my gratitude towards other members of the Computer Process Control (CPC) group for their help and support. I would particularly like to thank Dr. Rahul Raveendran for sharing his knowledge and expertise with me. He had provided me with ideas, valuable comments and feedbacks throughout my research work. I am extending my heartfelt thanks to my colleagues for their help and support, especially: Dr. Yanjun Ma, Anudari Khoshbayar, Oguzhan Dogru, Dr. Jayaram Velluru, Arun Senthil, Jingyi Wang, Faraz Amjad, Hareem Shafi, Yashas Mohankumar, Dr. Nabil Magbool Jan, Yousef Alipori , Dr. Fadi Ibrahim and several present and past members of the CPC group.

I am truly grateful to the industrial partners for giving me the opportunity to work on real industrial systems for my research.

I would like to acknowledge the financial support from Natural Sciences and Engineering Research Council (NSERC) of Canada.

A special thanks to my family. Words cannot express how grateful I am to my parents for all the love and encouragement they have given me. I would also like to thank my beloved husband, Unni and my brother, Arjun for supporting me through this journey. To my little girl, Ishita, thank you for always cheering me up. Finally, I thank god almighty for letting me through all the difficulties and successfully complete this research.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Causality analysis

Causality analysis has a broad range of applications ranging from neuroscience to economics [1], [2]. In process industry as well, causality analysis plays an important role. Maintaining normal operation is one of the primary challenges in process industries. Industrial processes often tend to drift from normal operations. These abnormalities in the normal operation can adversely affect plant performance and quality of products. The process variables often interact with each other, thus making it difficult to identify the variable causing the abnormality in the plant operations. Causality analysis helps in understanding relationships among the variables of the system and identifying the root cause of an abnormality. Causality analysis techniques fall into two main categories, namely knowledge based techniques and data-driven techniques. Knowledge about the process can help in analyzing the cause-effect relations among the variables. However, knowledge based causal relation evaluation is time-consuming and in most cases, the detailed knowledge of the process is not available. Causality study based on data-driven models can help circumvent this challenge. Data-driven causality analysis helps in identifying the cause and effect relationships among the variables and subsequently aids in the construction of a causal graph. Structural knowledge about a system helps in answering questions such as how does changes in one variable reflect in another variable?, if there is a fault in one variable, how is it

going to affect another variable? and how this change traverses through the system? Numerous data-driven approaches have been developed to identify cause and effect relationships among the variables, they include Granger causality [3], coherence-based methods [1], [4] and entropy-based methods [5], [6].

The main focus of this research is on Granger causality. Existing formulations of Granger causality have many limitations. Granger causality definitions rely heavily on observed variables. Nevertheless, in reality not all variables of the analysed system are measured and available which undermines the performance of Granger causality. Guo et.al., 2008 [7] proposed the concept of partial Granger causality to account for the effect of confounding or unknown variables. The method is based on the intuition that effect of confounding variables will be reflected in the correlation between prediction errors of MVAR models of measured variables. Thus, it removes the effect of unknown external variables. The underlying assumption of partial Granger causality is that the confounding variables have equivalent effects on all the measured variables. Another drawback of the Granger causality concept defined earlier is its applicability on stationary data only. To deal with the non-stationary data, Hesse et al. (2003) [8] adopted a windowing technique based on the assumption that in a very small window, non-stationary data is considered to be stationary.

Sometimes, causality analysis in time domain cannot capture causal connections among variables. Particularly, in time series data having cyclic patterns, examining Granger causality in spectral domain will give more accurate results than the time-domain analysis. The frequency domain representation of Granger causality called spectral Granger causality [9] is used in such situations and it is widely applied in neuroscience. In this thesis we are particularly interested in causality analysis of multi-model systems. Application of traditional Granger causality techniques for multi-modal systems and subsequent determination of causal structure of such a process is a very tiring task.

Inference of causal structure using traditional Granger causality technique for bi-

(a) Mode 1 (b) Mode 2

Figure 1.1: Causal graphs extracted from two modes of the same process. r causes z in Fig 1.1a, while x causes z through a third variable r in Fig 1.1b

variate system involves statistical tests. To construct the causal structure of a multivariate system, a series of statistical tests need to be carried out. This makes the determination of causal structure in multivariate system a tedious task. Each one of these tests performed for the multivariate case is similar to the test done for a bivariate case. Now, imagine the number of statistical tests that need to be done for an actual industrial processes where the number of variables can be very large. The proposed method suggests a much simpler statistical test to check for the significance of the causal connections. Besides, many industrial processes have more than one mode of operation. In such cases, a single VAR model cannot fit the data accurately. Switched VAR model is usually used to represent such data. The existing causality methods for multi-model systems have many drawbacks. A major drawback being that, the existing causality methods do not guarantee that the causal structures extracted from different modes are the same. This in turn could lead to inconsistent results. To understand this better, consider a multi-model system with three variables x, z and r. Suppose, the system has two modes of operations and the causal structure extracted from the two modes are shown in Fig 1.1. The two figures give contradictory results, as in the first mode there is no causal relation between x and r, whereas the causal graph from mode two shows x to be causing r. In such a situation, no final conclusion can be drawn regarding the cause and effect relationships among the variables. The proposed method helps to overcome this drawback of the existing methods. Additionally, the proposed Bayesian approach for Granger causal-

ity can be extended to the case when the data is contaminated with outliers. The effect of outliers is minimized by modeling the prediction error of the VAR model as a t-distribution.

1.2 Thesis Contributions

1. Development of a Granger causality technique which can be applied to multi-modal systems. It uses a switched VAR model such that model switches depending on the mode of operation of the system.

2. Variational Bayesian approach is used to infer the causal relations in each mode such that the causal structures extracted from different modes are consistent.

3. Variational Bayesian approach helps to develop a penalized approach for estimating the parameters of the VAR model so that the irrelevant causal connections vanish.

4. Our method provides a simple statistical test to check the significance of the causal connections.

5. The proposed method can be extended to find the causal structures of the systems where the process data collected is contaminated with outliers.

6. The performance of the proposed methods is tested on simulation and industrial examples.

1.3 Thesis Outline

The remainder of this thesis is organized as follows: In chapter 2, a brief overview of the data-driven causality methods, namely coherence-based methods, entropy-based methods and Granger causality. In chapter 3, the existing Granger causality methods for multi-model systems and their shortcomings are discussed. The proposed approach

given in this chapter is able to overcome many of these drawbacks and also has many added advantages. In this thesis, the data-driven models associated with Granger causality are expressed graphically using Bayesian networks (BNs) which are a special type of probabilistic graphical models (PGMs). Furthermore, to make the inference algorithm tractable, all the Bayesian networks considered in this thesis belong to a special class of Bayesian networks known as conjugate exponential family graphical models (CEFGMs). The efficiency of the method is tested using simulation data.

Chapter 4, discusses an extension of the above method to address data with outliers and develops a robust Granger causality technique for multi-model systems. The effect of outliers is alleviated by modeling the noise as a t-distribution variable. The efficacy of this method is evaluated using the same simulation example. Furthermore, the dependence of several parameters on the performance of the method is also studied.

In chapter 5, a comparative study between the above two methods is done to demonstrate the better performance of the robust method for an industrial data corrupted with outliers. Finally, chapter 6 concludes the thesis.

Chapter 2

Background

The motivation behind causality analysis is to identify variables affecting variables of interest such as quality of the product, a safety variable or any other variable which should be well maintained within a specified range. Causality analysis also helps in the construction of a causal map for a process plant by inferring the cause and effect relationships among the variables which can play a vital role in the root cause analysis of plant disturbances. This chapter provides a brief overview of the popular data-driven causality methods, namely coherence-based methods, entropy-based methods and Granger causality.

2.1 Coherence-based methods

Coherence measures the correlation among time series signals as a function of its frequency components. However, coherence does not provide any temporal relationships among the variables considered. Hence, it cannot be used in cause and effect analysis. To provide insights into the functional connections among the variables, the idea of directed coherence (DC) evolved. DC splits coherence into feedforward and feedback interactions, which helps to unravel the cause and effect relationships among the variables. The coherence formulation can help understand this concept clearly.

Consider that the time series data of three random variables x, z and r are available. Let vector Y be constituted of the variables x, z and r. The VAR model for the three-

variable system is given as the following,

$$y(t) = \sum_{l=1}^{L} W(l)y(t-l) + e(t) \tag{2.1}$$

where $y(t)$ is the vector comprising of $x(t), z(t)$ and $r(t)$. $W(l)$ is the coefficient matrix at lag l and $l = [1, 2, ..., L]$, $y(t-l)$ comprises of the values of variables x, z and r at lag l and $e(t)$ represents the noise of the process and has three components $e_1(t), e_2(t)$ and $e_3(t)$ which represent the process noise associated with the prediction models of x, z and r respectively. Thus, the above expression can be expanded as given below,

$$\begin{bmatrix} x(t) \\ z(t) \\ r(t) \end{bmatrix} = \sum_{l=1}^{L} W(l) \begin{bmatrix} x(t-l) \\ z(t-l) \\ r(t-l) \end{bmatrix} + \begin{bmatrix} e_1(t) \\ e_2(t) \\ e_3(t) \end{bmatrix} \tag{2.2}$$

where

$$W(l) = \begin{bmatrix} W(l)_{11} & W(l)_{12} & W(l)_{13} \\ W(l)_{21} & W(l)_{22} & W(l)_{23} \\ W(l)_{31} & W(l)_{32} & W(l)_{33} \end{bmatrix} \tag{2.3}$$

$W(l)_{11}, W(l)_{12}, ...W(l)_{33}$ are the elements of the coefficient matrix at lag l. The values of these elements indicate the presence of causal relation between particular input variable and corresponding output variable. Coherence-based methods determine causality in the frequency domain. To perform spectral analysis at frequency f, the frequency transformation of the VAR model (equation 2.1) is carried out which results in the following equation,

$$y(f) = W(f)y(f) + e(f) \tag{2.4}$$

where $y(f)$ and $e(f)$ are the Fourier transform of Y and E respectively and $W(f)$ is sum of the Fourier transformations of the coefficient matrices for all the lags and is given as the following [1],

$$W(f) = \sum_{l=1}^{L} W(l)z^{-l}|_{z=e^{-i2\pi f}} \tag{2.5}$$

where i is the imaginary unit, $W(l)$ is coefficient matrix at lag l and f is the frequency. Now, the transfer function matrix $H(f)$ is obtained by rearranging equation 2.4 as the following,

$$y(f)\left[I - W(f)\right] = e(f) \tag{2.6}$$

$$\implies y(f) = \bar{W}^{-1}(f)e(f) = \left[I - W(f)\right]^{-1} e(f) = H(f)e(f) \tag{2.7}$$

where

$$H(f) = \begin{bmatrix} H(f)_{11} & H(f)_{12} & H(f)_{13} \\ H(f)_{21} & H(f)_{22} & H(f)_{23} \\ H(f)_{31} & H(f)_{32} & H(f)_{33} \end{bmatrix} \tag{2.8}$$

where the elements of transfer matrix, $H(f)_{11}, H(f)_{12}, ..., H(f)_{33}$ are called the directed transfer functions (DTFs). The first concept which was developed to evaluate causal relations among variables from the VAR model (equation 2.7) after frequency transformation is directed coherence (DC). The directed coherence from variable z to x of the above mentioned three-variable system is given as the following,

$$DC_{z\to x}(f) = \frac{\sigma_{22}H_{12}(f)}{\sqrt{\sum_m^{1,2,3} \sigma_{mm}^2 |H_{1m}(f)|^2}} \tag{2.9}$$

where H_{12} is the element of the transfer function matrix $H(f)$ and is called as the directed transfer function (DTF) from z to x. H_{12} shows the causal influence of past of z on x. In the same lines, directed transfer functions terms H_{x11}, H_{12} and H_{13} of the denominator of above equation show the causal influence of past of x, z and r respectively on x. σ_{mm}^2 represents the diagonal elements of the covariance matrix σ given as the following,

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 \end{bmatrix} \tag{2.10}$$

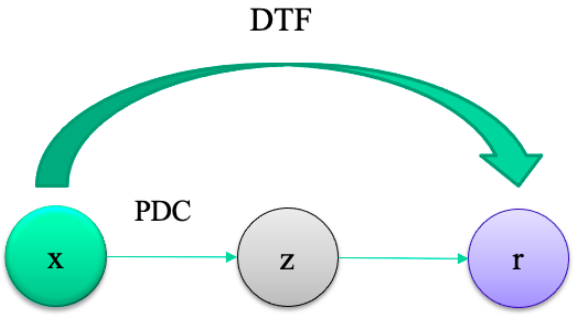

Figure 2.1: Direct and indirect effects captured by DTF and PDC

Under the assumption that Σ is diagonal, the square of DC given in equation 2.9 can be interpreted as the ratio of power of x contributed by past of z to the total power of x at frequency f [10]. This provides a directionality to the relation between x and z. Further, to overcome the restriction of diagonal covariance, Kamiński and Blinowska [11] proposed a normalized directed transfer function (DTF) given below which avoids the covariance term,

$$DTF_{z\to x}(f) = \frac{H_{12}(f)}{\sqrt{\sum_m^{1,2,3}|H_{1m}(f)|^2}} \tag{2.11}$$

where $H_{1,1}, H_{12}$ and H_{13} are directed transfer functions which give the causal relation of past of x, z and r on x respectively. The term in the denominator is used to normalize the DTF. Under unit variance condition, normalized DTF can be considered to be equal to the fraction of total power of x which is contributed by the past of z. DTF measures the total effect of one variable on another variable. For instance, for the three-variable system considered, when there exists no direct relation between x and r as given in Fig 2.1, DTF still measures the indirect effect of x on r through the third variable z. To measure the direct effect of one variable on another variable in such multivariate systems, the concept of partial directed coherence (PDC) was later introduced [1]. Partial directed coherence from z to x for the same three-variable

system is represented as the following,

$$PDC_{z\to x}(f) = \frac{\bar{W}_{12}(f)}{\sqrt{\bar{w}_z^H(f)\bar{w}_z(f)}} \tag{2.12}$$

where $\bar{W}(f) = I - W(f)$ and I being an identity matrix with dimensions equal to the dimension of $W(f)$. $\bar{W}(f)$ is derived as the following,

$$\implies \bar{W}(f) = \begin{bmatrix} 1 - W(l)_{11} & -W(l)_{12} & -W(l)_{13} \\ -W(l)_{21} & 1 - W(l)_{22} & -W(l)_{23} \\ -W(l)_{31} & -W(l)_{32} & 1 - W(l)_{33} \end{bmatrix} \tag{2.13}$$

$$= [\bar{w}_1(f)\bar{w}_2(f)\bar{w}_3(f)] \tag{2.14}$$

where $\bar{w}_1(f), \bar{w}_2(f), \bar{w}_3(f)$ represent the three columns of $\bar{W}(f)$ and superscript H represents the Hermitian transpose. $\bar{W}_{12}$ is the element of the matrix $\bar{W}(f)$, which is the Fourier transform of the VAR coefficient matrix. $PDC_{z\to x}(f)$ compares the effect of past of z on the current value of x to the effect of past z on all the other variables. The term in the denominator normalizes the expression such that the value of PDC ranges from 0 to 1. Unlike the DTF, PDC measures only the direct effect of one variable on another.

2.2 Entropy based methods

Mutual information and transfer entropy are the entropy based methods used for detection of cause and effect relationship among variables. These concepts are developed from Shannon entropy which quantifies the uncertainty of variable x, based on its probability density function $p(x)$. Shannon entropy is mathematically represented as the following,

$$H_x = \sum_x p(x) \log \frac{1}{p(x)} \tag{2.15}$$

The difference between the Shannon entropies of two probability density functions is defined as Kullback entropy which is given as the following,

$$K_x = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{2.16}$$

where $q(x)$ and $p(x)$ are the two probability density functions assumed for x. Mutual Information is a special kind of Kullback entropy which measures amount of dependency between two random variables x and z as the following,

$$M_{xz} = \sum_{x,z} p(x,z) \log \frac{p(x,z)}{p(x)p(z)} \tag{2.17}$$

where $p(x)$ and $p(z)$ are the marginal probability density functions of x and z respectively, while $p(x,z)$ is the joint probability density of x and z. Now, if x and z are independent, then numerator and denominator inside the log term become equal which makes the mutual information zero. However, mutual information does not give any idea about the direction of influence between the two variables. Some sense of directionality can be achieved from time lagged mutual information expressed as

$$M_{xz}(h) = \sum_{x,z} p(x(t), z(t-h)) \log \frac{p(x(t), z(t-h))}{p(x(t))p(z(t-h))} \tag{2.18}$$

where h and t represent the time lag and the current time instant respectively. $p(x(t), z(t-h))$ is the joint probability density of $x(t)$ and $z(t-h)$, while $p(x(t))$ and $p(z(t-h))$ are marginal probability density functions of $x(t)$ and $z(t-h)$ respectively. The value of h which gives the highest mutual information is considered as time lag of z from x. Now, suppose we find that history of z influences the current value of x. In some cases, the two variables may just be correlated and this is the reason why we see the influence of past values of z on x. In reality, the current value of x may also get all the information from its own past values and no additional information is supplied by the history of z. To make a clear distinction among such cases, transfer entropy is used, as it takes into account history of the variable itself. The expression of transfer entropy is the following,

$$T_{z \to x} = \sum_{x,z} p(x(t), x(t-1)^k, z(t-1)^l) \log \frac{p(x(t)|x(t-1)^k, z_(t-1)^l)}{p(x(t)|x(t-1)^k)} \tag{2.19}$$

where k and l are the number of previous time instants considered for x and z respectively i.e. $x(t-1)^k = [x(t-1), x(t-2), ..., x(t-k)]$ and $z(t-1)^l = [z(t-1), z(t-$

$2), ..., z(t-l)]$. $p(x(t), x(t-1)^k, z(t-1)^l)$ is the joint probability density of predicted value of x, k past values of x and l past values of z. $p(x(t)|x(t-1)^k, z(t-1)^l)$is the conditional probability density function of predicted value of x given k and l past values of x and z respectively. Similarly $p(x(t)|x(t-1)^k)$ is conditional probability density function of $x(t)$ given its k past values in the immediate past. Transfer entropy measures the effect of past of z on the future value of x and thus helps in evaluating the causal relation between the two variables.

2.3 Granger causality

Among the data-driven causality analysis techniques, Weiner-Granger causality also often simply referred to as Granger causality is probably the most popular technique.

2.3.1 Bivariate case

Consider time series data of two variables x and z, the variable z is said to Granger cause the variable x if the past states of both x and z combined give a better prediction of x than just the past states of x alone. In a simple system of two variables x and z, two separate prediction models are constructed for x. The first model consists of just the past values of x as input and the second model has past values of both variables x and z as inputs [3]. A reduction in the prediction error variance with the inclusion of past values of z in the model is said to indicate that z Granger causes x. The mean square of prediction error of the two prediction models is subjected to statistical tests to assess the improvement in the prediction accuracy of x when the past values of z are included in the model. Consider the two prediction models for x as follows,

$$x(t) = \sum_{l=1}^{L} W(l)'_{11} x(t-l) + e_{1,R}(t) \tag{2.20}$$

$$x(t) = \sum_{l=1}^{L} W(l)_{11} x(t-l) + \sum_{l=1}^{L} W(l)_{12} z(t-l) + e_{1,UR}(t) \tag{2.21}$$

where $x(t)$ is the predicted value of variable x at time instant t, while $x(t-l)$ and $z(t-l)$ are the past values of x and z respectively at lag l. L represents the extend of the time lag. Prediction models given in equation 2.20 and equation 2.21 are called restricted and unrestricted models respectively as the the first model is more restricted since it has just the past values of one variable (here x), while the second model has past values of both the variables x and z. $W(l)'_{11}$ is the coefficient of restricted prediction model at lag l, while $W(l)_{11}$ and $W(l)_{12}$ represent the coefficients of unrestricted prediction model at lag l. $e_{x,R}$ and $e_{x,uR}$ are the noise of the two models with subscripts R and UR indicating restricted and unrestricted respectively. The magnitude of Granger causal relation $F_{z\to x}$ [9] is defined as the logarithm of the F-statistic as follows,

$$F_{z\to x} = \ln \frac{var(e_{1,R})}{var(e_{1,UR})} \tag{2.22}$$

where the F-statistic is the ratio of prediction error variances of the restricted and unrestricted models given as $var(e_{x,R})$ and $var(e_{x,UR})$ respectively. If $F_{z\to x} > 0$, it implies Granger causality. However, to conclude if the causal connection is relevant, additional statistical tests have to be conducted. One such test involves the F statistic given in equation 2.22, which under null hypothesis follows a F-distribution. The null hypothesis of the statistical test is that the coefficients associated with the past values of z are jointly equal to zero. The rejection of this null hypothesis implies that z Granger causes x.

2.3.2 Multivariate case

The VAR model for a multivariate system is represented as follows,

$$y(t) = \sum_{l=1}^{L} W(l)y(t-l) + e(t) \tag{2.23}$$

where $y(t) \in R^D$ is the observation at time instant t with dimension D, $W(l) \in R^{D\times D}$ is the coefficient matrix at lag l and $e(t) \in R^D$ is a noise which follows a Gaussian distribution with zero mean. If the values of the ij^{th} elements of the coefficient

matrices $W(1), W(2), ..., W(L)$ i.e. for all the lags l are zero, it would imply that j^{th} component of $y(t)$ does not Granger cause the i^{th} component of $y(t)$. The relevance of the causal connections can be checked using statistical tests to discard insignificant causal connections. Now, the statistical test for a multivariate case is very tedious when compared to the bivariate case discussed earlier. A technique of conditional Granger causality is used in the multivariate case. Conditional Granger causality can essentially be considered as an extension of bivariate Granger causality to a multivariate case. To understand conditional Granger causality better, consider the simplest multivariate system with three variables x, z and r, then $y(t)$ in equation 2.23 is given as follows,

$$y(t) = \begin{bmatrix} x(t) \\ z(t) \\ r(t) \end{bmatrix} \tag{2.24}$$

Then, the VAR model for the three-variable system can be written as follows,

$$\begin{bmatrix} x(t) \\ z(t) \\ r(t) \end{bmatrix} = \sum_{l=1}^{L} \begin{bmatrix} W(l)_{11} & W(l)_{12} & W(l)_{13} \\ W(l)_{21} & W(l)_{22} & W(l)_{23} \\ W(l)_{31} & W(l)_{32} & W(l)_{33} \end{bmatrix} \begin{bmatrix} x(t-l) \\ z(t-l) \\ r(t-l) \end{bmatrix} + \begin{bmatrix} e_1(t) \\ e_2(t) \\ e_3(t) \end{bmatrix} \tag{2.25}$$

To inspect the causal relation between x and z, the unrestricted and restricted regression models of the x component of the VAR model need to be constructed which are analogous to equations 2.20 and 2.21 respectively. The two prediction models are as follows,

$$x(t) = \sum_{l=1}^{L} W(l)'_{11} x(t-l) + \sum_{l=1}^{L} W(l)'_{13} r(t-l) + e_{1,R}(t) \tag{2.26}$$

$$x(t) = \sum_{l=1}^{L} W(l)_{11} x(t-l) + \sum_{l=1}^{L} W(l)_{12} z(t-l) + \sum_{l=1}^{L} W(l)_{13} r(t-l) + e_{1,UR}(t) \tag{2.27}$$

where $W(l)'_{11}$ and $W(l)'_{13}$ are coefficients of restricted regression model, while $W(l)_{11}$, $W(l)_{12}$ and $W(l)_{13}$ represent the coefficients of the unrestricted model. The regression

models for evaluating conditional Granger causality between x and z consist of the past values of the third variable r. Thus, the effect of past values of r is accounted for in both the equations. Except for the past values of z, both the models contain past values of x and r. In such a case, if the prediction accuracy of the unrestricted regression model is higher that the restricted model, it implies that, the improvement is solely due to inclusion of the past values of z in the unrestricted model. The expression of conditional Granger causality $F_{z\to x|r}$ [12] which gives the magnitude of Granger causality from z to x given r for the three-variable system is given as the following,

$$F_{z\to x|r} = \ln \frac{var(e_{1,R})}{var(e_{1,UR})} \tag{2.28}$$

$var(e_{1,R})$ and $var(e_{2,UR})$ represent the variances of the restricted and unrestricted model errors respectively. To construct the causal structure of a multivariate system, a series of statistical tests needs to be carried out. This makes the determination of causal structure in multivariate system a tedious task. For instance, for the earlier three-variable system, if we want to check whether x causes r, a different reduced regression model containing just the past values of x and z needs to be constructed like the one given in equation 2.26 and subsequently, the statistical test has to be conducted. This process has to be repeated for evaluating the causal relation between each pair of variables. Each one of these tests performed for the multivariate case is similar to the test done for a bivariate case.

2.4 Conclusions

This chapter of the thesis gives a background on data-driven causality techniques, particularly coherence-based methods, entropy-based methods and Granger causality. Further, this chapter gives the basic formulation of these techniques for causality analysis.

Chapter 3

A variational Bayesian approach for causality analysis in multi-modal systems

3.1 Introduction

Granger causality which is widely used for causal inference in time series data has several shortcomings. This may greatly restrict the applicability of the method on real systems. Major drawbacks of the classical definition are covariance stationary assumption of data, applicability to linear systems and the dependence on selection of observed variables. Over the years, several extensions have been made for the basic Granger causality technique to overcome these drawbacks. In this chapter, a novel Granger causality technique for multi-model systems using the variational Bayesian approach is proposed. The chapter also gives a brief literature review of the existing multi-model Granger causality techniques, lists their drawbacks before moving onto the formulation of the proposed method.

Numerous extensions are available in the literature to implement Granger causality to multi-model systems. Freiwald et al.(1999) [13] in their work inferred causal relations in a multi-model system by identifying locally linear neighborhoods for the non-linear data and applied Granger causality in each of these linear neighbourhoods. In general, the methods for determining Granger causality in non-linear systems

which operate around multiple steady states fall mainly into two categories, namely information-theoretic approach-based methods and nonlinear predictor construction based methods. Information-theoretic based methods such as transfer entropy [14] and conditional mutual information [15], [16] rely on the measurement of entropy to infer non linear Granger causality, while kernel method based causality methods [17] fall under the latter category. Kernel methods are based on the idea that after transforming data points from a lower dimensional feature space to a higher dimensional feature space, linear relations may exist among the data points. Then, linear Granger causality analysis can be applied on the new higher dimensional feature space [17]. Furthermore, a correntropy-based partial directed coherence(PDC) called Kernel PDC (KPDC) [18] combines the concepts of kernels and partial directed coherence to estimate Granger causality in multi-model systems. Coherence is the measure of linear relation among variables in the frequency domain similar to correlation in time domain. For a multivariate case, partial coherence is used in place of coherence. Both coherence and partial coherence are symmetric measures and do not provide any directionality information. Directed coherence (DC)[10] and partial directed coherence (PDC), split coherence and partial coherence terms into feed-forward and feedback directed influences respectively. However, PDC can be used only in the case when a linear relationship exists among the variables. Therefore, transforming the data into a higher dimensional space where the variables are linearly associated and subsequent calculation of PDC helps to calculate Granger causality for multi-model systems. Calculation of PDC needs the determination of coefficients of the VAR model in the higher dimension space. The VAR model coefficients cannot be determined directly as non-linear transformation is not known explicitly. Thangirala and Kannan [18] used the concept of correntropy to estimate the coefficients of the VAR model. Correntropy is a similarity measure (correlation) in the higher dimension feature space and is defined as the expectation of the inner product of the vectors in the higher dimension space. In KPDC, the covariance between the variables at different time lags

which is given by correntropy is used to calculate the coefficients of the VAR model. There exist many other kernel based approaches to compute multi-model Granger causality. Unlike KPDC which uses a Gaussian kernel, the nonlinear Granger causality technique proposed by Ancona et.al. [19] uses a specific type of kernel functions called radial basis functions. A more generalized approach for multi-model Granger causality which does not place restrictions on the type of kernel function was developed later [20]. This method uses variance operator which gives the variance of the variables in the higher dimensional feature space. A significant reduction in the variance indicates the presence of Granger causal relation.

The main drawback of the entropy-based methods is that their application to multivariate systems is computationally expensive. The entropy based methods are practically applied mainly to bivariate systems. For instance, in transfer entropy which is a popular information-theoretic based method, conditional probability density functions (PDFs) need to be determined. For a multivariate system, large amounts of data need to be used for accurate calculation of the joint PDFs; furthermore the dimensionality of the PDF increases drastically when the dimensionality of the system increases. This leads to an increase in the computational load. For kernel based methods, the computational load is a function of number of training instances. When the number of training data increases the computational load increases significantly. This makes the implementation of kernel methods on real industrial processes with a huge amount of process data difficult. Furthermore, kernel based methods involve the transformation into another higher dimensional feature space, this is not required in the proposed method. Unlike these multi-model Granger causality methods, the proposed method also could be more easily applied on large datasets of multivariate systems.

Auto-regressive (AR) models have been widely used for identifying causal connections among the variables. In multivariate systems, the AR model is replaced with a vector auto regressive (VAR) model[1], [21]. The coefficients of the VAR model

indicate the presence or absence of causal relations among the variables. Significance testing of the coefficients gives the relevance of the causal relations. When the system is multi-model, a single linear VAR model cannot capture the complete system dynamics. In such cases, a VAR model which switches among different operating points is used. To the best of our knowledge, a switched VAR model has not been used in the study of Granger causality analysis. In this work, a switched VAR model will be used to infer Granger causality in multi-model systems.

Parameter estimation of the switched VAR model can be done in a number of ways such as maximum likelihood estimation (MLE) and maximum a posteriori (MAP) etsimation. However, these methods only give a point estimate of the parameters. In most of the real processes where data is corrupted corrupted by noise, estimation of uncertainty of the parameters is also required. The uncertainty associated with the estimates can be quantified by adopting a Bayesian approach where parameters are considered to be random variables. The Bayesian approach adopted in this thesis involves the graphical representation of the data-driven models using probabilistic graphical models. Bayesian approach estimates the posterior distributions for the parameters instead of their point estimates. Bayesian approach has the added advantage over the point estimation approaches that process knowledge can be incorporated in the form of prior distributions of unknown model parameters and latent variables. The exact determination of posterior distribution in the Bayesian approach using the Bayes's rule is generally difficult as some of the integrals are intractable. Monte Carlo based sampling methods can be used to approximate such integrals and subsequently estimate the posterior distribution. An alternate approach for determining posterior distribution of parameters is variational inference [22], [23].

Variational inference involves approximating the true posterior distribution as a distribution which is easier to handle. A VAR model tends to overfit a model as it contains many model parameters. Hence, different model structures with different order and lag combinations need to be estimated and validated to determine the

best VAR model structure. The variational Bayesian approach helps in overcoming this difficulty by regularizing the model structure and thereby removing the insignificant model parameters. Furthermore, the variational Bayesian approach reduces the computational load as it determines only an approximate distribution instead of the actual posterior distribution.

Causality studies have already been performed using the variational Bayesian approach [24]. In this work, we attempt to estimate the Granger causality among variables in a multivariate multi-model system using variational Bayesian parameter estimation technique through a switched VAR model. The main advantage of the variational Bayesian approach used in this thesis over other nonlinear methods is that it is able to maintain constant causal structure across different modes. It achieves it by placing a soft constraint using a Normal-Gamma prior on the corresponding elements of the coefficient matrices from all the VAR models. The statistical tests involved in the traditional Granger causality techniques for a multivariate system like conditional Granger causality are tedious as it uses a series of F-tests. For which, the proposed method provides an alternative way of defining a single metric for evaluating the significance of the causal connections.

The rest of the chapter is organized as follows. Section 3.2 presents the model description in detail. Section 3.3 discusses the model estimation approach. Section 3.4 provides the implementation steps and section 3.5 presents the simulation case study. In section3.6, we present the concluding remarks.

3.2 Model description

3.2.1 VAR model

For determining the Granger causal connections in a multivariate system $y(t) \in R^D$, the linear vector auto regressive (VAR) model is constructed as follows,

$$y(t) = \sum_{l=1}^{L} W(l)y(t-l) + e(t) \tag{3.1}$$

where l represents the lag and it varies from 1 to maximum lag L, t represents the time instants and it ranges from 1 to N and $y(t)$ is measured variable at time t. $W(l) \in R^{D\times D}$ is the coefficient matrix of the VAR model at lag l and $e(t) \in R^D$ is the noise associated with the process which is assumed to follow a Gaussian distribution.

3.2.2 Mixture VAR model

Most of the industrial processes operate in more than one mode, due to changes in operating conditions such as changing feed rates, varying production targets, varying catalyst conditions etc. In such cases more than one linear VAR model need to be used to model the actual process. Switched linear VAR model assumes the following structure,

$$y(t) = \sum_{l=1}^{L} W(l)^s y(t-l) + e^s(t) \tag{3.2}$$

where $W(l)^s \in R^{D\times D}$ is the regression parameters for the s^{th} local model at lag l which is given as the following,

$$W(l)^s = \begin{bmatrix} w(l)^s_{11} & \dots & w(l)^s_{1D} \\ \vdots & \ddots & \vdots \\ w(l)^s_{D1} & \dots & w(l)^s_{DD} \end{bmatrix} \tag{3.3}$$

where the values of s and sampling time instant t range from 1 to ub and 1 to N respectively. The values of the elements of $W(l)^s$ give indication of causal relation between a particular input in the regression vector and the corresponding output variable. For instance $w(l)^s_{ij} = 0$ implies that there exists no causal relation between

the j^{th} input in the regression vector and i^{th} output for s^{th} local model at lag l. Each linear VAR model is represented by a probability distribution (typically multivariate Gaussian distribution) and the entire multi-mode process is expressed as a weighted sum of the multi-variate normal distributions. Such a probabilistic model is called a mixture vector auto regressive model (MVAR) [25]. The formulation of a typical MVAR model for time series $y(t) \in R^D$ is given as follows,

$$p(y(t)|Y^{t-1}) = \sum_{s=1}^{ub} \alpha_{s,t} 2\pi^{-D/2} \det((\delta^s)^{-1} I_D)^{-1/2} \times \exp\{-\frac{1}{2}(y(t) - \mu_{s,t})^T ((\delta^s)^{-1} I_D)^{-1} (y(t) - \mu_{s,t})\} \quad (3.4)$$

where $p(y(t)|Y^{t-1})$ represents the conditional distribution of y(t) given all its past values represented as Y^{t-1}. $\alpha_{s,t}$ represents the mixing weights and it satisfies $\sum_{s=1}^{ub} \alpha_{s,t} = 1$ at any time t. $(\delta^s)^{-1} I_D$ is the noise covariance, where δ^s is a scalar quantity and I_D is the identity matrix of dimension D. The mean, μ_s, is the conditional mean given as the following,

$$\mu_s = \mu_{0,s} + \sum_{l=1}^{L} W(l)^s y(t-l) \quad \text{where } s = 1, 2, ..., ub \quad (3.5)$$

The unknown parameters set of the s^{th} local model of the MVAR model is $\Phi^s = W^s, \Sigma^s$ where $W^s = (W(1)^s, W(2)^s, ...W(L)^s)$. There are different choice of weights in the literature such as constant or time varying weights [25].

3.2.3 Bayesian Mixture VAR model

In this thesis, a Bayesian approach is used which assumes the unknown parameters and mixing weights as random variables and assigns prior distributions to the variables. A higher order MVAR model may be used without loss of generality. However, for the sake of simplicity, we restrict the illustrations and case studies to first order MVAR model. For a first order model, equation 3.2 simplifies to the following form,

$$y(t) = W^s y(t-1) + e^s(t) \quad (3.6)$$

where, $W^s(orW(1)^s) \in R^{D\times D}$ are the regression parameters for the s^{th} local model. To distinguish the rows and columns of W^s, in our derivation, the number of columns is indicated as M even though D is equal to M in our case i.e. $W^s(orW(1)^s) \in R^{D\times M}$. W^s is given as the following,

$$W^s = \begin{bmatrix} w^s_{11} & \dots & w^s_{1M} \\ \vdots & \ddots & \vdots \\ w^s_{D1} & \dots & w^s_{DM} \end{bmatrix} \tag{3.7}$$

In this work, the causal structure is assumed to be same across all the operating modes. In terms of the switched VAR model considered, it would mean that the coefficient matrices $W^1, W^2, ..., W^{ub}$ will have zero and non-zero elements at the same locations. To achieve the same sparse structure for the coefficients matrices across modes, we assign a single joint prior on the corresponding elements of the coefficient matrices from all the operating modes. A sparsity enforcing prior like a Normal-Gamma prior is used to represent the joint prior. For instance, the prior distributions of the dm^{th} element of the coefficient matrices from all the modes are assumed to follow a Normal-Gamma distribution as the following,

$$\begin{gathered} [w^1_{dm}, w^2_{dm}, ..., w^{ub}_{dm}] = W_{dm} \sim \mathcal{N}(0, \beta^{-1}_{dm} I) \\ \text{where } \beta_{dm} \sim \Gamma(a^*, b^*) \end{gathered} \tag{3.8}$$

where W_{dm} is the set of the dm^{th} elements from all the coefficient matrices, a^* and b^* are the shape parameter and rate parameter respectively. β_{dm} is the precision parameter of the normal distribution. While I is identity matrix in general, here it is just one. The priors of the MVAR coefficients are assumed to have Gaussian distributions with zero mean. The precision parameter (inverse of variance) of the normal distribution is assumed to follow a gamma distribution. Adopting a Normal-Gamma prior for the coefficients will introduce a penalty on the lower valued coefficients which will introduce a sparsity in the coefficient matrix by reducing the insignificant coefficients to zero. Now, since the similar positioned elements across coefficient matrices have the same Normal-Gamma prior they will be simultaneously all zero or all non-zero, this

allows for a consistent causal structure across all modes. Thus, the Normal-Gamma prior will act as a soft constraint which ensures that the causal structure across the operating modes remains the same. Under the above assumption, the prior distribution of the set of coefficient matrices $W = [W^1, W^2, ..., W^{ub}]$ given the set of precision parameters $\beta = [\beta_{11}, \beta_{12}, ..., \beta_{DM}]$ can be factorized as the following,

$$p(W|\beta) = \prod_{d=1}^{D} \prod_{m=1}^{M} p(W_{dm}|\beta_{dm}) = \prod_{d=1}^{D} \prod_{m=1}^{M} \mathcal{N}(W_{dm}|0, \beta_{dm}^{-1}) \tag{3.9}$$

where W_{dm} represents the dm^{th} element of all the coefficient matrices in set W and its precision parameter is β_{dm}. The precision parameter β_{dm} follows a gamma distribution with a^* and b^* as shape and rate parameters respectively. Each regression parameter from all the local models is considered to follow Gaussian distribution with zero mean and a certain value (β_{dm}) of precision. Then, the joint distribution of β which is the set of all the precision parameters $[\beta_{11}, \beta_{1,2}, ..., \beta_{DM}]$ can be factorized as the following,

$$p(\beta|a^*, b^*) = \prod_{d=1}^{D} \prod_{m=1}^{M} p(\beta_{dm}) = \prod_{d=1}^{D} \prod_{m=1}^{M} \Gamma(\beta_{dm}|a^*, b^*) \tag{3.10}$$

where a^* and b^* are shape and rate parameters of the prior gamma distributions.

Measurement noise of the s^{th} local model is considered to be a normal distribution variable with zero mean and precision δ^s. The noise precision δ^s is considered as an unknown constant. Then, the probability that the observed data $Y = [y(1), y(2), ..., y(N)]$ is generated by the s^{th} local model is given as the following,

$$p(Y|W^s, \delta^s, S = s) = \prod_{t=1}^{N} p(y(t)|W^s, \delta^s, S = s) = \prod_{t=1}^{N} \mathcal{N}(y(t)|0, \delta^{s^{-1}} I_D) \tag{3.11}$$

where $p(Y|W^s, \delta^s, S = s)$ is the conditional distribution of Y given coefficient matrix W^s and precision of the noise δ^s for the s^{th} local model. I_D is the identity matrix of dimension D.

Now, we move onto the prior distributions of the latent variables. The model identity S is the hidden variable in the proposed model. The prior distribution of

model identity S being s can be factorized as the following,

$$p(S = s) = \sum_{t=1}^{N} p(S(t) = s) = \sum_{t=1}^{N} \alpha^{S(t)=s} \tag{3.12}$$

where S is the model indicator random variable which takes value s that ranges from 1 to ub. The prior probability of the model indicator variable S taking the value s is derived by taking the sum of the prior probabilities of local model s from time instants 1 to N. Finally, significant coefficients are assigned to each of the local models. The significance coefficient α^s of local model s is derived as the following,

$$\alpha^s = \frac{\sum_{t=1}^{N} \alpha^{S(t)=s}}{N} \tag{3.13}$$

The initial significance coefficient of local model s being assigned as α^s is the average of prior probabilities of model indicator variable $S = s$ over N time instants.

Furthermore, a symmetric Dirichlet distribution prior is assigned to the set of significant coefficients $\alpha = \left[\alpha^1, \alpha^2, ..., \alpha^{ub}\right]$ which is given as the following,

$$p(\alpha) = Dir(\alpha|\alpha^* m^*), m^* = \left[\frac{1}{ub}, ..., \frac{1}{ub}\right] \tag{3.14}$$

where α^1, α^2 and α^{ub} are significant coefficients of local models $1, 2$ and ub respectively. α^* is the only hyperparameter of the Dirichlet distribution and is a scalar value. Since we assume the Dirichlet distribution to be symmetric, a parameter vector $\alpha^* m^*$ is used such that all the elements of scale vector m^* are equal.

The Bayesian network (BN) for the model described above is given in Fig 3.1. The nodes inside circles are random variables, while the others are deterministic. The nodes enclosed within the squares repeat themselves by the number given in its lower left corner. The BN considered belongs to a special subclass of BN called as conjugate exponential family graphical model (CEFGM). In CEFGM, the prior distributions of all the random variable nodes belong to the exponential family of distributions. Under the independence assumption in the VB approach (equation 3.15), the prior distributions are conjugate to their likelihoods; then prior and posterior belong to the same family of distributions.

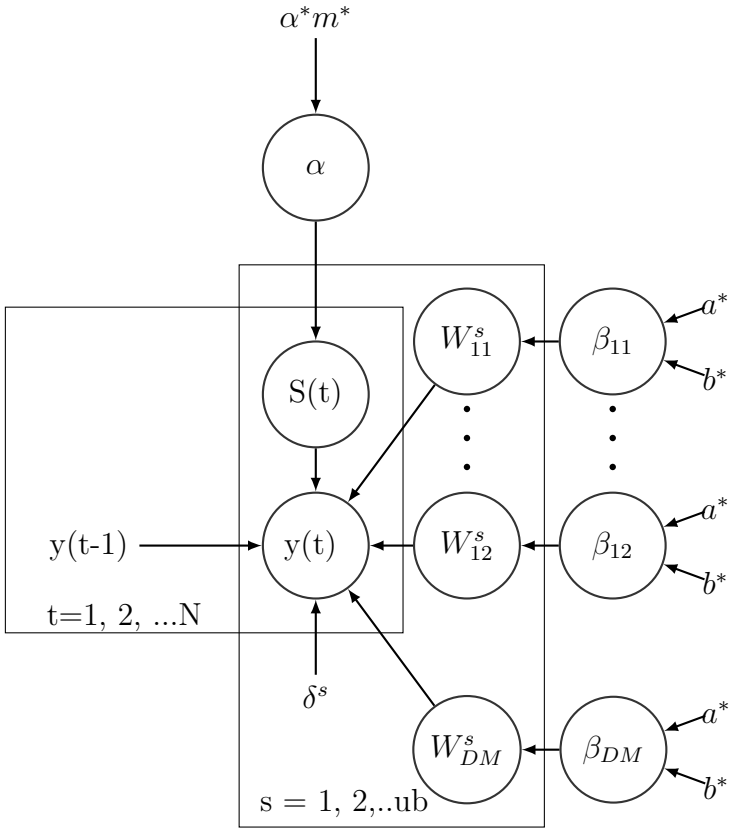

Figure 3.1: Bayesian Network for the proposed model

3.3 Model Estimation

Variational Bayesian Expectation Maximization (VBEM) Theory

The set of unknown parameters of the model are represented as $\Phi = [\Phi^1, \Phi^2, ..., \Phi^{ub}]$ and it includes the set of coefficient matrices of all the local models, $W = [W^1, W^2, ..., W^{ub}]$, set of precision of the local model parameters $\beta = [\beta_{11}, \beta_{1,2}, ..., \beta_{dm}]$, and set of model significance coefficients $\alpha = \left[\alpha^1, \alpha^2, ..., \alpha^{ub}\right]$. The latent variable is the model identity S. The joint posterior distribution of ϕ and S using the Bayes rule is given as the following,

$$\begin{aligned} p(\phi, S|Y, M) =& \frac{p(Y|\Phi, S, M)p(\Phi, S|M)}{p(Y|M)} \\ =& \frac{p(Y|\Phi, S, M)p(\Phi, S|M)}{\sum_S \int_\Phi p(Y|\Phi)p(\Phi, S|M))} \end{aligned}$$

where $p(\phi, S|Y, M)$ represents the posterior distribution of unknown parameters (Φ) and latent variable (S) given the data Y and the model structure M. Switched VAR model is considered as the model structure. $p(\Phi, S|M)$ represents the joint prior distribution of the unknown parameters and latent variables, $p(Y|\Phi, S, M)$ represents the likelihood of the data Y and $p(Y|M)$ is the model evidence which is the likelihood that Y is generated by model structure M. Additionally, we are also interested in obtaining the model evidence given a model structure as it helps in determining the best model for a given data. However,the actual posterior distribution is often difficult to determine as the calculation of model evidence is often intractable. Hence, the variational Bayesian expectation maximization (VBEM) approach is used as it helps to determine the posterior distribution and model evidence in an approximate manner [22], [23]. The variational Bayesian approach approximates the actual posterior distribution $p(\phi, S|Y, M)$ as the following,

$$p(\phi, S|Y, M) \sim q(\Phi)q(S) \tag{3.15}$$

where the approximate posterior distributions of unknown parameters set Φ and latent variable S are given as $q(\phi)$ and $q(S)$ respectively. This approximation is made

for tractability. The log of model evidence can be expanded as follows through some mathematical manipulations,

$$\begin{aligned}\ln p(Y|M) =& \sum_S \int_\Phi q(\Phi)q(S) \ln \frac{p(Y,S,\Phi|M)}{q(\Phi)q(S)} d\Phi \\ &+ \sum_S \int_\Phi q(\Phi)q(S) \ln \frac{q(\Phi)q(S)}{p(\Phi,S|Y,M)} d\Phi \\ =& L(q(S),q(\Phi)) + KL(q(\Phi)q(S)||p(\Phi,S|Y,M))\end{aligned} \tag{3.16}$$

where the second term is the KL divergence between approximated posterior distributions ($q(\Phi)q(S)$) and actual posterior distribution ($p(Y,S,\Phi|M)$). It is always a positive quantity and hence the first term $L(q(S),q(\Phi))$ lower bounds the log of model evidence. KL divergence approaches zero when the approximated posterior distribution becomes equal to the actual posterior distribution. Equation 3.16 shows that the minimization of KL divergence can be interpreted as the maximization of the lower bound of the log model evidence.Thus, to find the most appropriate approximated posterior distribution, instead of minimizing KL divergence, the lower bound $L(q(S),q(\Phi))$ is maximized.

3.3.1 Lower bound expression

Further, the approximated posterior distribution of the unknown parameters set Φ is further factorized as follows,

$$q(\phi) = q(W|S)q(\beta)q(\alpha) \tag{3.17}$$

where $W = [W^1, W^2, ..., W^{ub}]$, $\beta = [\beta_{11}, \beta_{1,2}, ..., \beta_{dm}]$, and $\alpha = \left[\alpha^1, \alpha^2, ..., \alpha^{ub}\right]$ and S is the model identity. The initial step in VBEM is to assume the structure for the approximated posterior distributions for the unknown parameters. The D-separation principle helps to factorize the approximated posterior distributions further. The rows of the regression parameter matrix W^s of a local model s are independent given Y since each row of W^s is a parent to only a particular dimension of Y and no two rows of W^s share a common child. Then, $q(W|S=s)$ which is the approximate posterior

distribution of coefficient matrix of the s^{th} local model and W^s can be expressed as the product of distribution of rows of W^s as the following,

$$q(W|S=s) = q(W^s) = \prod_{d=1}^{D} q(W_d^s|\hat{W}_d^s, \Sigma_{\hat{W}_d^s}) = \prod_{d=1}^{D} \mathcal{N}(W_d^s|\hat{W}_d^s, \Sigma_{\hat{W}_d^s}) \tag{3.18}$$

where W_d^s represents the d^{th} row of W^s of local model s. It is assumed to be a Gaussian distribution with mean and covariance being equal to $\hat{W}_d^s$ and $\Sigma_{\hat{W}_d^s}$ respectively. Similarly, $q(\beta)$ can be expanded as the following,

$$q(\beta) = \prod_{d=1}^{D}\prod_{m=1}^{M} q(\beta_{dm}|a, b_{dm}) = \prod_{d=1}^{D}\prod_{m=1}^{M} \Gamma(\beta_{dm}|a, b_{dm}) \tag{3.19}$$

where a and b_{dm} represent the shape and rate parameters respectively. β_{dm} is the parent of w_{dm}^s. Additionally it does not share its child with any other $\beta_{ab\neq dm}$ which implies that β_{ab} is independent to β_{dm}. The noise precision of the s^{th} local model, δ^s, is assumed to be an unknown constant. However, we could also define a distributions for δ^s and further define distribution for its hyperparameters and so on.

The approximated posterior distribution of latent variable $S = s$ is as the following,

$$q(S=s) = \sum_{t=1}^{N} q(S(t)=s) = \sum_{t=1}^{N} \alpha_{new}^{S(t)=s} \tag{3.20}$$

where $\alpha_{new}^{S(t)=s}$ represents the approximate posterior probability of local model s at time instant t. It should satisfy the following constraint,

$$\sum_{s=1}^{ub} q(S(t)=s) = \sum_{s=1}^{ub} \alpha_{new}^{S(t)=s} = 1 \tag{3.21}$$

since we are assuming the data at any time instant t to be generated by any one of local models considered. This in turn implies the following,

$$\sum_{t=1}^{N}\sum_{s=1}^{ub} q(S(t)=s) = \sum_{t=1}^{N}\sum_{s=1}^{ub} \alpha_{new}^{S(t)=s} = N \tag{3.22}$$

The expression for the significance coefficient for the s^{th} local model, α^s, is derived as the following,

$$\alpha^s = \frac{\sum_{t=1}^{N} \alpha_{new}^{S(t)=s}}{N} \tag{3.23}$$

Thus, the updated significance coefficient of the s^{th} local model is obtained by taking the average of the approximate posterior probability of model indicator variable s ($\alpha_{new}^{S(t)=s}$) over N time instants. The approximated posterior distribution of the set of significance coefficients of models, α, is considered to be independent to each other and follows a dirichlet distribution as the following,

$$q(\alpha) = \prod_{s=1}^{ub} q(\alpha^s) = \prod_{s=1}^{ub} q(\alpha^S|\alpha_{new}^* m_S) = \prod_{s=1}^{ub} Dir(\alpha^S|\alpha_{new}^* m_S) \tag{3.24}$$

where $\alpha_{new}^* m_s$ is the parameter of the approximate posterior Dirichlet distribution of significance coefficient α^s.

The lower bound is given as follows,

$$L(q(S), q(\phi)) = \sum_S \int_{W,\beta,\alpha} q(W|S)q(\beta)q(\alpha)q(S) \ln \frac{p(Y, W, \beta, \alpha, S|\delta, \alpha^*, a^*, b^*)}{q(W|S)q(\beta)q(S)} \tag{3.25}$$

The above equation can be expanded further into the following form,

$$\begin{aligned} L(q(S), q(\phi)) = & \sum_{s=1}^{ub} \int_\beta q(\beta) \sum_{d=1}^{D} \int_W q(W_d|S=s) \ln \frac{p(W_d^s|\beta_d)}{q(W_d|S=s)} \\ & + \sum_{d=1}^{D} \sum_{m=1}^{M} \int_{\beta_{dm}} q(\beta_{dm}|a, b_{dm}) \ln \frac{p(\beta_{dm}|a^*, b^*)}{q(\beta_{dm}|a, b_{dm})} \\ & + \sum_{s=1}^{ub} q(\alpha^s) \ln \frac{p(\alpha^s|\alpha^*)}{q(\alpha^s)} \\ & + \sum_{t=1}^{N} \sum_{s=1}^{ub} q(S(t)=s) \int q(\alpha^s) \ln \frac{p(S|\alpha^s)}{q(S(t)=s)} \\ & + \sum_{s=1}^{ub} \int q(W|S=s)q(S=s) \ln p(Y|W, \delta, S=s) \end{aligned} \tag{3.26}$$

The first term of the expression can be expanded as the following,

$$\sum_{s=1}^{ub}\int_{\beta} q(\beta)\sum_{d=1}^{D}\int_{W} q(W_d|S=s)\ln\frac{p(W_d^s|\beta_d)}{q(W_d|S=s)} \tag{3.27}$$

$$=\sum_{S=1}^{ub}\int q(\beta)[-\sum_{d=1}^{D} KL(q(W_d^s|\hat{W}_d^s,\Sigma_{\hat{W}_d^s})||p(W_d|[0]_{1\times M},(diag([\beta_{d1},...,\beta_{dM}]^T))^{-1}))] \tag{3.28}$$

$$=\frac{1}{2}\sum_{s=1}^{ub}\sum_{d=1}^{D}\ln|\Sigma_{\hat{W_d}^s}|+\frac{ub}{2}\sum_{d=1}^{D}\sum_{m=1}^{M}(\psi(a)-\ln b_{dm})+\frac{ubDM}{2} \tag{3.29}$$

$$-\frac{1}{2}\sum_{s=1}^{ub}\sum_{d=1}^{D} tr[\lambda_d(\Sigma_{\hat{W_d}^s}+(\hat{W}_d^{sT}\hat{W}_d^s)]$$

where

$$\lambda_d = diag\left([\frac{a}{b_{d1}},...,\frac{a}{b_{dM}}]\right) \tag{3.30}$$

In CEFGMs, an explicit expression for the lower bound can be derived. The detailed expression for the lower bound is given in table A.1 of the appendix.

3.3.2 Posterior updates

In the VBEM algorithm, the lower bound is maximized iteratively through an expectation step or E-step and a maximization step or M-step such that with each iteration the approximated posterior distributions of the unknown parameters and latent variables are updated. Substituting these updates in the lower bound expression will maximize the lower bound iteratively. The VB E-step involves maximizing the lower bound with respect to $q(S)$ while keeping $q(\phi)$ constant. In the VB M-step, the lower bound is maximized with respect to $q(\phi)$ while keeping the distribution of hidden variable $q(S)$ constant. For instance, the derivative of lower bound with respect to

$q(W|S=s)$ is given as the following,

$$\frac{\partial L}{\partial q(W|S=s)} = \int q(\beta)\ln p(W|\beta)d\beta - \int q(\beta)\ln q(W|S=s)d\beta - 1 \tag{3.31}$$
$$+\sum_{t=1}^{N} q(S(t)=s)\ln p(y(t)|W^s, S(t)=s, \delta^s, \alpha) = 0$$

$$\implies \Sigma_{\hat{W}_d{}^s} = [\lambda_d + \delta^s \sum_{t=1}^{N} \alpha_{new}^{S(t)=s} y(t-1)y(t-1)^T]^{-1} \tag{3.32}$$

$$(\hat{W}_d{}^s)^T = \Sigma_{\hat{W}_d{}^s}[\sum_{t=1}^{N} \alpha_{new}^{S(t)=s}\delta^s y_d(t)y(t-1)] \tag{3.33}$$

The d^{th} row of coefficient matrix W^s follows a multivariate Gaussian distribution of covariance $\Sigma_{\hat{W}_d{}^s}$ and mean $\hat{W}_d{}^s$. $y_d(t)$ is the d^{th} component of $y(t)$. The remaining updated approximated posterior distributions are given in appendix A.2. The maximization involves taking the derivative of the lower bound with respect to the particular approximated posterior distribution and then equating it to zero. The point estimate of noise precision is obtained by taking the derivative of the lower bound with respect to the precision variable and equating it to zero.

3.3.3 Hyperparameter selection

Rather than assuming random values for the hyperparameters, it is better to infer the hyperparameter values from the given data. One of the methods of hyperparameter selection is cross validation. In cross validation, the data is divided into training and validation sets. The model parameters are identified using the training set for different hyperparameter values and the model is validated using the validation set. The hyperparameters which give the best validation performance is retained. Log likelihood of the model parameters in the validation data set is used as the validation criterion in this thesis and it is as follows,

$$\sum_{t=1}^{N_{val}} \log \sum_{s=1}^{ub} \hat{\alpha}^s \mathcal{N}(y(t)|\hat{W}^s y(t-1), \hat{\delta}^{s-1} I_D) \tag{3.34}$$

where N_{val} is the number of data points in the validation set, $\hat{W}^s$ and $\hat{\delta}^s$ are derived from the update equations and the mixing coefficients $\hat{\alpha}^s$ are the initial guess values of

weights α^s. Bayesian optimization using 'Bayesopt' function in MATLAB is used to obtain the optimal hyperparameter values. Since 'Bayesopt' performs minimization of the objective function, the negative of the log likelihood function is passed as the objective function. The significance of hyperparameter selection becomes clear from the update equation given in 3.33. λ_d adds a penalty to each column of the coefficient matrix W. For instance, the penalty added to the m^{th} column is given as the following,

$$\begin{aligned}\frac{a}{b_{dm}} =& \frac{a^* + \frac{ub}{2}}{b^* + \frac{1}{2}\sum_{s=1}^{ub}\left[\hat{W}_{dm}^{s\ 2} + \Sigma_{\hat{W}_{dm}^s}\right]} \\ =& \frac{a^* + \frac{ub}{2}}{b^* + \frac{1}{2}\sum_{s=1}^{ub} E((W_{dm}^s)^2)}\end{aligned} \tag{3.35}$$

where $E((W_{dm}^s)^2)$ is the posterior expectation of $(W_{dm}^s)^2$ for local model s. Now, it is observed that for a fixed value of b^*, a decrease in a^* value imposes heavier penalty on lower valued coefficients and lower penalty on higher valued coefficients (Fig 3.2). Hence, the rate factor a^* is the only parameter in this work which is estimated using Bayesian optimization. The best value of a^* is chosen to be between 10^{-8} and 10^8 and b^* is fixed at 10^{-8}.

3.4 Implementation of causality analysis

The implementation of the VBEM approach for the causality analysis is explained in this section. It involves the maximization of lower bound iteratively. After each iteration, the parameters of the approximated posterior distributions and noise precision set δ are updated. The iteration is stopped once the value of lower bound increases only by a negligible value. The detailed implementation steps are given in table 3.1. The implementation of the VBEM approach for the causality analysis is explained in this section. It involves the maximization of lower bound iteratively. After each iteration, the parameters of the approximated posterior distributions and noise precision set δ are updated. The iteration is stopped once the value of lower bound increases

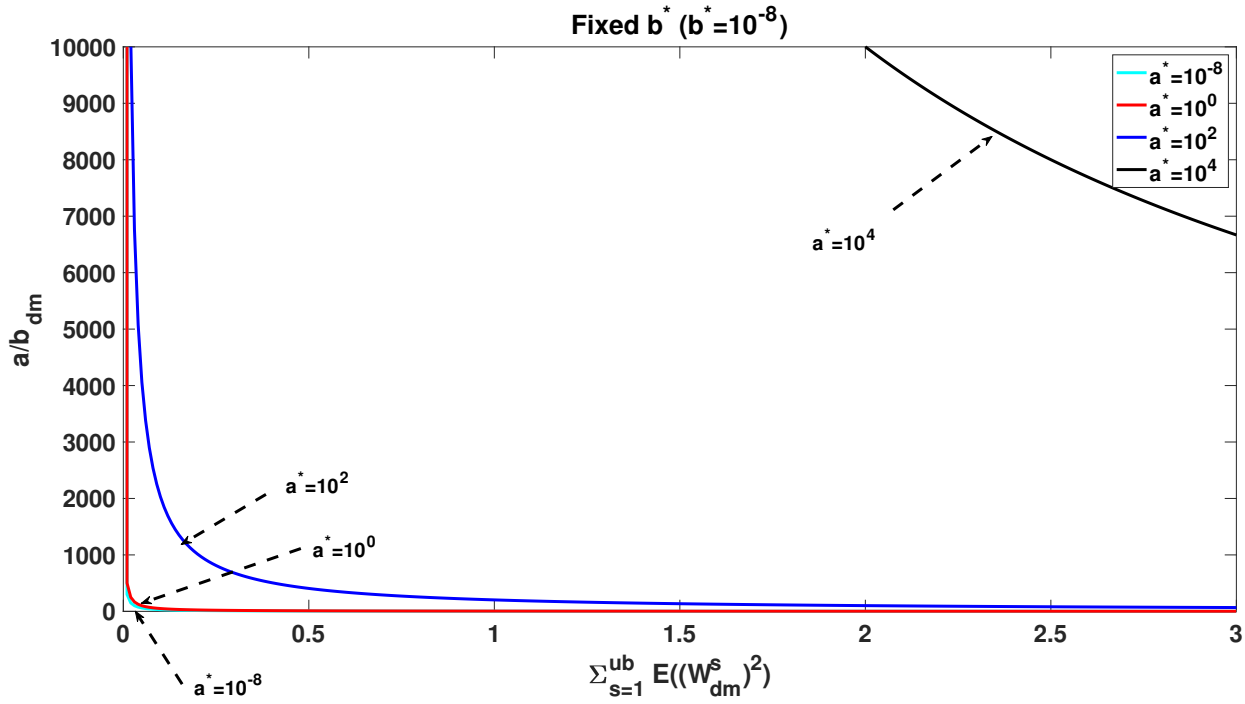

Figure 3.2: The graph showing the effect of decrease in a^* values on the penalty for a fixed b^*. As a^* decreases, heavier penalty is imposed on lower valued coefficients of the switched VAR model.

only by a negligible value. The detailed implementation steps are given in table 3.1. Once W has been determined, relevance of each element in W can be checked using the estimate of its precision parameter β. The posterior distribution of the precision parameter of the dm^{th} element of the coefficient matrices $W^1, W^2, ..., W^{ub}$ is a gamma distribution and its expected value is given as the following,

$$\beta_{dm} = \frac{a}{b_{dm}} \tag{3.36}$$

which is same as equation (3.35) where a and b_{dm} are the shape and rate parameters of the approximated posterior gamma distribution of precision parameter β_{dm}. Higher expected value (β_{dm}) implies that the sum of the posterior expectation of $(W^s_{dm})^2$ from all the local models ($\sum_S E((W^s_{dm})^2)$) is close to zero. This implies that values of the dm^{th} element of the coefficient matrix are close to zero in all the local models. Thus, by setting a lower threshold value on the inverse of the expected value of precision parameter, the relevant parameters in the coefficient matrix set W can be

Table 3.1: Implementation steps

Steps	
1	Fix values for MaxIter, threshold ϵ and ub
2	Perform Bayesian optimization to determine a suitable a^* value
3	Assign initial guess values for the remaining hyperparameters of the priors and the approximate posterior distributions
4	Compute lower bound $L(k)$ where k=1, using the initial guess values given in step 3
5	Begin For loop. For idx=1:MaxIter
6	Take first derivative of lower bound with respect to $q(\Phi)$ keeping $q(S)$ constant and equate it to zero
7	Update the parameters of $q(\phi)$
8	Take first derivative of lower bound with respect to $q(S)$ keeping $q(\Phi)$ constant and equate it to zero
9	Update the parameters of $q(S)$
10	If reminder of (idx/10)=0 then k=k+1
11	Recompute the lower bound L(k), using the previously updated parameters
12	If $\|(L(k) - L(k-1)\|/\|L(k-1)\| \leq \epsilon$
13	Break For loop
14	Else
15	End both of the If loops
16	update idx=idx+1 and repeat steps from 6 to 15
17	End For

differentiated from the irrelevant ones.

3.5 Simulation case study

Simulation case study is used to verify the VBEM approach for identifying causal relations in multi-mode systems. The data of a multi-mode system with ub number of local models is generated as the following,

$$y(t) = W^s y(t-1) + e^s(t), t \in [1, 2, ..., N], s \in [1, 2, ..., ub] \tag{3.37}$$

where W^s is the coefficient matrix for the s^{th} local model and N is the number of samples.

For our simulation example, we set $N = 3000$ and $ub = 3$. 1000 data points were generated for each local model. The causal connections in all three models were considered to be the same, but the strengths of the causal connections were allowed to be different in each of the modes. The true coefficient matrix representing the causal relations for one of the simulation example is below,

$$W^1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.3 & 0 \\ -0.9 & 0 & 0 \end{bmatrix}, W^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0.9 & 0 & 0 \end{bmatrix}, W^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -0.9 & 0 \\ 0.9 & 0 & 0 \end{bmatrix}$$

The additive noise e^s was generated from a normal distribution with mean 0 and variance 1. The proposed method was able to identify causal structure in each mode accurately.

Further simulations were done to check the accuracy of the proposed method when changes were made in the number of local models (ub), dimension of data, noise variance, sparsity percentages and b^* values. The accuracy is defined as the number of causal connections identified correctly. The simulation was repeated for 50 different causal connections. The relevance metric defined earlier in section 3.4 was used to check the accuracy of the method. The simulation details are summarized in table 3.2. Switched VAR model considered here is a dynamic system as the values of coefficient matrix W^s changes after 1000 time instants. The elements of the parameter matrix

Table 3.2: Simulation details

Attribute	Value
Number of local models	2 to 5
Number of different causal cases	50
Model order	1
Dimension of input and output	2 to 6
Threshold of the relevance metric	10^{-3}
Total time instants	3000
Sparsity	$\sim U(0.1, 0.4)\& \sim U(0.4, 0.9)$
Parameter, W	$\sim U(-1.9, -1)\& \sim U(1, 1.9)$
Noise mean	0
Noise variance	0.1, 0.3, 0.5, 0.7, 1

W^s, were drawn from the uniform distributions such that the local models are stable. The accuracy of the proposed method for different number of local models is given in Fig 3.3. Thus the method is sensitive to the number of local models considered. As the number of local models ub increases, the accuracy increases.

In the second simulation study, the dimension of the data was changed keeping the number of local models constant (ub=3). Accuracy of the method was checked for 50 different causal connections while dimension was varied from 2 to 6 and the results are shown in Fig 3.4. Overall, there is a decrease in accuracy with increase in dimension. This can probably be attributed to local optima convergence of the model. As the dimension increases, generating initial guesses close to global optima becomes more challenging. There seems to be an exception only for the case when dimension increases from 3 to 4. In the proposed work, different initial guesses were assumed arbitrarily and the accuracy of the algorithm for each of these guesses was evaluated. The arbitrary guess which gave the most accurate result was chosen.

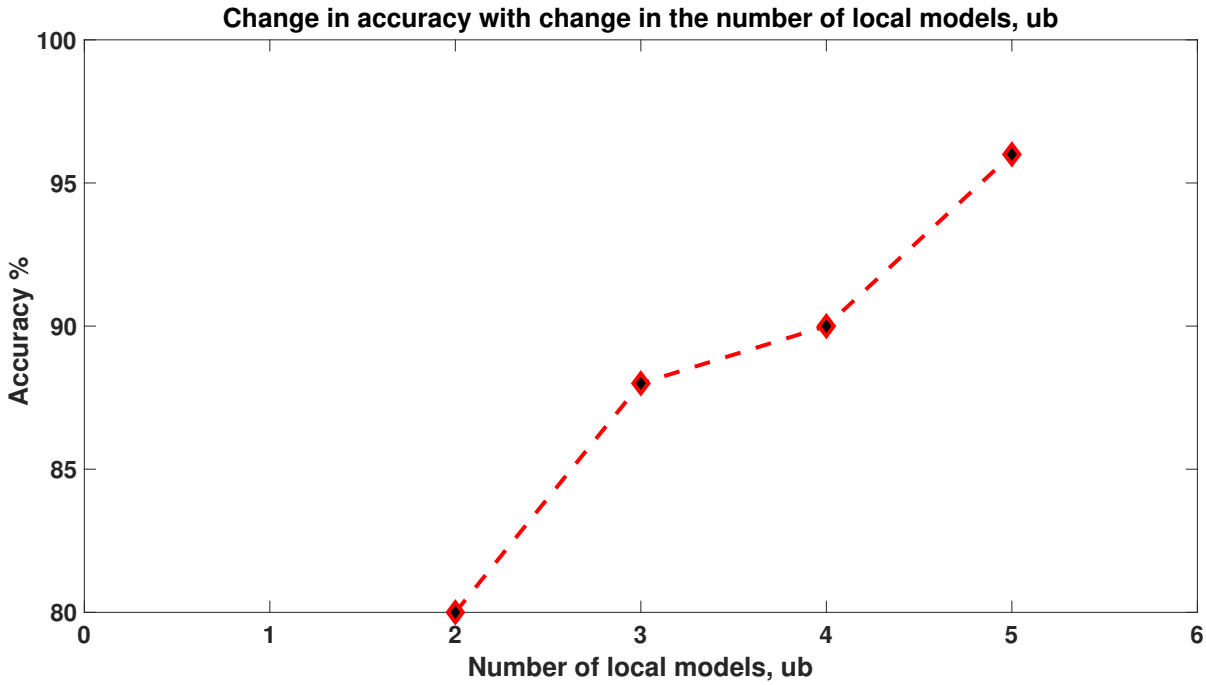

Figure 3.3: Accuracy of the proposed method for different number of local models

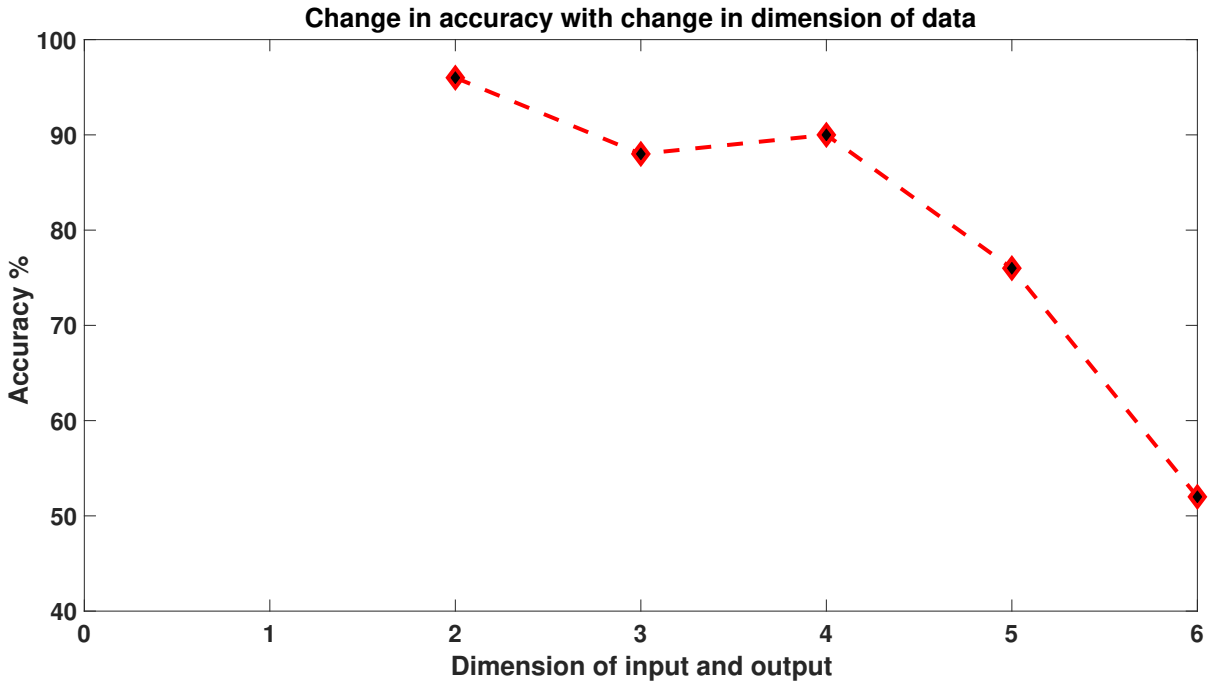

Figure 3.4: Accuracy of the proposed method for different dimensions of data

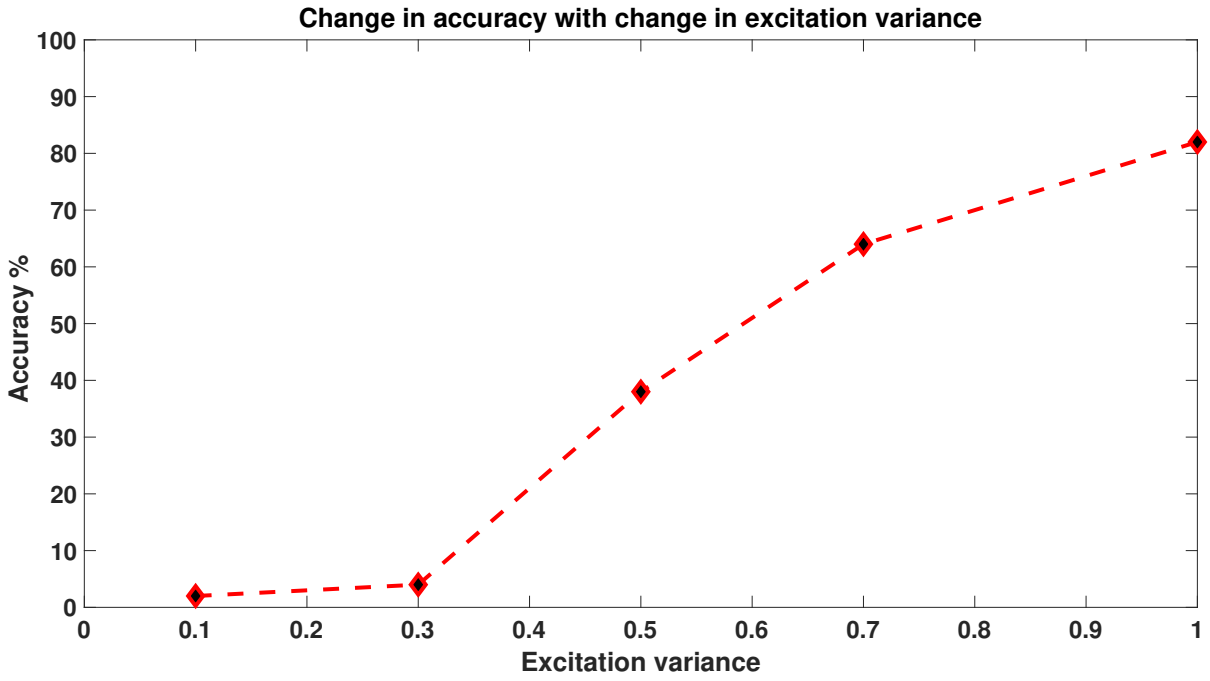

Figure 3.5: Accuracy of proposed method for different noise variance

Table 3.3: Accuracy results for different sparsity of coefficient matrix W

Sparsity	Number of causal cases	Accuracy Percentage
$\sim U(0.4, 0.9)$	15	93.33
$\sim U(0.1, 0.4)$	15	66.67

The noise acts as both disturbance and excitation in the above example. The accuracy of the method increases with increase in the variance of the disturbance (Fig 3.5). The penalty (equation 3.36) added to lower valued coefficients is seen to increase with increase in the noise variance. This increased the accuracy of the method for higher noise variance as it helped to eliminate insignificant causal connections.

The change in sparsity of the coefficient matrix had an effect on the accuracy of the method. Two different cases of sparsity were considered in the work and the results are given in table 3.3. The last set of simulations were done to study the effect of change of b^* values on the accuracy (Fig 3.6). b^* is the shape parameter of the

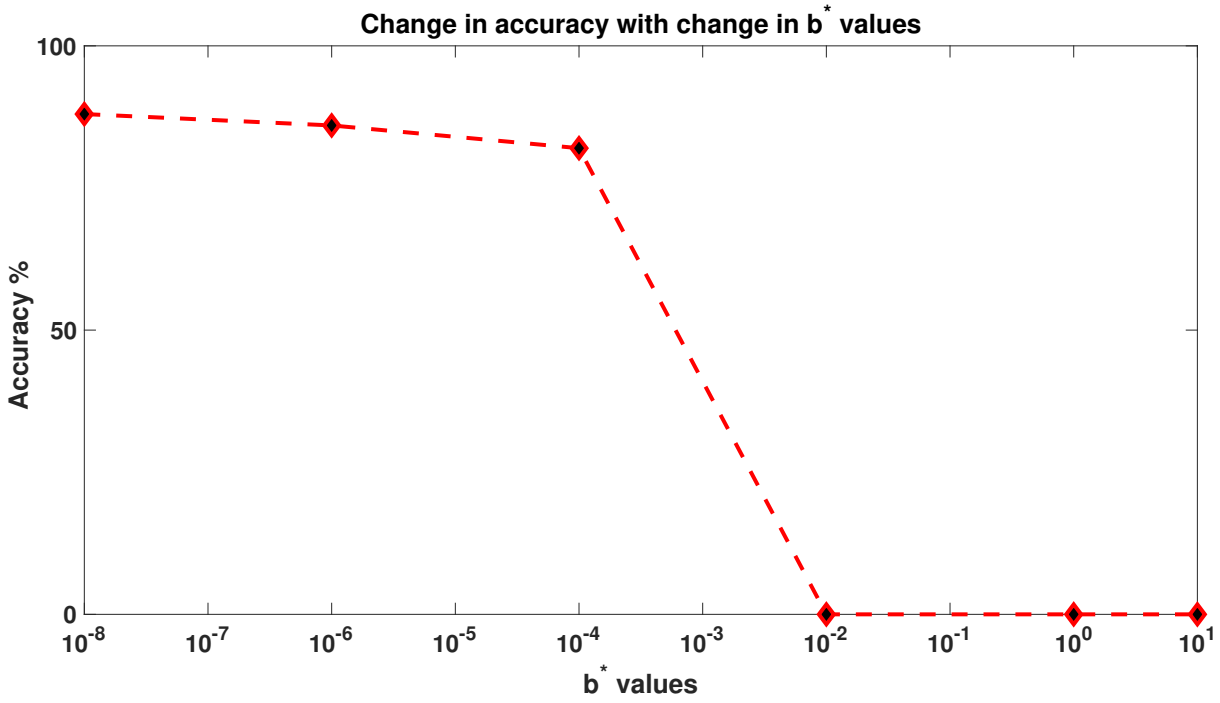

Figure 3.6: Accuracy of the proposed method for different b^* values

prior gamma distributions of the precision parameters $[\beta_{11}, \beta_{12}, ..., \beta_{DM}]$. It can be concluded that accuracy decreases as the b^* value increases. Thus, the choice of the hyperparameter b^* plays a crucial role in the accuracy of the method. Lower values of b^* will impose heavier penalty on the lower valued coefficients which in turn improves the accuracy of the method.

3.6 Conclusions

The existing Granger causality methods do not ensure the causal structure extracted from each mode of operation to be same. This can lead to inconclusive result on the cause-effect relationship among the variables of the multi-mode system when each mode has the same causal structure. Moreover, in these approaches the statistical tests to check for the significance of the causal connections are laborious. To overcome these drawbacks, a novel variational Bayesian approach is proposed to infer the causal relations in multi-modal systems, which introduces the same sparsity across coefficient

matrices across modes through a Normal-Gamma prior. This ensures that the causal structure extracted from each mode is consistent. The proposed method can be easily extended to a multivariate system and also provides a simpler statistical test for checking the significance of the causal connections. The method when implemented on simulation example was able to infer the causal connections with good accuracy. Additionally, it is observed that the accuracy of the proposed method is dependent on several parameters such as dimension of data, number of local models chosen, sparsity and hyperparameter values.

Chapter 4

A robust variational Bayesian approach for causality analysis in multi-modal systems

4.1 Introduction

Granger causality is a popular data-driven technique which infers causal relations among variables in a process system. The primary step in Granger causality analysis is the construction of time-series models for the process. However, the accuracy of the data-driven models obtained greatly depends on the quality of the data used for the identification process. Data derived from many industrial processes tend to have data points called outliers, which lie outside the normal range of the data and this can compromise the quality of the data. Instrument failures, process disturbances, human errors and errors during the transmission of data can give rise to outliers in the data. Presence of outliers leads to poor parameter estimates and subsequently inaccurate Granger causality analysis. In this chapter, the method developed in the earlier chapter is extended to infer Granger causality relations among variables in a nonlinear systems when process data contains outliers.

Numerous methods have been developed in the past for outlier detection [26] and process identification in the presence of outliers. Process identification by data after outlier removal can lead to loss of important process information in some cases. So

it is important to develop a model which can also describe the data with outliers. There exist both deterministic and probabilistic approaches to identify models that are robust against outliers. The deterministic approaches such as M estimator [27] which minimizes a weighted sum of square of residuals to reduce the effect of outliers, while the probabilistic approaches use noise models such as mixture of Gaussian distributions [28], t-distribution [29] and Laplace distribution [30] to handle the outliers. Granger causality involves the estimation of vector auto-regressive (VAR) model. The magnitudes of the coefficients represent the strength of causal connections among the variables of the considered process system. Probabilistic approach for robust identification and inference of Granger causality in multi-model systems are the main focus of this chapter.

Noise with mixture of Gaussian distributions

The outliers fall into two types, namely scale outliers and location outliers. Scale and location outliers are generated by shift in scale (variability) and location (mean) respectively [28],[30]. Scale outliers are usually modelled as a mixture of two Gaussian distributions with same mean but different covariance matrices such that one covariance matrix is inflated to include outliers i.e noise term ϵ_k is expressed as the following equation which is taken from [30],

$$\epsilon_k \sim \delta\mathcal{N}(0, \rho^{-1}\sigma_\epsilon^2) + (1-\delta)\mathcal{N}(0, \sigma_\epsilon^2) \tag{4.1}$$

where σ_ϵ^2 is the noise covariance, $0 < \rho < 1$ is the inflation factor and δ is the unknown prior probability of occurrence of outliers.

Location outliers arise due to several reasons such as jammed measuring instruments to name one. The distribution of noise term in presence of location outliers is usually modelled as the following equation which is taken again from [30],

$$\epsilon_k \sim \delta[\mathcal{N}(\Gamma, \sigma_\epsilon^2) + \mathcal{N}(-\Gamma, \sigma_\epsilon^2] + (1-\delta)\mathcal{N}(0, \sigma_\epsilon^2) \tag{4.2}$$

where σ_ϵ^2 is the covariance, Γ indicates the shift in location of the outlier and δ is again the prior probability of occurrence of outliers.

Noise with Student's t-distribution

Another common method to deal with outliers is to consider a t-distribution for the prediction error. T-distribution differs from a Gaussain disribution in that it has heavier tails to accommodate large-valued outliers, which make it more robust to outliers compared to a Gaussian distribution. The heavier tails in t-distribution help to handle outliers unlike a Gaussian distribution. In the past, robust parameter estimation has been carried out using t-distribution in many types of problems such as linear regression [31] and mixture probabilistic principle component regression models for soft-sensor development [32] to name a few. The probability density of noise ϵ is assumed to follow a t-distribution with mean μ, covariance σ^2 and degree of freedom ν is as the following [30], [33],

$$P(\epsilon|\mu,\sigma^2,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu\sigma^2)^{1/2}}\left(1+\frac{1}{\nu}\left(\frac{\epsilon-\mu}{\sigma}\right)^2\right)^{-(\frac{\nu+1}{2})} \tag{4.3}$$

where $\Gamma(t)$ is the Gamma function which is given as the following,

$$\Gamma(t) = \int_0^\infty z^{t-1}e^{-z}dz \tag{4.4}$$

The degree of freedom, ν, controls the width of the tails of the t-distribution and σ corresponds to the scale of the distribution. When $\nu \to \infty$, the t-distribution becomes a Gaussian distribution. Fig 4.1 shows t-distribution for different ν values. The mean and scale parameters are fixed at 0 and 1 respectively. From the figure, it is clear that as ν value decreases, the tails get heavier which can account for the outliers with larger values.

Another property of t-distribution which makes it useful for system identification is that it can be decomposed into a scaled Gaussian distribution and a gamma dis-

tribution as the following [30],

$$p(\epsilon|0,\sigma^2,\nu) = \int_0^\infty p(\epsilon|0,\sigma^2,r)p(r|\nu)dr \tag{4.5}$$

where $p(\epsilon|0,\sigma^2,r)$ is the scaled Gaussian distribution with scale r and $p(r|\nu)$ is the gamma distributions, mathematically expressed as the following,

$$\epsilon|0,\sigma^2,r \sim \mathcal{N}(0,\sigma^2/r) \tag{4.6}$$

$$r|\nu \sim \Gamma(\frac{\nu}{2},\frac{\nu}{2}) \tag{4.7}$$

The noise variable in the proposed approach is considered to follow a multivariate t-distribution.

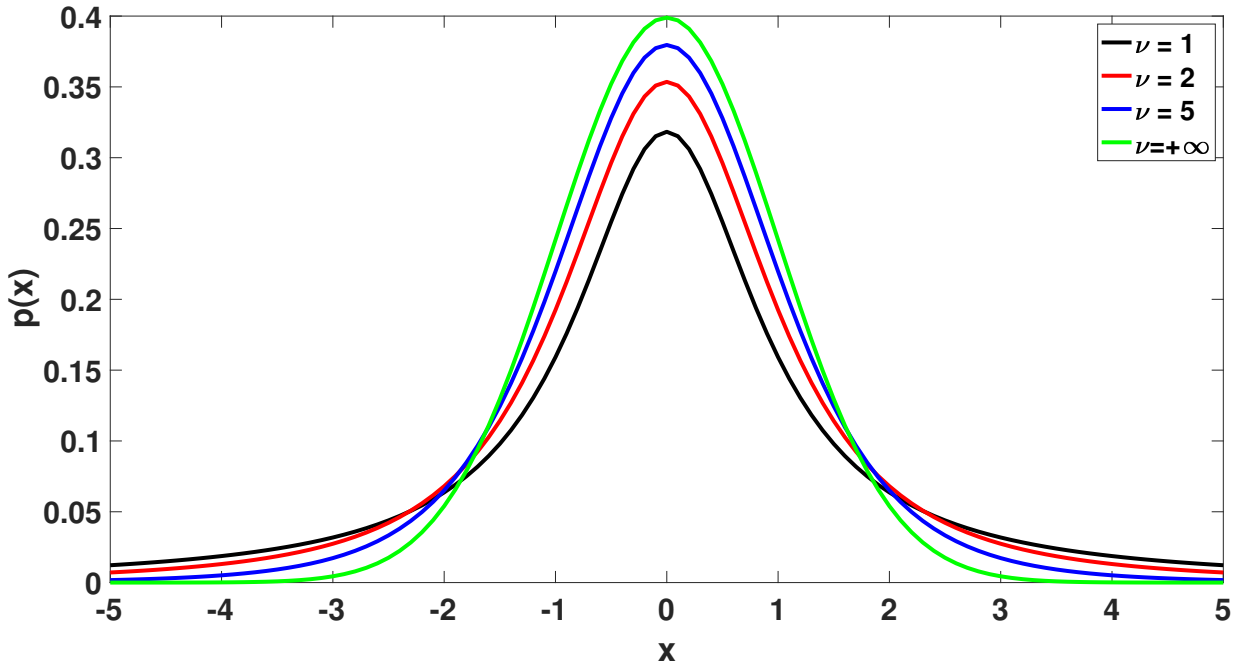

Figure 4.1: t-distribution for different degree of freedom value

Some new techniques have been developed to estimate VAR parameters and infer Granger causality more reliably in the presence of outlier noise. Granger causality analysis in Lp ($p \leq 1$) norm space [34] is one such method. Traditionally, Granger causality analysis involves the minimization of the L2 norm of the prediction error which is the objective function with respect to the coefficients of the VAR model.

It is done by taking first derivative of the objective function with respect to the coefficients and equating the derivative to zero. By doing so, the optimum values of the coefficients are obtained, which subsequently helps to determine the prediction errors and help in the estimation of Granger causality. In the presence of outliers, the L2 norm based objective function exaggerates the effect of outliers further as a square of the prediction error is used as the objective function. This method is unable to incorporate prior knowledge of the system in the estimation algorithm and also require more complicated statistical tests. The proposed Bayesian approach helps to overcome these drawbacks.

A single linear VAR model cannot give good parameter estimation for complex multi-mode processes. A switched VAR model is used in the proposed method, whose structure is given in detail in section 4.2. Many methods are available in literature to determine multi-model Granger causality relations in times series data contaminated with outliers. The method of robust time varying generalized partial directed coherence (rTV-gPDC) [35] uses a time varying multivariate auto-regressive model (TVAR) whose parameters are determined using Kalman filter. An outlier-free observation is estimated and it is used in place of the outlier observation. Such an observation is then used in a Kalman filter algorithm and this makes the algorithm robust against outliers. The coloured noise of the TVAR model is modelled as a time varying moving average (TVMA) model. The time varying nature of the parameters makes the TVAR and TMVA models more suitable for non-linear process when the process operates around multiple steady states. The gPDC obtained using the TVAR coefficients gives the causal relations among the variables of the system. Fujita et.al. [36] developed a robust statistical test for a VAR model using a likelihood ratio test statistic. The switched VAR considered in this current work has noise which is modelled as a t-distribution to account for the outliers in the data. Subsequently, a Variational Bayesian inference is used for the parameter estimation of model which helps in evaluating the Granger causality relations among the variables. There are

some advantages of using variational Bayesian method over the existing multi-model or time-varying methods. First, the variational approach proposes an easier statistical test to determine the significance of causal connections. Second, the Bayesian approach regularizes the model structure which helps in circumventing the tedious process of finding the best model by trying different combinations of model order and lag. Last, the variational Bayesian approach ensures that the causal structure extracted from each mode is consistent. This is achieved by choosing a Normal-Gamma prior for the identically positioned elements of coefficient matrices for all of the VAR models considered. It acts as a soft constraint during optimization using the variational Bayesian approach.

The rest of the chapter is organized as follows. Section 4.2 presents the proposed model followed by section 4.3 which discusses the proposed model estimation approach. Sections 4.4 and 4.5 give the implementation steps and case study results respectively. Finally, section 4.6 provides the concluding remarks.

4.2 Model description

In a multi-mode system, a switched VAR model needs to be used such that the process switches among different VAR models depending on the mode of operation of the process system. The switched VAR model considered in this chapter is shown below,

$$y(t) = \sum_{l=1}^{L} W(l)^s y(t-l) + e^s(t) \tag{4.8}$$

where $y(t) \in R^D$ is the observation at time t, l represents the lag and it ranges from 1 to L. $W(l)^s \in D \times D$ corresponds to the VAR coefficients of the s^{th} local model at time lag l. s and t range from 1 to ub and 1 to N respectively. $e^s(t) \in R^D$ is the process noise associated with the s^{th} local model. A probabilistic model is used to represent each VAR model (usually Gaussian distribution) and the resulting switched VAR model is represented as the weighted sum of the individual probabilistic models

(mixture VAR model). However, in the presence of data contaminated with outliers, we assume the noise associated with the process to follow a t-distribution. This, in-turn leads to the representation of the mixture VAR (MVAR) model as a sum of t-distributions as given below,

$$p(y(t)|Y^{t-1}) = \sum_{s=1}^{ub} \alpha_{s,t} t(y(t)|\mu_s, (\delta^s)^{-1} I_D, \nu^s) \tag{4.9}$$

where $p(y(t)|Y^{t-1})$ is the conditional distribution of $y(t)$ given its $t-1$ past values, Y^{t-1}. $\alpha_{s,t}$ is the mixing weight for the s^{th} local model. $(\delta^s)^{-1} I_D$ is the covariance where δ^s is a scalar value and $I_D \in R^{D \times D}$ is the identity matrix. ν^s is the degrees of freedom respectively and μ_s is the conditional mean which is given as the following,

$$\begin{gathered} \mu_s = \mu_{0,s} + \sum_{l=1}^{L} W(l)^s y(t-l) \\ \text{where } s = 1, 2, ..., ub \end{gathered} \tag{4.10}$$

In the above model, $W(l)^s$ are the unknown parameters. In a Bayesian mixture VAR model which is used in this thesis, they are considered to be random variables with probability distributions.

4.2.1 Proposed model

The proposed model in this work is a Bayesian mixture VAR model which switches from one mode to another as given in equation 4.8. In this work, the lag parameter is restricted to one in the derivations and case studies for simplicity of presentation but they can be extended to more lags following the same procedure as given below. The switched VAR model given in equation 4.8 can be modified for unit lag as the following,

$$y(t) = W^s y(t-1) + e^s(t) \tag{4.11}$$

Therefore, the coefficient matrix of the model reduces to the following form,

$$W^s = \begin{bmatrix} w^s_{11} & \dots & w^s_{1M} \\ \vdots & \ddots & \vdots \\ w^s_{D1} & \dots & w^s_{DM} \end{bmatrix} \tag{4.12}$$

where D and M are the number of columns and rows of the coefficient matrix of the s^{th} local model, W^s. For a VAR model, the number of rows and columns are the same. However, throughout this chapter, to distinguish between rows and columns, they are indicated as D and M respectively. The first step in a Bayesian approach is the construction of a Bayesian network by defining the prior distributions of the unknown parameters and latent variables.

The proposed work ensures that the causal structures extracted from all the operating modes are consistent, only the magnitudes or the strengths of the causal relations vary from one mode to another. This is achieved by assuming the identically positioned elements of coefficient matrices from all modes to follow the same normal-gamma prior. Consider the multivariate system with D variables discussed before has ub modes of operation. This would imply that ub number of coefficient matrices W^1, W^2 , ..., W^{ub} are present and it is assumed that any dm^{th} element of W^1, W^2, ..., W^{ub} follows the same Normal-Gamma distribution. The Normal-Gamma prior for the dm^{th} element of the coefficient matrices W^1, W^2 , ..., W^{ub} represented as $w^1_{dm}, w^2_{dm}, ..., w^{ub}_{dm}$ respectively is given as the following,

$$\begin{gathered}[w^1_{dm}, w^2_{dm}, ..., w^{ub}_{dm}] = W_{dm} \sim \mathcal{N}(0, \beta^{-1}_{dm} I) \\ \text{where } \beta_{dm} \sim \Gamma(a^*, b^*)\end{gathered} \tag{4.13}$$

where β_{dm} is the precision parameter of dm^{th} element of the coefficient matrices W^1, W^2 , ..., W^{ub}. The precision parameter β_{dm} is assumed to follow a gamma distribution with a^* and b^* as its shape and rate parameters respectively. The normal-gamma prior introduces a penalty on the lower valued coefficients, which will make the insignificant lower valued coefficients to converge to zero. Now, since identically positioned elements, for instance, the dm^{th} element of the coefficient matrices W^1, W^2 , ..., W^{ub} are considered to have the same Normal-Gamma prior and hence if the causal effect between m^{th} variable on the d^{th} variable is small in one mode, then the assumption would ensure that the causal relation remains insignificant in all other modes as well. Subsequently, $w^1_{dm}, w^2_{dm}, ..., w^{ub}_{dm}$ would all converge to zero

simultaneously. This would in-turn lead to identical sparsity structures in the coefficient matrices from all the modes. The above assumption helps to factorize the prior distribution of W as the following,

$$p(W|\beta) = \prod_{d=1}^{D}\prod_{m=1}^{M} p(W_{dm}|\beta_{dm}) = \prod_{d=1}^{D}\prod_{m=1}^{M} \mathcal{N}(W_{dm}|0, \beta_{dm}^{-1}) \tag{4.14}$$

where $W = [W^1, W^2, ..., W^{ub}]$ and W_{dm} is the set of element in the d^{th} row and m^{th} column from coefficient matrices W^1 to W^{ub}. The precision parameter of the dm^{th} element is assumed to follow a gamma distribution. The prior distribution of set of precision parameters $\beta = [\beta_{11}, \beta_{1,2}, ..., \beta_{dm}]$ can be factorized as the following,

$$P(\beta|a^*, b^*) = \prod_{d=1}^{D}\prod_{m=1}^{M} P(\beta_{dm}) = \prod_{d=1}^{D}\prod_{m=1}^{M} \Gamma(\beta_{dm}|a^*, b^*) \tag{4.15}$$

where a^* and b^* are the shape and rate parameters respectively.

The process noise is considered to follow a t-distribution as the following,

$$e^s(t) \sim t(0, (\delta^s)^{-1} I_D, \nu^s) \tag{4.16}$$

where δ^s is the precision of the local model s that is a scalar quantity, I_D is an identity matrix of dimension D and ν_s is the degree of freedom for the s^{th} local model. Further, the t-distribution can be decomposed into a scaled Gaussian distribution and a gamma distribution as the following,

$$t(e^s|0, (\delta^s)^{-1} I_D, \nu^s) = \int_0^\infty \mathcal{N}(e^s|0, \frac{(\delta^s)^{-1}}{R^{S(t)=s}} I_D)\Gamma(R^{S(t)=s}|\frac{\nu^s}{2}, \frac{\nu^s}{2}) dR^{S(t)=s} \tag{4.17}$$

where

$$\begin{aligned} \mathcal{N}(e^s|0, \frac{(\delta^s)^{-1}}{R^{S(t)=s}} I_D) &= \frac{(R^{S(t)=s}\delta^s)^{D/2}}{2\pi^{D/2}} \exp - \frac{R^{S(t)=s}\delta^s (e^s T e^s)}{2} \\ \Gamma(R^{S(t)=s}|\frac{\nu^s}{2}, \frac{\nu^s}{2}) &= \frac{1}{\Gamma(\frac{\nu^s}{2})} (\frac{\nu^s}{2})^{\frac{\nu^s}{2}} (R^{S(t)=s})^{\frac{\nu^s}{2}} \exp^{-\frac{\nu^s}{2} R^{S(t)=s}} \end{aligned} \tag{4.18}$$

S represents the model identity, $R^{S(t)=s}$ and δ^s are the scale and precision parameters for the s^{th} local model. The prior distribution of scale R of the noise vector is expanded as the following,

$$p(R|S) = \prod_{s=1}^{ub}\prod_{t=1}^{N} \Gamma(R^{S(t)=s}|\frac{\nu^{S(t)=s}}{2}, \frac{\nu^{S(t)=s}}{2}) \tag{4.19}$$

No two local models are considered to have the same scale for its noise vector. Additionally within each local model, the scale of noise is assumed to vary with time. The noise precision is considered to be an unknown parameter in this chapter, which will be estimated.

The joint distribution of the observed data Y is expressed as the product of N multivariate t- distributions as given below,

$$p(Y|W^s, \delta^s, \nu^s, S = s) = \prod_{t=1}^{N} p(y(t)|W^s, \delta^s, \nu^s, S = s) = \prod_{t=1}^{N} t(y(t)|0, \delta^{s-1} I_D, \nu^s) \tag{4.20}$$

$p(y(t)|W^s, \delta^s, \nu^s, S = s)$ is the conditional probability distribution for which $y(t)$ is generated by the s^{th} local model.

Now, we move to defining the prior distributions of the latent variable of the model. The prior distribution of the model identity S, which is a latent variable, is given below,

$$p(S = s) = \sum_{t=1}^{N} p(S(t) = s) = \sum_{t=1}^{N} \alpha^{S(t)=s} \tag{4.21}$$

where $p(S = s)$ is prior probability of local model s which is obtained by summing up the individual prior probabilities ($\alpha^{S(t)=s}$) of local model s for N time instants. Now, the significance coefficient of local model s, α^s, is given as follows,

$$\alpha^s = \frac{\sum_{t=1}^{N} \alpha^{S(t)=s}}{N} \tag{4.22}$$

which implies the significance coefficient of local model s is obtained by taking the time average of the prior probabilities of local s. It is assumed that as prior each of the local model is equally probable. This is achieved by assuming a symmetric Dirichlet function as prior for the significance coefficients setα which is given as the following,

$$p(\alpha) = Dir(\alpha|\alpha^* m^*), m^* = \left[\frac{1}{ub}, ..., \frac{1}{ub}\right] \tag{4.23}$$
$$\text{where } \alpha = \left[\alpha^1, \alpha^2, ..., \alpha^{ub}\right]$$

α^1, α^2 are the significance coefficients of local models 1, 2 respectively. α^* is the hyperparameter and $\alpha^* m^*$ is parameter vector of the Dirichlet distribution. Since α^* is scalar and elements of scale vector of the prior, $m*$, are identical, the resulting Dirichlet distribution is symmetric. The above assumptions on the prior distributions help in the construction of the Bayesian network shown in Fig 4.2. The random variables are represented as circular nodes. The nodes inside the rectangle repeats by the number given in its bottom corner. The nodes are connected based on their parent-child relations.

4.3 Estimation

The unknown parameters set is $\Phi = [W, \beta, R, \alpha]$ and latent variable is the model identity S. Given data $Y = [y(1), y(2), ..., y(N)$ and model structure M which is the switched VAR model, the posterior distribution of ϕ and S is given by Bayes rule as the following,

$$p(\Phi, S|Y, M) = \frac{p(Y|\Phi, S, M)p(\Phi, S|M)}{p(Y|M)} = \frac{p(Y|\Phi, S, M)p(\Phi, S|M)}{\int_\Phi \int_S p(Y|\Phi)p(\Phi, S|M)} \tag{4.24}$$

where $p(Y|\Phi, S, M)$ represents the likelihood of data Y, $p(\Phi, S|M)$ represents the joint prior of parameters and latent variables, and finally, $p(Y|M)$ represents the model evidence. However, the exact determination of posterior distribution is often impossible as the model evidence $P(Y|M)$ which gives the likelihood of the data Y being generated by the model M is intractable. In such situations, the approximate posterior distributions of the unknown parameters and latent variable are evaluated using the variational Bayesian expectation maximization (VBEM) approach. The VBEM approach assumes the actual posterior distribution to be approximated as the following factorization,

$$p(\Phi, S|Y, M) \sim q(\Phi)q(S) \tag{4.25}$$

where $q(\Phi)$ and $q(S)$ represent the approximate posterior distributions. The above approximation will ensure the tractability using the VBEM algorithm. The VBEM

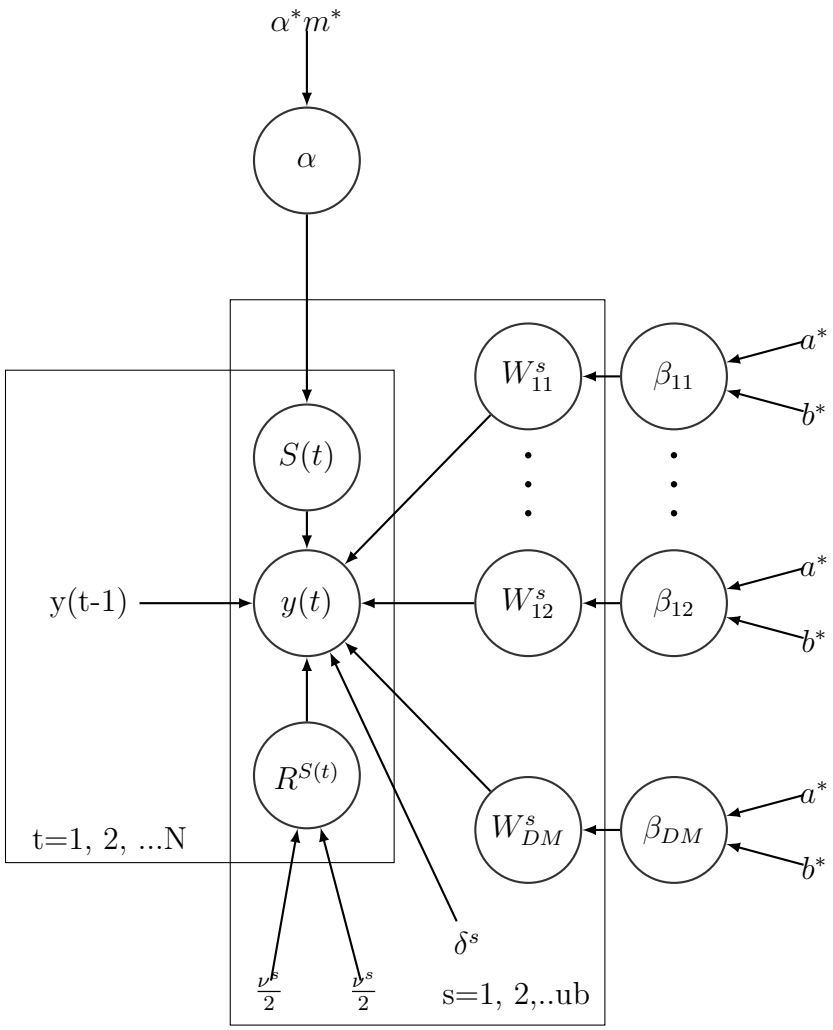

Figure 4.2: Bayesian Network for the proposed model

approach involves the maximization of the lower bound of the log of model evidence iteratively. The lower bound of log of model evidence is derived from the model evidence in the following manner,

$$\begin{aligned}\ln p(Y|M) =& \sum_S \int_\Phi q(\Phi)q(S) \ln \frac{p(Y,S,\Phi|M)}{q(\Phi)q(S)} d\Phi \\ &+ \sum_S \int_\Phi q(\Phi)q(S) \ln \frac{q(\Phi)q(S)}{p(\Phi,S|Y,M)} d\Phi \\ =& L(q(S),q(\Phi)) + KL(q(\Phi)q(S)||p(\Phi,S|Y,M))\end{aligned} \tag{4.26}$$

where the first term $L(q(S),q(\Phi))$ is the lower bound of the model evidence and the second term $KL(q(\Phi)q(S)||P(\Phi,S|Y,M))$ is KL divergence between the approximate and true posterior distributions. Our objective is to obtain an approximate posterior distribution which is as close to the actual posterior distribution as possible. This can be achieved by minimizing the KL divergence term. The model evidence is a fixed value given a particular model structure. Additionally, KL divergence is always a positive quantity, which in turn would imply that, minimizing KL divergence is equivalent to maximizing the lower bound. The lower bound $L(q(S),q(\Phi))$ is iteratively maximized with respect to $q(\Phi)$ and $q(S)$.

4.3.1 Approximate posterior distribution

The approximate posterior distribution of unknown parameters $q(\Phi)$ is further factorized as follows,

$$q(\Phi) = q(W,\beta,R,\alpha) = q(W|S)q(\beta)q(R|S)q(\alpha) \tag{4.27}$$

where $W = [W^1, W^2, ..., W^{ub}], \beta = [\beta_{11}, \beta_{12}, ..., \beta_{DM}]$, $R = [R^1, R^2, ..., R^{ub}]$ and $\alpha = [\alpha^1, \alpha^2, ..., \alpha^{ub}]$. The above approximation is to make posterior distribution approximation tractable. Further factorization of the approximate posterior distributions $q(W), q(\beta), q(R)$ and $q(\alpha)$ is done using the D-separation rules. Each row of the coefficient matrix W is independent given Y, as each row of W acts as parent to a particular dimension of Y and no two rows of the matrix W share a common child.

From the Bayesian network given in Fig. 4.2, W is dependent on β as they share a parent child relationship. However, from equation 4.27, it assumed that W, β, R and α are independent to each other in their posterior. This is the trade off that needs to be made to make the calculation of approximate posterior distributions tractable. The expansion of approximated posterior probability of the coefficient matrix of the s^{th} local model, $q(W|S=s)$ is given as follows,

$$q(W|S=s) = \prod_{d=1}^{D} q(W_d^s|\hat{W}_d^s, \Sigma_{\hat{W}_d^s}) = \prod_{d=1}^{D} \mathcal{N}((W_d^s|\hat{W}_d^s, \Sigma_{\hat{W}_d^s}) \tag{4.28}$$

Each row W_d^s of the coefficient matrix W^s is assumed to follow a Gaussian distribution with mean $\hat{W}_d^s$ and covariance $\Sigma_{\hat{W}_d^s}$. Now, $q(\beta)$ is further expanded as follows,

$$q(\beta) = \prod_{d=1}^{D}\prod_{m=1}^{M} q(\beta_{dm}) = \prod_{d=1}^{D}\prod_{m=1}^{M} \Gamma(\beta_{dm}|a, b_{dm}) \tag{4.29}$$

where a and b_{dm} are the shape and rate parameters respectively. The above expansion is possible as the set of dm^{th} elements of coefficient matrices, W_{dm}, has a different parent represented as β_{dm} and none of the parents have a common child. The approximate posterior distribution $q(R|S)$ of the scale R of the noise vector can be expanded as follows,

$$q(R|S) = \prod_{s=1}^{ub}\prod_{t=1}^{N} \Gamma(R^{S(t)=s}|{\nu_\alpha}^{S(t)=s}, \nu_\beta^{S(t)=s}) \tag{4.30}$$

where ${\nu_\alpha}^{S(t)=s}$ and $\nu_\beta^{S(t)=s}$ are shape and rate parameters of the gamma distribution respectively. At each local model s, $R^{S(t)=s}$ has $y(t)$ as child and none of the parents have a common child, which makes them independent according to the D-separation principle. Since they are independent to each other, the above separation can be made. The model indicator variables S has the approximate posterior distribution as follows,

$$q(S=s) = \sum_{t=1}^{N} q(S(t)=s) = \sum_{t=1}^{N} \alpha_{new}^{S(t)=s} \tag{4.31}$$

where $\alpha_{new}^{S(t)=s}$ is the approximate posterior probability of model s at time t. At any time instant, it is assumed that the data is generated by any one of the ub number of local models. This imposes an additional constraint as follows,

$$\sum_{s=1}^{ub} \alpha_{new}^{S(t)=s} = 1 \tag{4.32}$$

which in turn would imply that

$$\sum_{t=1}^{N} \sum_{s=1}^{ub} \alpha_{new}^{S(t)=s} = N \tag{4.33}$$

The significance coefficient for the s^{th} local model is as the following,

$$\alpha^s = \frac{\sum_{t=1}^{N} \alpha_{new}^{S(t)=s}}{N} \tag{4.34}$$

No two significance coefficients have the same parent and none of the parents have a common child. As a result, the significance coefficient set α has the approximate posterior which can be expanded as follows,

$$q(\alpha) = \prod_{s=1}^{ub} q(\alpha^s) = \prod_{s=1}^{ub} q(\alpha^s | \alpha_{new}^* m_s) = \prod_{s=1}^{ub} Dir(\alpha^s | \alpha_{new}^* m_s) \tag{4.35}$$

where $\alpha_{new}^* m_s$ is the parameter of the Dirichlet distribution of α^s.

4.3.2 Lower bound

The next step in the VBEM framework is the calculation of a lower bound for the model evidence. The lower bound can be expanded as follows,

$$
\begin{aligned}
L(q(\phi), q(S)) = & \sum_{s=1}^{ub} \int_{\beta} q(\beta) \sum_{d=1}^{D} \int_{W} q(W_d|S=s) \ln \frac{p(W_d^s|\beta_d)}{q(W_d|S=s)} \\
& + \sum_{d=1}^{D} \sum_{m=1}^{M} \int_{\beta_{dm}} q(\beta_{dm}|a, b_{dm}) \ln \frac{p(\beta_{dm}|a^*, b^*)}{q(\beta_{dm}|a, b_{dm})} \\
& + \sum_{t=1}^{N} \sum_{s=1}^{ub} q(S(t)=s) \int q(\alpha^s) \ln \frac{p(S(t)=s|\alpha^s)}{q(S(t)=s)} \\
& + \sum_{s=1}^{ub} q(\alpha^s) \ln \frac{p(\alpha^s|\alpha^*)}{q(\alpha^s)} \\
& + \sum_{s=1}^{ub} \sum_{t=1}^{N} \int q(R|S(t)=s) \ln \frac{p(R|S(t)=s)}{q(R|S(t)=s} \\
& + \sum_{s=1}^{ub} \int q(W|S=s) q(S=s) q(R|S=s) \ln p(Y|W, \delta, S=s)
\end{aligned} \tag{4.36}
$$

Except for the last term in the above expression, rest of the terms can be expressed as expectations of KL divergences. For instance, the first term can be expanded as the following,

$$
\sum_{s=1}^{ub} \int_{\beta} q(\beta) \sum_{d=1}^{D} \int_{W} q(W_d|S=s) \ln \frac{p(W_d^s|\beta_d)}{q(W_d|S=s)} \tag{4.37}
$$

$$
\begin{aligned}
= & \sum_{S=1}^{ub} \int q(\beta) [- \sum_{d=1}^{D} KL(q(W_d^S|\hat{W}_d^s, \Sigma_{\hat{W}_d^s}) || p(W_d|[0]_{1xM}, (diag([\beta_{d1}, ..., \beta_{dM}]^T))^{-1}))] \\
= & \frac{1}{2} \sum_{S=1}^{ub} \sum_{d=1}^{D} \ln |\Sigma_{\hat{W_d}^{S(t)}}| + \frac{ub}{2} \sum_{d=1}^{D} \sum_{m=1}^{M} (\psi(a) - \ln b_{dm}) + \frac{ubDM}{2} \\
& - \frac{1}{2} \sum_{S=1}^{ub} \sum_{d=1}^{D} tr[\lambda_d (\Sigma_{\hat{W_d}^{S(t)}} + (\hat{W}_d^{S(t)^T} \hat{W}_d^{S(t)})]
\end{aligned} \tag{4.38}
$$

where

$$
\lambda_d = diag\left([\frac{a}{b_{d1}}, ..., \frac{a}{b_{dM}}] \right) \tag{4.39}
$$

The explicit expression for lower bound is given in table B.1 of appendix.

4.3.3 Updates of posterior distribution

The approximate posterior distributions of the unknown parameters and hidden variables are obtained by maximizing the lower bound with respect to each of the approximate posterior distributions of unknown parameters and hidden variables one at a time in an iterative manner and updating the hyper parameters continuously. The derivatives of lower bound are taken with respect to each of these approximate posterior distributions and are equated to zero. For instance, the updated hyperparameters of the approximate posterior distribution $q(W|S)$ is obtained as the following,

$$\frac{\partial L}{\partial q(W|S=s)} = \int q(\beta)\ln P(W|\beta)d\beta - \int q(\beta)\ln q(W|S=s)d\beta - 1 \tag{4.40}$$
$$+\sum_{t=1}^{N} q(S(t)=s)\ln p(y(t)|W^s, S(t)=s, \delta^s, \alpha) = 0$$

$$\implies \Sigma_{\hat{W}_d{}^s} = [\lambda_d + \delta^s \sum_{t=1}^{N} \alpha_{new}^{S(t)=s} y(t-1)y(t-1)^T]^{-1} \tag{4.41}$$

$$(\hat{W}_d{}^s)^T = \Sigma_{\hat{W}_d{}^s}[\sum_{t=1}^{N} \alpha_{new}^{S(t)=s}\delta^s y_d(t)y(t-1)] \tag{4.42}$$

where

$$\lambda_d = diag\left([\frac{a}{b_{d1}}, ..., \frac{a}{b_{dM}}]\right) \tag{4.43}$$

$\Sigma_{\hat{W}_d{}^s}$ and $\hat{W}_d{}^s$ represent the covariance and mean of multivariate Gaussian distribution followed by the d^{th} row vector of coefficient matrix W^s. $y_d(t)$ represents the d^{th} element of observation $y(t)$. δ^s is the precision parameter of the s^{th} local model and $\alpha_{new}^{S(t)=s}$ is the approximate posterior probability of indicator variable S taking value s at time t. The updates for the hyper parameters of the approximated posterior distributions are given in table B.2 of the appendix.

4.3.4 Hyperparameter selection through Bayesian optimization

In some situations, when the prior process knowledge is not informative, it is better to determine the hyperparameters of the prior distributions through cross validation, as

it is proposed in the previous chapter. The first step in cross validation is separation of given data into training and validation sets. Subsequently, for different choices of hyperparameters, the parameters of the model are determined using training data and are validated using the validation data. The log likelihood function given in equation 4.44 is the validation criterion which is maximized through cross validation.

$$\sum_{t=1}^{N_{val}} \log \sum_{s=1}^{ub} \hat{\alpha}^s \mathcal{N}(y(t)|\widehat{W}y(t-1), \hat{\delta}^s I_D) \tag{4.44}$$

where N_{val} is the total number of data points in the validation set, $\widehat{W}$ and $\hat{\delta}^s$ are obtained from the update equations given in table B.2, I_D is an identity matrix of dimension D. $\hat{\alpha}^s$ is the intial guess for the significance coefficient of the local model s. Bayesian optimization is performed using the MATLAB built-in function 'Bayesopt' to perform cross validation by minimizing the negative log likelihood of the validation data. For a given b^* value, decrease in a^* value imposes a heavier penalty on smaller valued coefficients. This can be better understood using the expression of penalty which is added to the m^{th} column of the coefficient matrix given below,

$$\frac{a}{b_{dm}} = \frac{a^* + \frac{ub}{2}}{b^* + \frac{1}{2}\sum_{s=1}^{ub}\left[\hat{W}_{dm}^{s\ 2} + \Sigma_{\hat{W}_{dm}^s}\right]} \tag{4.45}$$

where the term in the bracket of the denominator represents the posterior mean of $(W_{dm}^s)^2$ which is represented as $E((W_{dm}^s)^2)$. Now, for a fixed value of $b*$, say 10^{-8}, the effect of decreasing a^* on the penalty $\frac{a}{b_{dm}}$ is shown in Fig 3.2. Clearly from Fig 3.2, for a certain small value of b^*, the penalty added to smaller valued coefficients increases when a^* value is decreased. Hence in this work, b^* is fixed at a small value of 10^{-8} and the best value of a^* is chosen between 10^{-8} to 10^8.

4.4 Implementation steps

Various steps involved in the implementation of the proposed method is summarized in table 4.1. Once the implementation is complete, our proposed approach helps in

Table 4.1: Implementation steps

Steps	
1	Fix values for MaxIter , threshold, ϵ and upper bound on number of local models, ub
2	Perform Bayesian optimization and determine the value of the hyperparameter a^* for a fixed value of b^*
3	Assign initial guess values for the rest of the hyperparameters of the prior and approximate posterior distributions of parameter set $\Phi = [W, \beta, R, \alpha]$ and latent variables in set S
4	Compute the lower bound L(k) when $k = 1$ using the parameters of the prior and initial parameters of approximate posterior distributions $q(\Phi)$ and $q(S)$
5	For idx=1:MaxIter
6	Update parameters of $q(S)$ by taking derivative of the lower bound $L(k)$ with respect to $q(S)$ keeping $q(\Phi)$ constant and equating it to zero
7	Update parameters of $q(\Phi)$ by taking derivative of the lower bound $L(k)$ with respect to $q(\Phi)$ keeping $q(S)$ constant and equating it to zero
8	If reminder of (idx/10)=0 then k=k+1
9	Recompute the lower bound L(k), using the previously updated parameters
10	If $\|(L(k) - L(k-1)\|/\|L(k-1)\| \leq \epsilon$
11	Break For loop
12	Else
13	End both If loops
14	idx=idx+1 and repeat steps from 6 to 13
15	End For

the formulation of a metric which helps to determine the relevance of the estimated causal connections. The expected value of the gamma distributed precision parameter of the dm^{th} element of the coefficient matrix W is given as follows:

$$\beta_{dm} = \frac{a}{b_{dm}} = \frac{a^* + \frac{ub}{2}}{b^* + \frac{1}{2}\sum_{s=1}^{ub} E(W_{dm}^s)^2} \tag{4.46}$$

where $E((W_{dm}^s)^2)$ represents the posterior mean of $(W_{dm}^s)^2$. When the value of the inverse of the metric β_{dm} is small, it would imply that the sum of posterior expectation of $(W_{dm}^s)^2$ from all the local models is close to zero. This in turn would imply that the values of W_{dm}^s in all the local models are close to zero. Thus, by assigning a lower threshold value to the inverse of the metric β_{dm}, the parameters with inverse of metric value lesser than the threshold value can be considered to be irrelevant.

4.5 Simulation case study

Multi-modal data contaminated with outliers is generated using the following model,

$$y(t) = W^s y(t-1) + e^s(t) \tag{4.47}$$

where W^s is the coefficient matrix which represents causal connections in the s^{th} local model. Model identity s and time instant t range from 1 to ub and 1 to N respectively. Noise e^s is considered to be a zero mean and unit variance Gaussian distribution.

For the above simulation, $ub = 3$ and $N = 3000$ were considered, such that there were 1000 data points for each local model. The generated data was corrupted with outliers. Outliers are data points outside the five sigma bound. A total of 50 different causal connections were considered for the simulation and subsequently accuracy of the method was checked when different parameters of the model, namely outlier percentage, number of local models (ub), dimension of data, noise variance and sparsity of the coefficient matrix W were varied. Accuracy is defined as the number of causal connections determined correctly. The specific values of the attributes of the model used in the simulation are given in table 4.2.

Table 4.2: Simulation details for the relevance study

Attribute	Value
Number of local models	3
Model order	1
Dimension of input and output	2 to 6
Threshold of the relevance metric	10^{-3}
Total time instants	3000
Sparsity	$\sim U(0.1, 0.4)\& \sim U(0.4, 0.9)$
Parameter, W	$\sim U(-1.9, -1)\& \sim U(1, 1.9)$
Noise mean	0
Noise variance	0.1 0.3 0.5 0.7 1

The first set of simulation studies performed demonstrates the accuracy of the method for different percentages of outliers in the data. Four different percentages of outliers, namely $5, 10, 15$ and 20 percent were considered for the simulation study. From Fig 4.3, it is clear that the accuracy of the proposed method is high even with a considerable percentage of outliers in the data.

The number of local models does not seem to have an apparent effect on the accuracy of the method (Fig 4.4) when ub ranges from 2 to 5. A slight increase in accuracy is observed when the number of local models increases from 5 to 6. However, higher number of local models are not preferred as the number of parameters to be estimated also increases; hence a optimum value has to be determined.

The accuracy of the method for different dimension of data is given in Fig 4.5. In general there is a decrease in accuracy with increase in dimension. This might be due to the difficulty for generating initial guesses close to global optima in higher dimensional problems. Due to this, the optimizer might converge to local optima. In this work, different initial guesses were chosen arbitrarily and subsequently, the

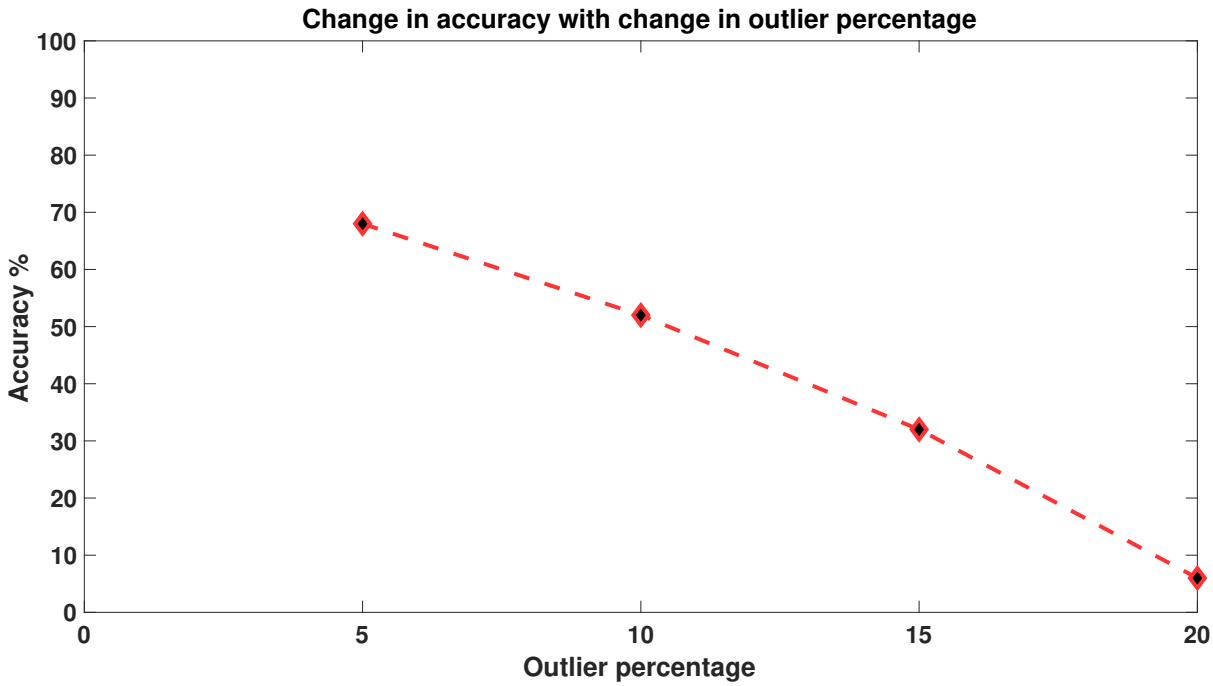

Figure 4.3: Accuracy of the proposed method for different percentages of outliers

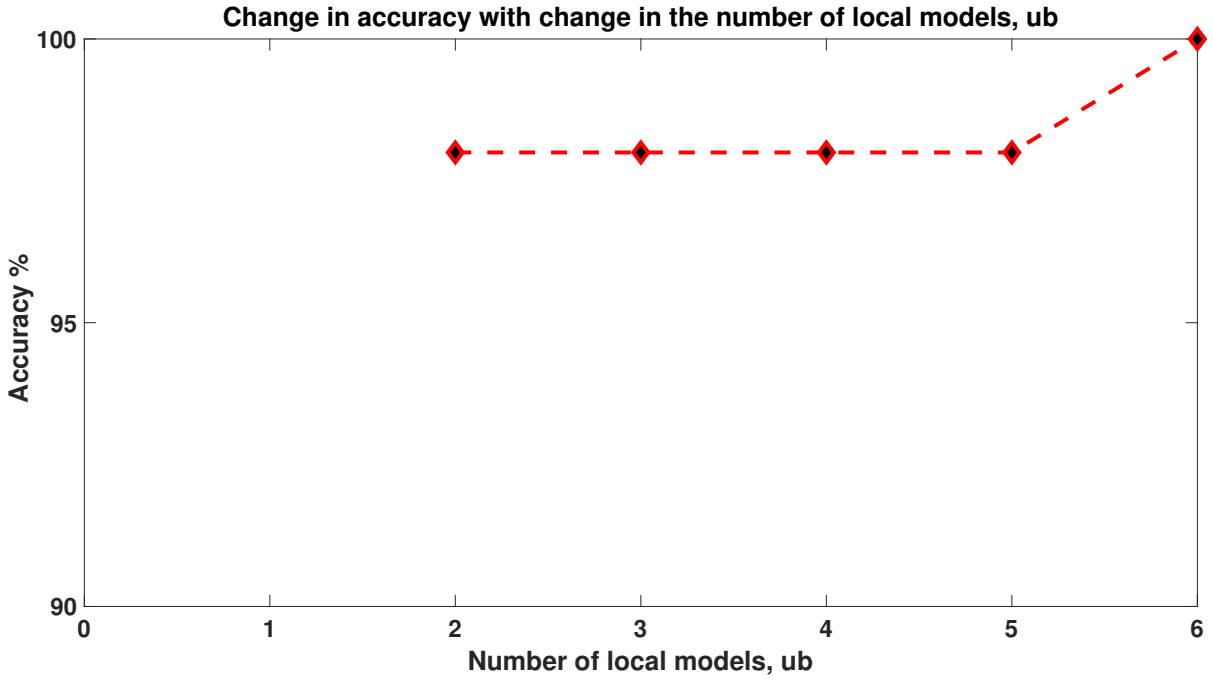

Figure 4.4: Accuracy of the proposed method for different number of local models, ub

accuracy of the method was calculated for each of these initial guesses to choose the best guess.

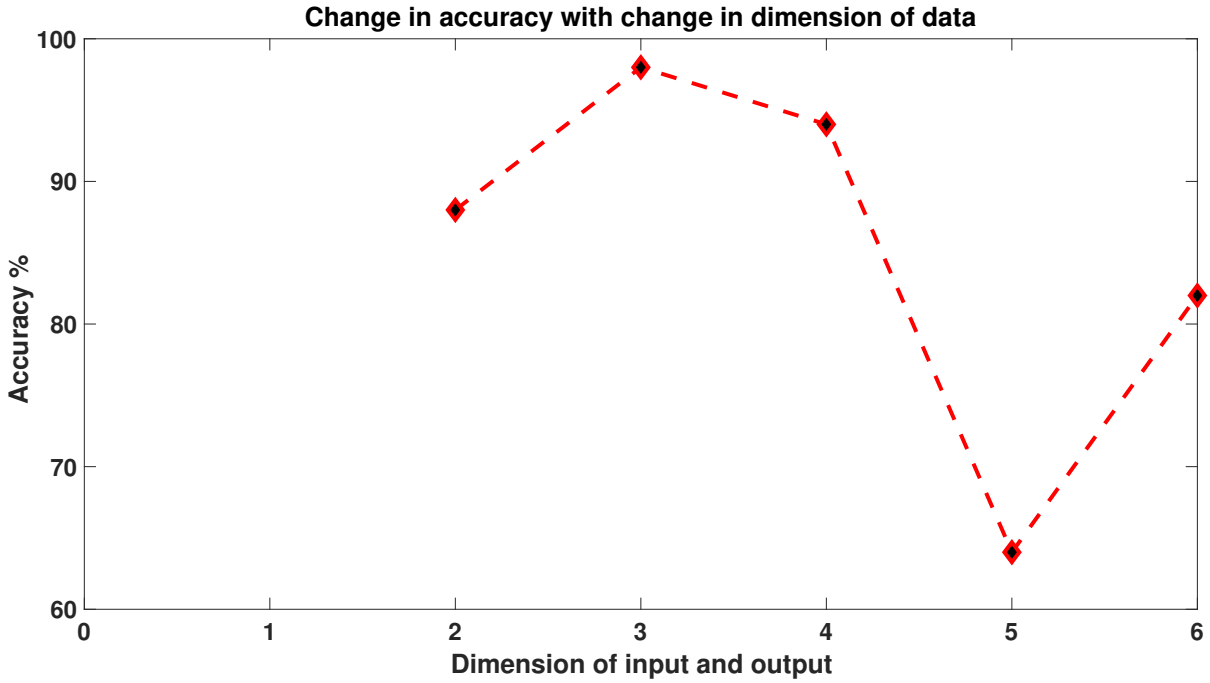

Figure 4.5: Accuracy of the proposed method for different dimensions of data

The noise plays a dual role in this simulation example as both disturbance and excitation. The variance of the disturbance is varied and the accuracy of the method for different variances is given in Fig 4.6. As the variance of the disturbance increases, the accuracy was observed to increase. It was observed that the inverse of expected value of precision parameter (inverse of β_{dm} given in equation 4.46) decreases with increase in noise variance. This would in turn mean that penalty added to the lower valued coefficients increases with variance. This helps in eliminating insignificant causal connections which increases the accuracy.

Sparsity also influences the accuracy of the method (table 4.3). As the sparsity of the coefficient matrix reduces, more parameters have to be estimated, which affects the accuracy of the method, causing it to decrease.

The last set of simulation studies were carried out to study the effect of hyperparameter values on the accuracy (Fig 4.7). Clearly low b^* values are preferred as they

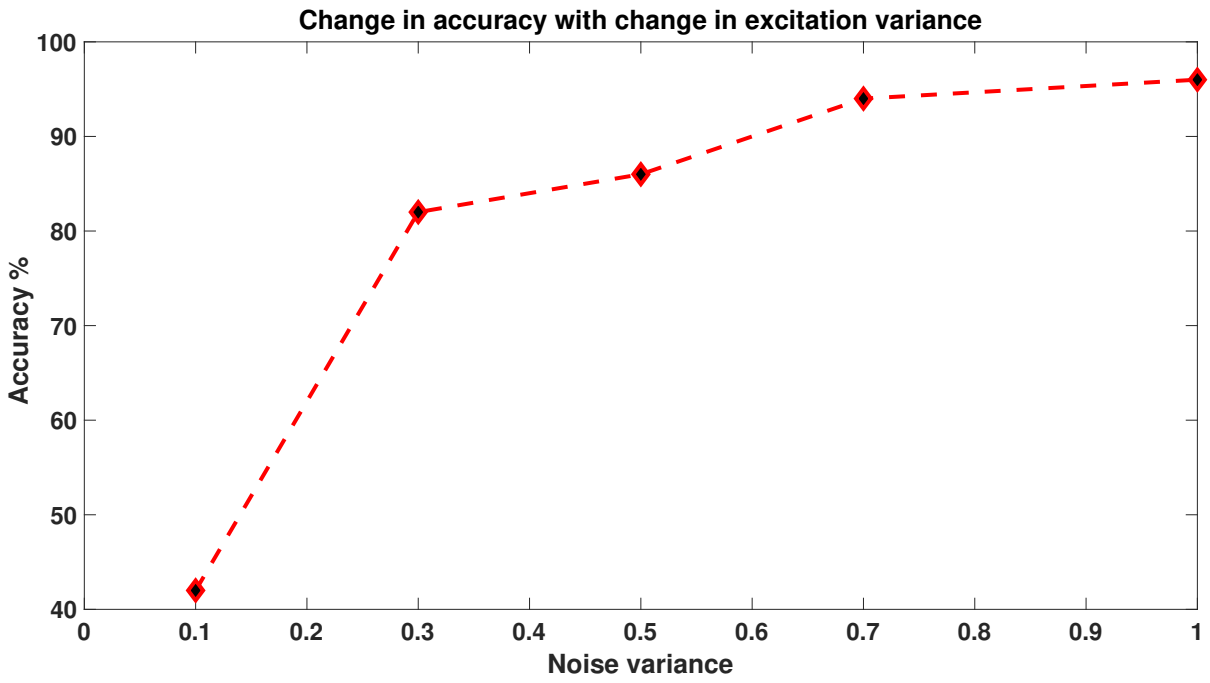

Figure 4.6: Accuracy of the proposed method for different noise variance

Table 4.3: Accuracy results for different sparsity of the coefficient matrix W

Sparsity	Number of causal cases	Accuracy Percentage
$\sim U(0.4, 0.9)$	15	86.67
$\sim U(0.1, 0.4)$	15	53.33

give higher accuracy. The possible reason is that penalty imposed on lower valued coefficients is decreased when b^* value increases. This can affect the accuracy of the method.

4.6 Conclusions

The failure to account for the influence of outliers can greatly reduce the performance of the causality analysis methods. Furthermore, the existing causality methods which are robust against outliers have complex statistical tests to determine the significance of the causal connections. In this chapter, a robust Granger causality technique

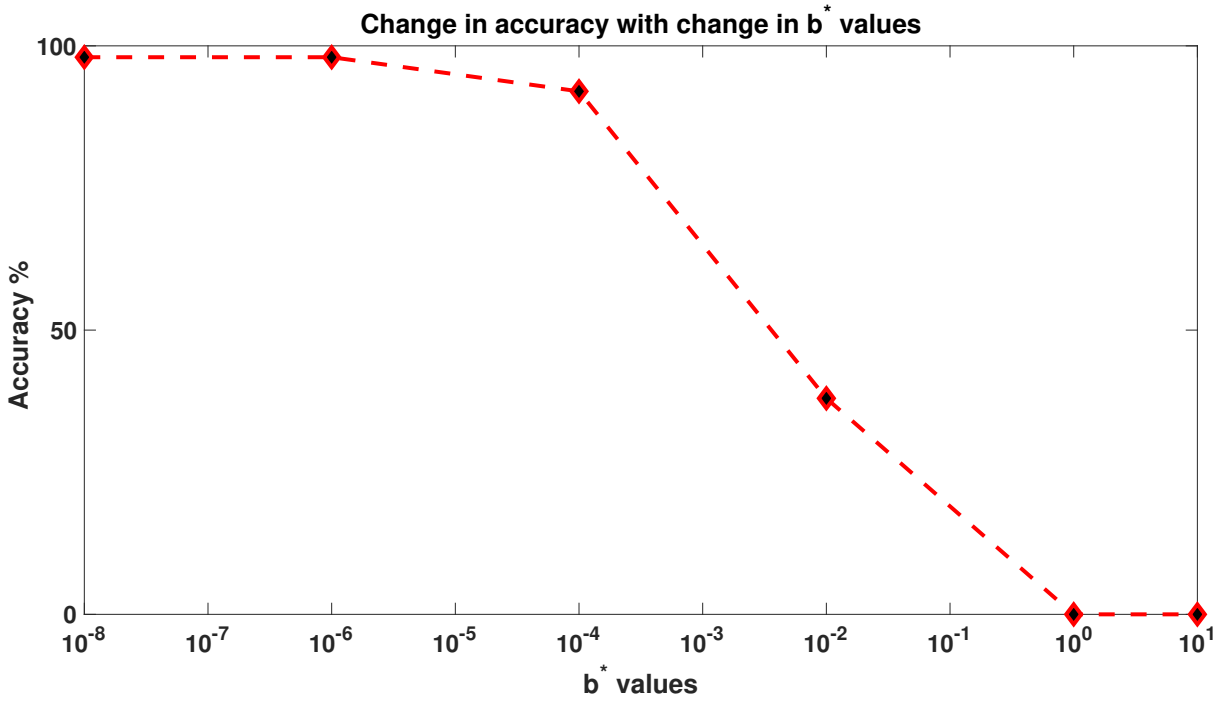

Figure 4.7: Accuracy of the proposed method for different b^* values

for multi-model systems using the variational Bayesian approach is proposed. As the data-driven models can be expressed as a Bayesian network, the outliers can be accommodated by modelling the prediction error by a t-distribution. This popular technique is used to account for the outliers in this work. In addition, this method also proposes a simple statistical test to check the significance of the causal connections. From the simulation results, it is evident that the proposed method is able to mitigate the effect of different percentages of outliers in the data to give accurate results.

Chapter 5

A comparative study of the two methods using an industrial example

The two proposed methods were implemented on a real industrial process. This chapter presents the results for the multi-model case first when no outliers were present, followed by providing a comparison of the performance of the two methods on the same industrial data after outliers were introduced. The industrial system considered is a refinery process which consists of a fluid catalytic cracking (FCC) unit and an unsaturated gas plant. These two units are part of any refinery settings and converts unmarketable gas oils into lighter oils with higher market value such as gasoline, fuel gas etc.

5.1 Process description

The FCC unit consists of a reactor-regenerator system followed by a fractionator. The gas oil feed is combined with fluidized solid catalyst (at high temperature) and sent from bottom of a reactor where the thermal cracking of the heavy gas oil into lighter components takes place. A simplified schematic of the process along with the stream numbers is given in Fig 5.1. The lighter products are separated from the catalyst using a riser termination device (RTD) device and a series of cyclone separators. The mixture of lighter products (stream 5) is sent to a fractionator and the

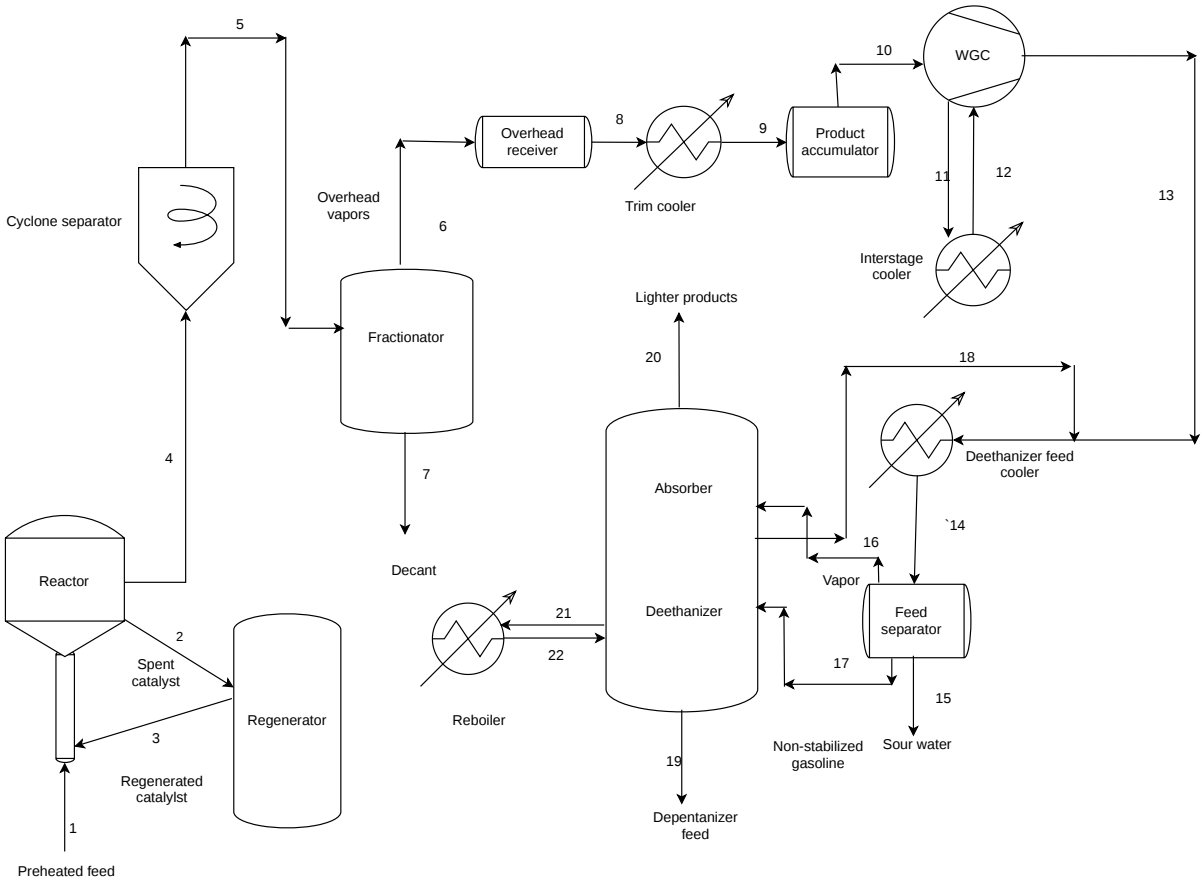

Figure 5.1: Simple schematic flowchart of the industrial process

deactivated/spent catalyst goes to a regenerator to burn off the coke deposited on the catalyst surface during thermal cracking process. Regenerated catalyst is then returned to the reactor riser feed mixing point. The fractionator system separates the mixture of lighter products into an overhead vapor stream (light ends and LPG products in a downstream plant), overhead liquid stream (olefin and gasoline products in a downstream plant), a liquid side draw (light cycle oil) and a bottom product (decant) based on the differences in volatility. Unsaturated gas plant separates the overhead stream from the FCC unit into commercially valuable products such as gasoline, fuel gas and olefin (which is a feed stream to downstream alkylation unit). The overhead vapors (stream 6) with some amount of overhead liquid from fractionator overhead receiver pass through a series of trim coolers and pass to a product accumulator. The accumulator separates the liquid and vapor from the trim coolers and ensures only the vapor stream to be sent to the wet gas compressor. The compressed vapor goes to an interstage cooler and the remaining vapor proceeds to the second stage of the compressor. Compressed vapors from the compressor (stream 13) and rich gasoline stream from the upper deethanizer section (stream 18) are sent to the deethanizer feed cooler. Cooled and partially compressed stream from the cooler gets separated into 3 streams (hydrocarbon vapors, gasoline and condensed sour water) after passing through the feed separator. The hydrocarbon vapor stream (stream 16) flows to the bottom of the absorber section of the deethanizer. The third liquid stream (stream 17) which is non stabilized gasoline enters the top tray of the lower deethanizer section. A part of non-stabilized gasoline entering the lower deethanizer section is partially vaporized by a reboiler which draws liquid from bottom section (stream 21) of the deethanizer section. The rate of vaporization is controlled by a temperature controller (TC) , which measures liquid temperature from the downspout of tray 9 of the 20 tray absorber section. It is usually maintained at 250F to ensure low content of the ethane in the bottom product (stream 19) of the deethanizer column which goes to the alkylation unit. The TC can also get its signal from deethanizer reboiler

outlet temperature.

The deethanizer tower is prone to frequent flooding followed by weeping incidents. One of the reasons for flooding which can be concluded from time trends of the variables is that, the increase in reboiler duty increases the temperature profile of the de-ethanizer column bottom. The increased reboiler duty will increase the temperature of the stream entering the de-ethanizer bottom, which in turn increases the temperature of the reboiler product stream (TI_{RB_P}) and de-ethanizer bottom (TI_{DE_O}). Increase in the temperature profile in the de-ethanizer tower increases the vapor flow up the column. When the vapor flow exceeds a threshold value, flooding occurs. After flooding the controller in the de-ethanizer tower tries to bring down the tower bottom temperature (indicated by two temperature indicators, TI_{TB1} and TI_{TB2}), which causes a reduction in vapor pressure in the tower and in turn leads to weeping in the de-ethanizer tower.

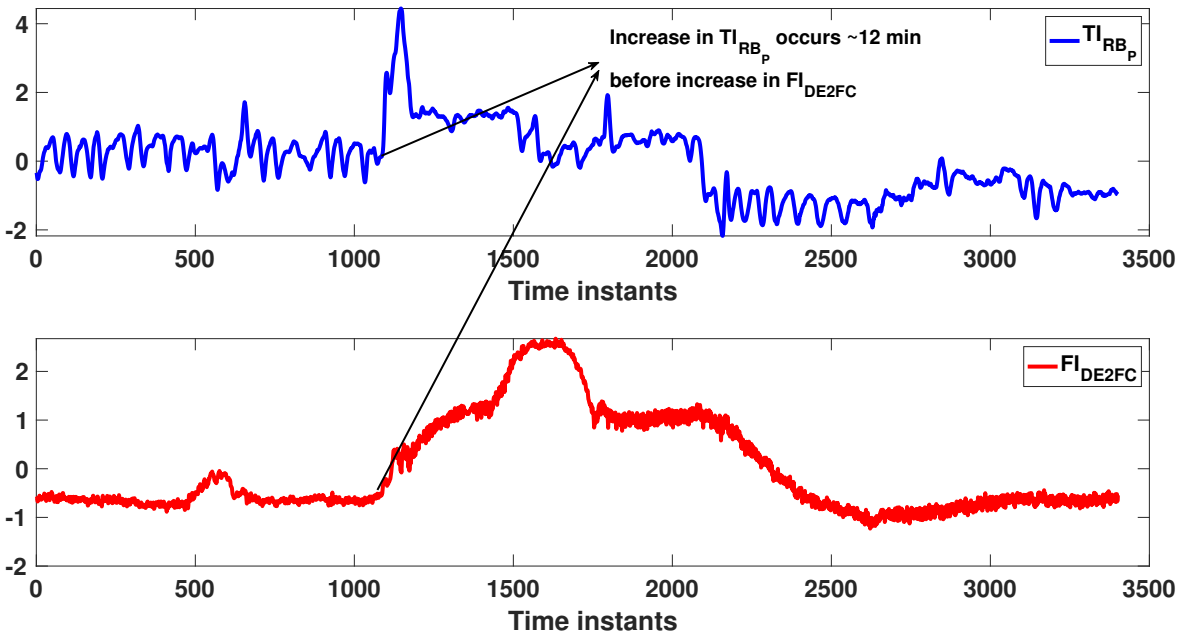

Figure 5.2: Time trends of TI_{RB_P} and FI_{DE2FC} during flooding event

Causality methods can be used to find the root causes of such flooding events as such a study can give the cause and effect graph which illustrates the causal

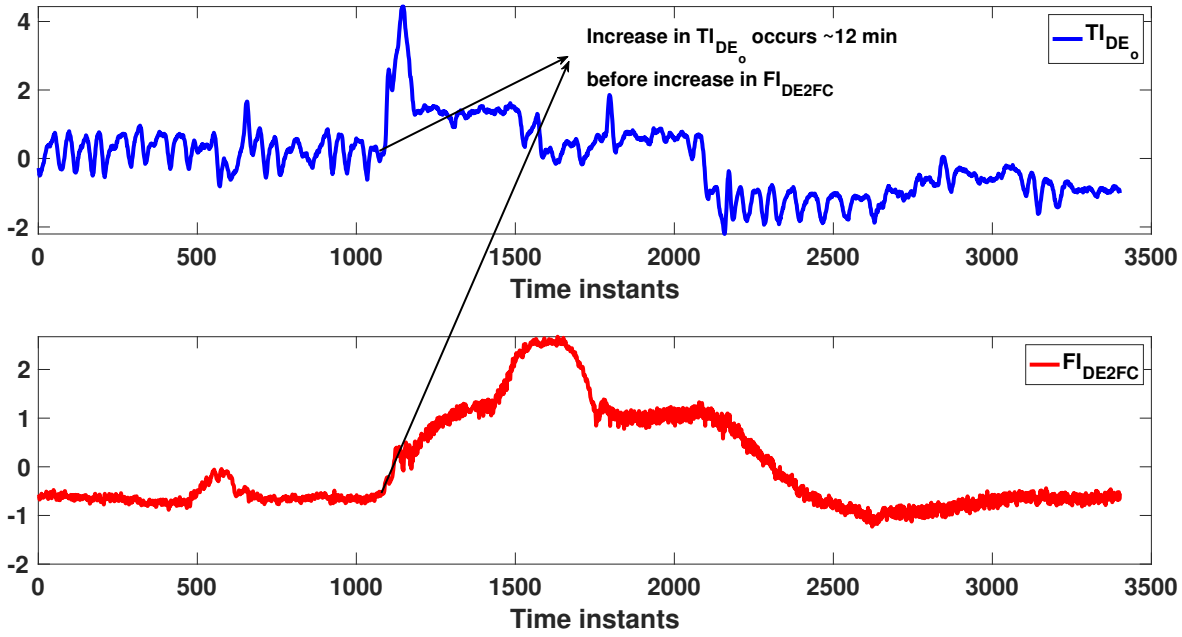

Figure 5.3: Time trends of TI_{DE_O} and FI_{DE2FC} during flooding event

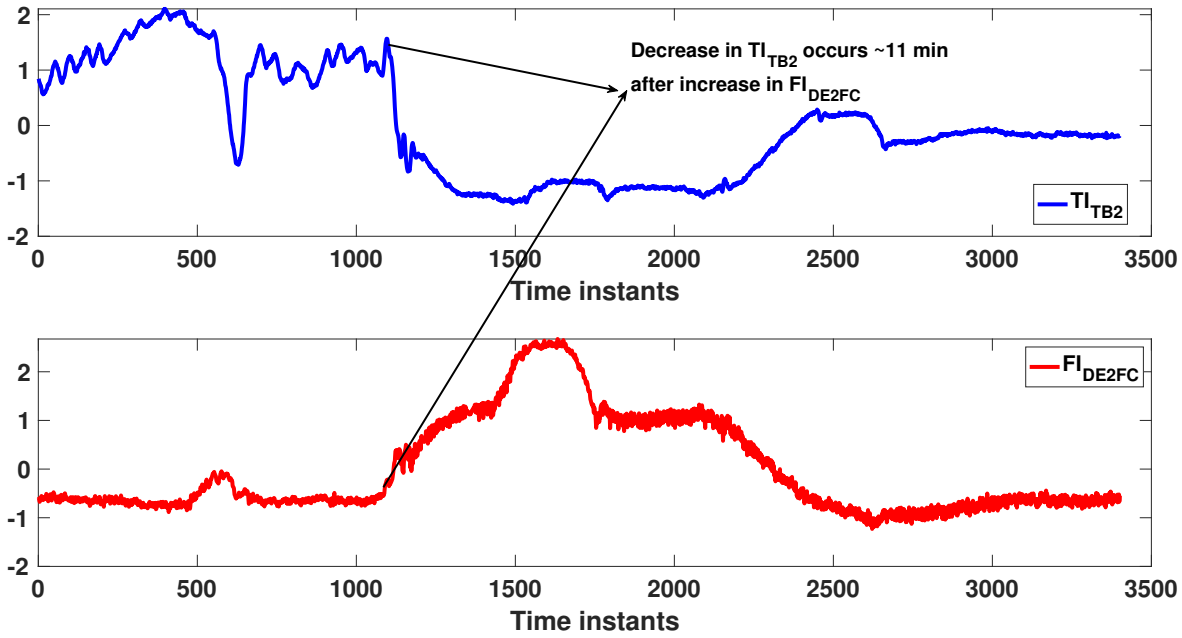

Figure 5.4: Time trends of TI_{TB2} and FI_{DE2FC} during flooding event

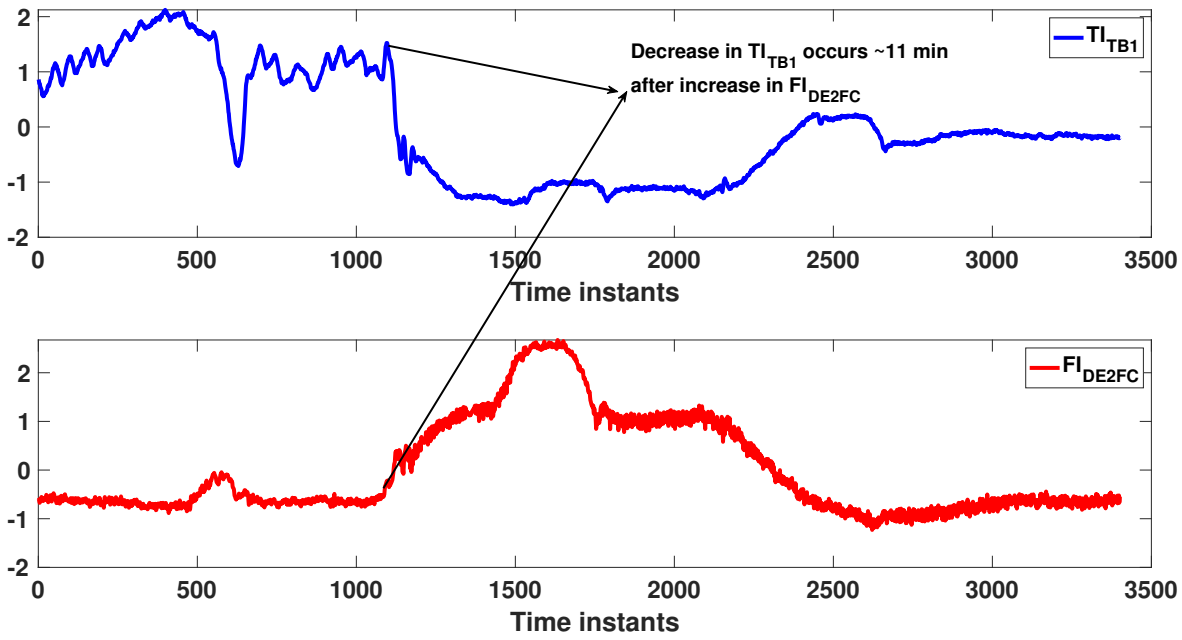

Figure 5.5: Time trends of TI_{TB1} and FI_{DE2FC} during flooding event

relationships among the variables. In this chapter we try to identify the cause and effect relations of a set of variables (given in Table 5.1) chosen from the deethanizer column with the flooding indicator (FI_{DE2FC}), which indicates the flow from upper deethanizer section to deethanizer feed cooler (stream 18). When this flow exceeds a certain value, it indicates flooding. The 12 variables chosen are temperature, flow and pressure indicators inside and immediately around the deethanizer column .

5.2 Data

Flooding data for the month of October 2018 with a sampling interval of 1 min from a refinery was available for analysis. For proprietary reason, all data have been normalized. The following steps were carried out in sequence as part of the data pre-processing: 1. Removal of missing data 2. Removal of NaN values 3. Removal of data points over the 3 sigma bounds . When the value of flow indicator (FI_{DE2FC}) which measures the flow of de-ethanizer feed to the deethanizer feed cool-

Table 5.1: List of selected variables for causality analysis.

No	Variable	Description
1	TI_{ABS_T}	TI of stream from top of the absorber
2	FI_{DE2FC}	FI in the stream from deethanizer to deethanizer feed coolers
3	FI_{FSA}	FI in the input stream to the absorber from the feed separator
4	TI_{ABS_B}	TI in the stream from bottom of the absorber
5	TI_{DE_T}	TI in the Stream from top of deethanizer
6	FI_{ABS2FS}	FI in the stream from absorber to feed separator
7	TI_{RB_P}	TI in the reboiler product stream
8	TI_{TB1}	First TI in the deethanizer tower bottom
9	TI_{TB2}	Second TI in the deethanizer tower bottom(above TI_{TB1})
10	TI_{DE_O}	TI in the outlet stream from the extreme bottom of deethanizer
11	PI_{DE}	PI at Deethanizer tray 1
12	TI_{DE2DP}	FI in the deethanizer bottom stream to depentanizer

ers crosses 16, it indicates flooding. A continuous data for 3400 time instants which contains data during both normal operation and flooding condition was chosen for the causality study. The causal relations among the variables were obtained using the Multi-Variate Granger Causality (MVGC) toolbox [21] and the proposed multi-mode causality method. Sixth order VAR models were considered for both the MVGC toolbox and the traditional Granger causality method based on the AIC criteria.

5.3 Results and discussions

5.3.1 Results of multi-model method

The cause and effect relationship among the variables in the analysis is represented as 2D colour intensity plots.

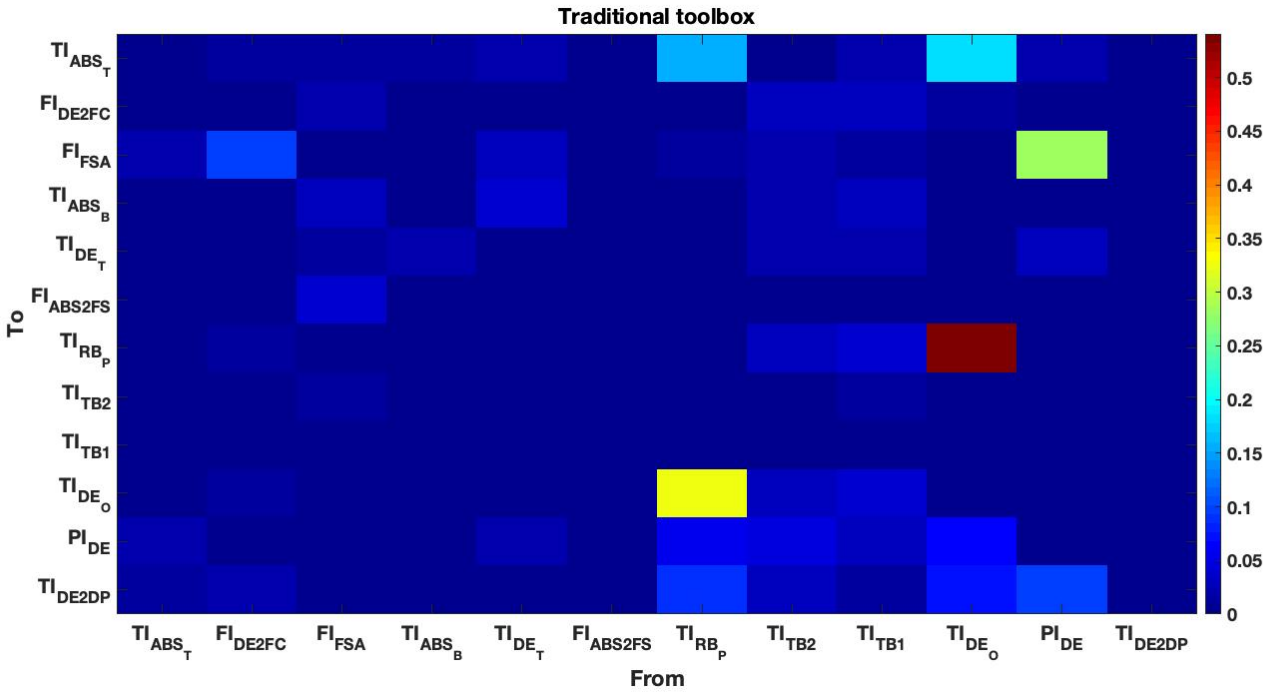

Figure 5.6: Traditional Granger Causality

The traditional method and MVGC toolbox were not able to capture the sequence of events leading to flooding. Both the MVGC toolbox and traditional method are based on the assumption that the data can be represented as a linear model satisfac-

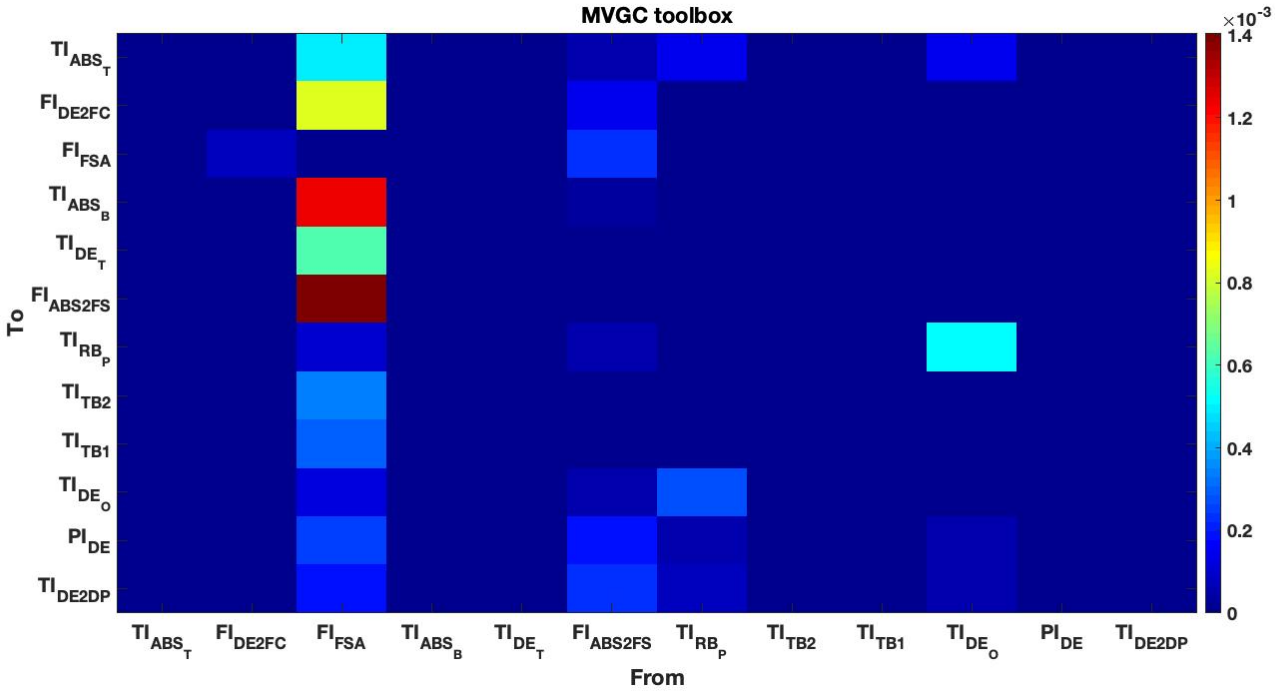

Figure 5.7: Multi Variate Granger Causality (MVGC) toolbox results

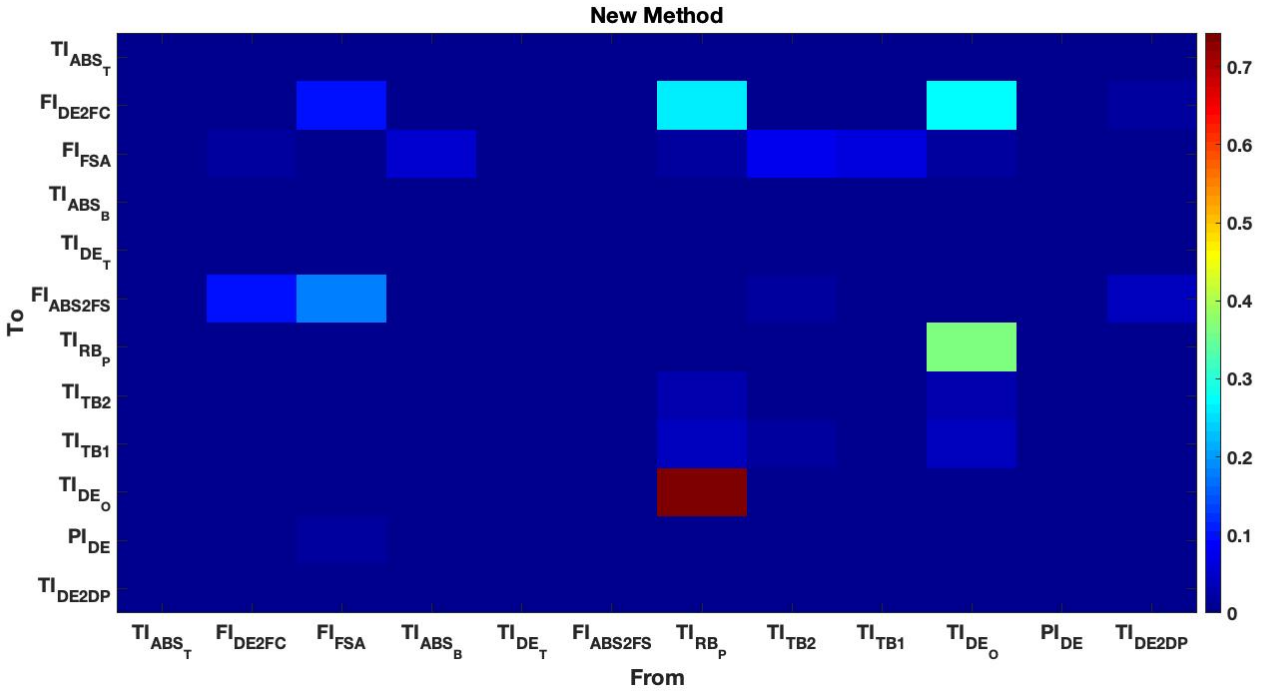

Figure 5.8: New method results

torily. Hence, they cannot handle data from processes which operate in more than one mode, which is the reason behind the poor performance of these methods. The causal intensity graph derived from the proposed method helps in the construction of the causal graph given in Fig 5.10.

A *ub* of 2 was chosen for our simulation. Fig 5.9 shows the variation of the significant coefficients with time. It was observed that only when the flooding is very high, the model switches to the second mode completely. Otherwise, the model is a mixture of the two VAR models considered.

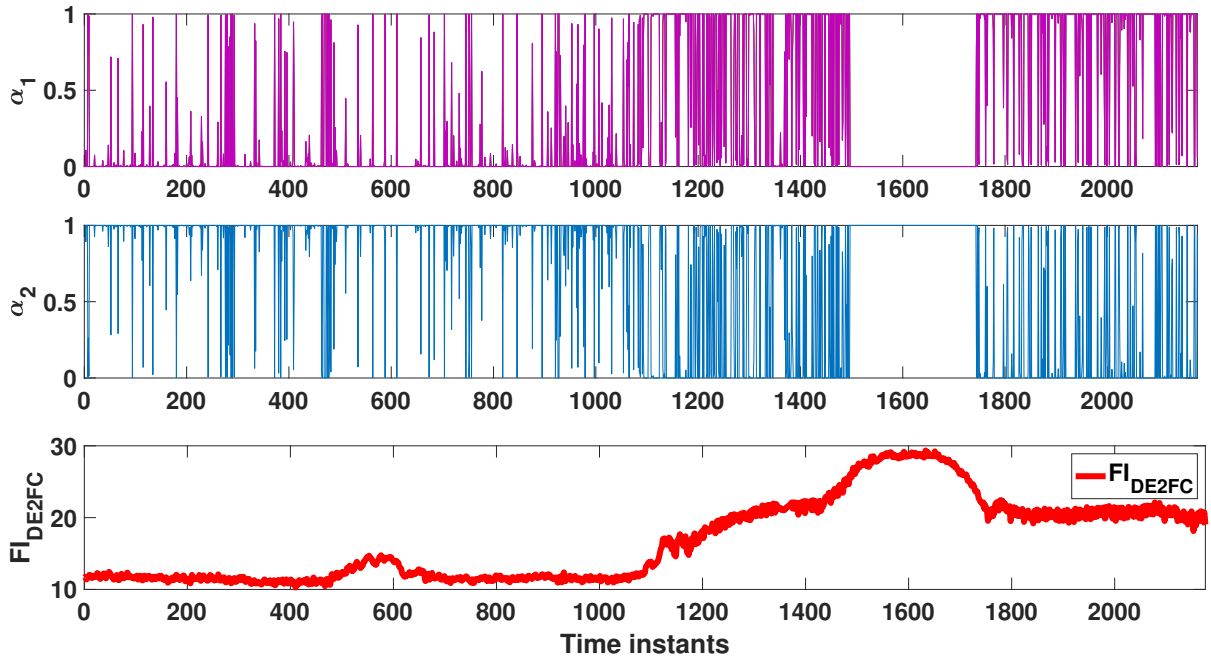

Figure 5.9: Significant coefficients and flooding indicator plots

The causal graph helps to conclude that, an increase in the reboiler duty increases the temperature profile of the bottom of the de-ethanizer column. The increased reboiler duty will increase the temperature of the stream entering the de-ethanizer bottom, which in turn would increase the temperature of the reboiler product stream (TI_{RB_P}) and de-ethanizer bottom temperature (TI_{DE_O}). The increase in the temperature profile in the de-ethanizer tower increases the vapor flow inside the column.

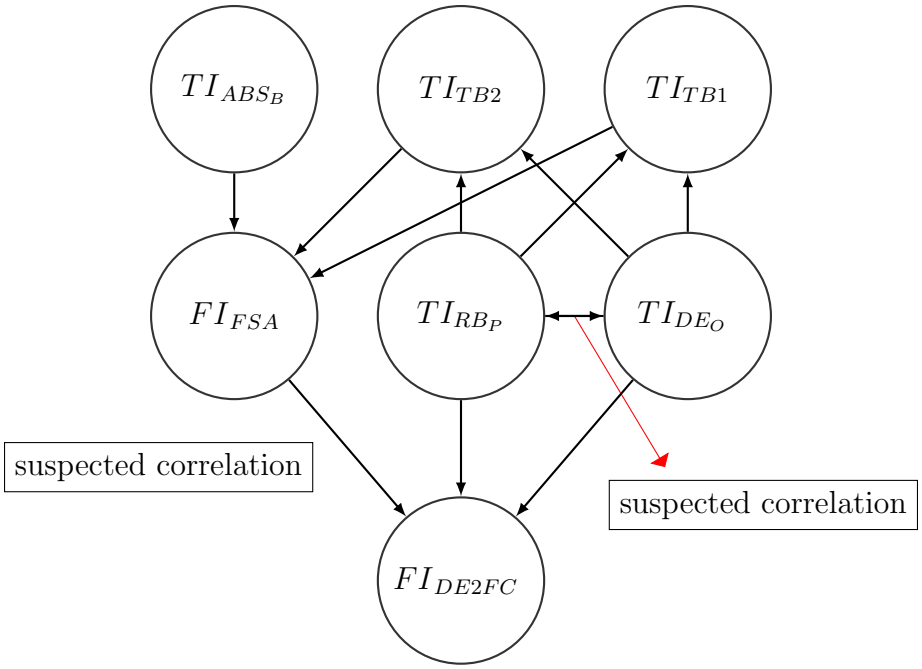

Figure 5.10: Identified causal graph

When the vapor flow exceeds a threshold value, flooding occurs, which is indicated by the flooding indicator. The reboiler product temperature (TI_{RB_P}) and de-ethanizer bottom temperature (TI_{DE_O}) cause change in TI_{TB1} and TI_{TB2} . From process knowledge, it is understood that the flooding event is followed by weeping of the de-ethanizer tower. This explains why TI_{RB_P} and TI_{DE_O} cause change in TI_{TB1} and TI_{TB2}. Decrease in TI_{TB1} and TI_{TB2} is indication of weeping. Now, a third variable, FI_{FSA} appears to be a causal variable of flooding. The weeping indicators (TI_{TB1} and TI_{TB2}) and temperature indicator TI_{ABS_B} seem to be the causal variables of FI_{FSA}. However, from time trends it is observed that the changes in TI_{ABS_B} and FI_{FSA} happen after the flooding event. The proposed method is not able capture this causal relation accurately. A possible reason for this being the presence of an unaccounted variable which is correlated to FI_{FSA} and is not considered in the causality analysis. This unaccounted variable might be a possible causal variable, however since it is not considered in the causality analysis, the variable correlated to it (here FI_{FSA}) is wrongly concluded to be a causal variable.

Now, we segregate the data to flooding and non-flooding data, and use only the flooding data to perform traditional Granger causality technique (Fig 5.11) and method using the MVGC toolbox (Fig 5.12). TI_{RB_P} which is the primary causal variable, now appears in the causal graph obtained using traditional Granger causality technique. However, the MVGC toolbox is still unable to identity the causal variables. We chose the best model order (using AIC criteria) from a upper limit of 10, maybe a higher upper limit has to be set for selecting the best model order for the MVGC toolbox.

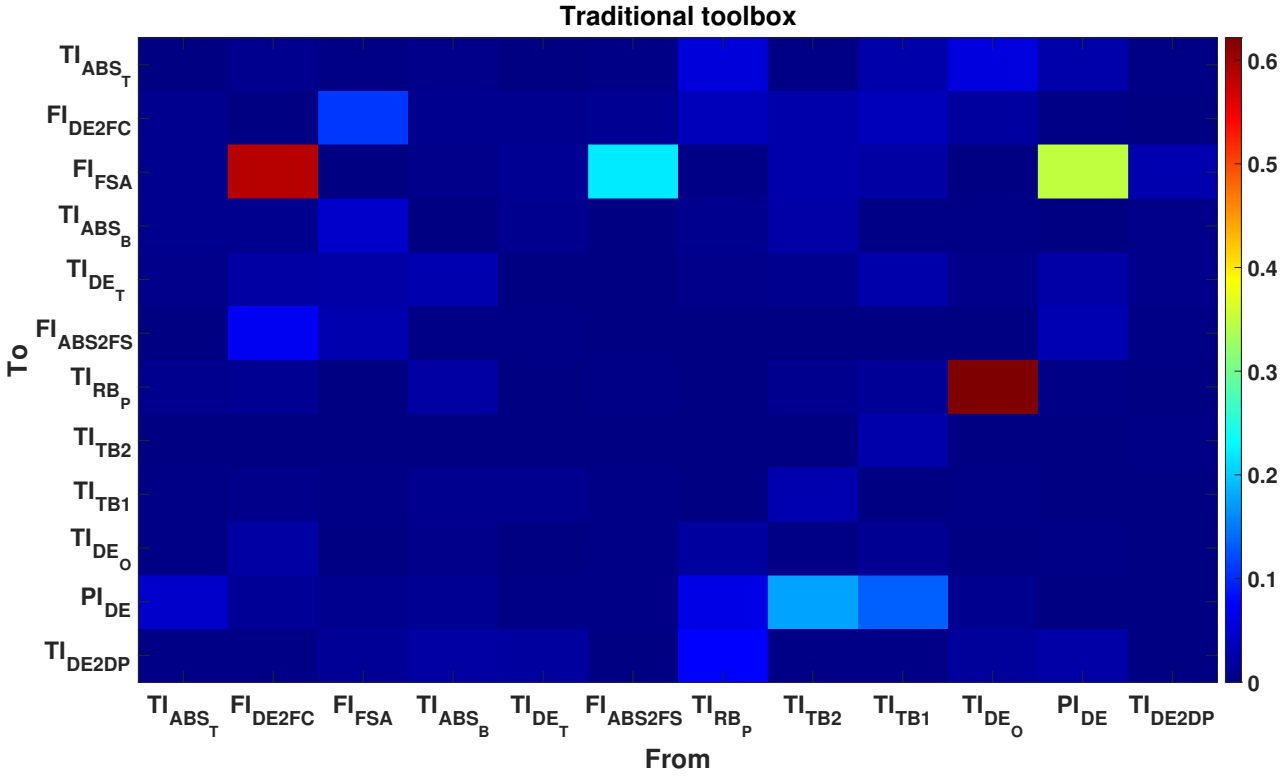

Figure 5.11: Traditional Granger causality results using only flooding data

Remarks

When compared to the traditional and MVGC toolbox results, the results from the new method are more accurate and aligned with the process knowledge of the system. It can be concluded that an increase in reboiler duty leads to an increase in the

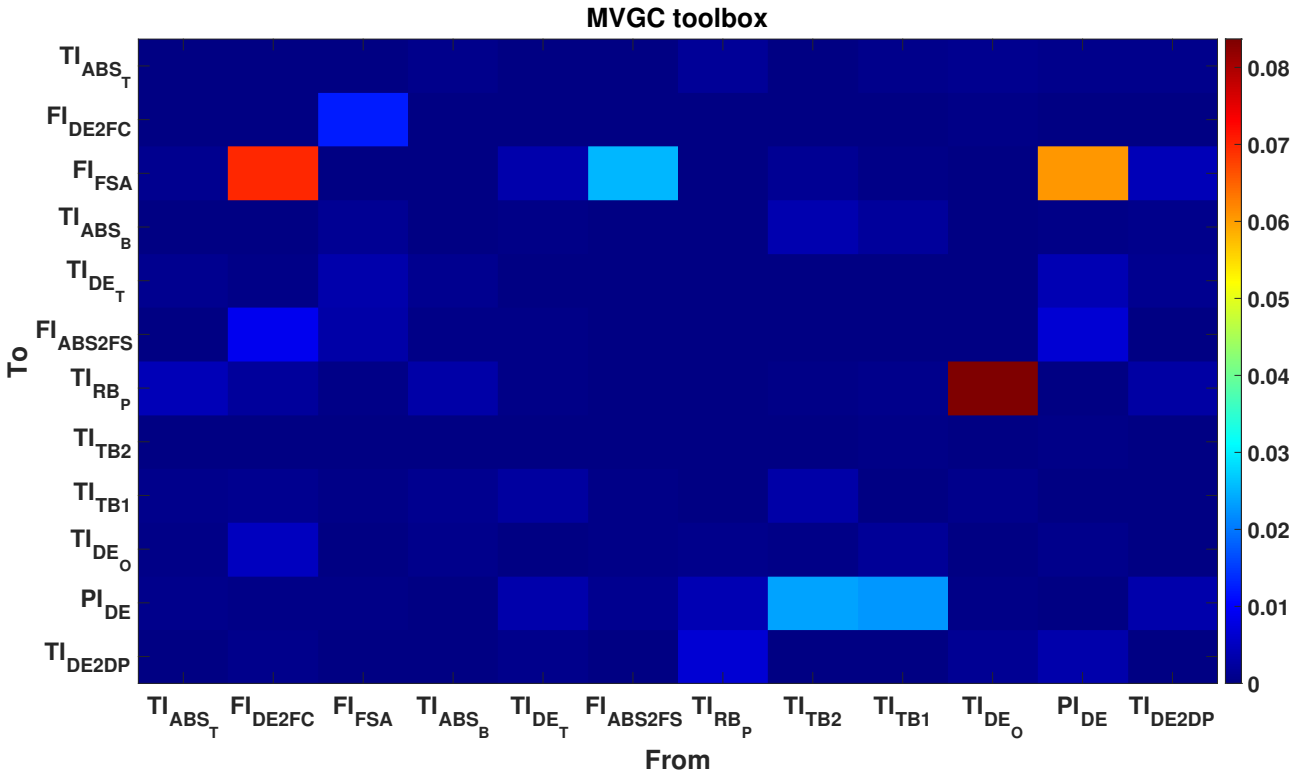

Figure 5.12: Multi Variate Granger Causality (MVGC) toolbox results using only flooding data

deethanizer bottom temperature which in turn leads to an excess vapor production causing flooding.

5.3.2 Comparison study of the two methods in presence of outliers

In this section, a comparison of the performances of the two methods mentioned in the previous chapters is done using the same industrial process data set. The actual industrial data did not have any outliers, hence outliers were introduced artificially to check the robustness of the two proposed methods. We tried to check if the robust method was able to infer the root causes of flooding in the situation when the process data is contaminated by outliers. Real industrial flooding data for the month of October 2018 with the maximum number of flooding events was chosen again. The sampling interval was 1 min. We performed data-preprocessing steps to remove missing data. For performing the causality analysis, the same set of 12 variables (table 5.1) in and around the deethanizer column were chosen. 3400 continuous data points containing both flooding and normal operation data were collected for analysis. Then, 20 percentage of the data was replaced by data points which lie outside the 5 sigma bounds of the original data. The causality matrices obtained from implementing the two proposed method on the data are shown in Fig. 5.13 and 5.14.

Clearly, there is a difference between the results obtained using the two methods. The robust method is able to identify the root cause of flooding more accurately in the presence of outliers. The increase in the reboiler product temperature TI_{RB_P}, which in turn is caused due to an increase in the reboiler duty was the root cause of the flooding. The multi-model method is unable to find the root cause of flooding as it cannot handle data contaminated with outliers.

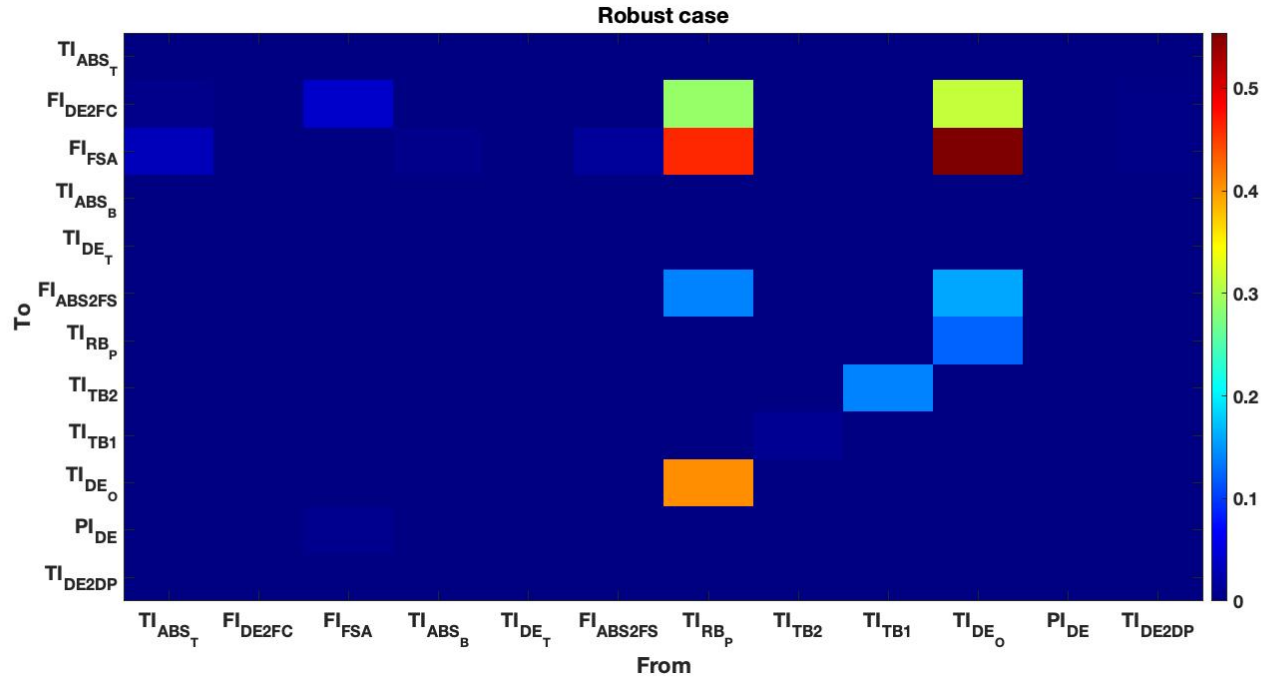

Figure 5.13: Robust multi-model method where the noise of the prediction model has a t-distribution

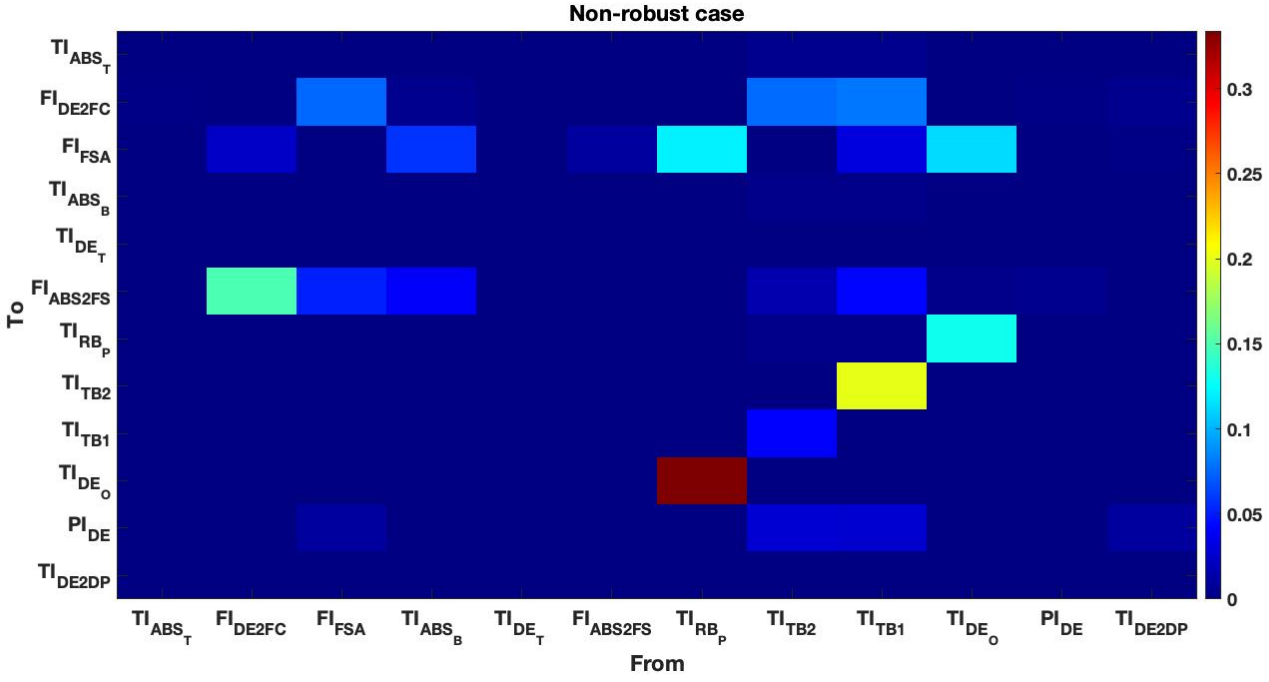

Figure 5.14: Multi-model method where the noise of the prediction model has a Gaussian distribution

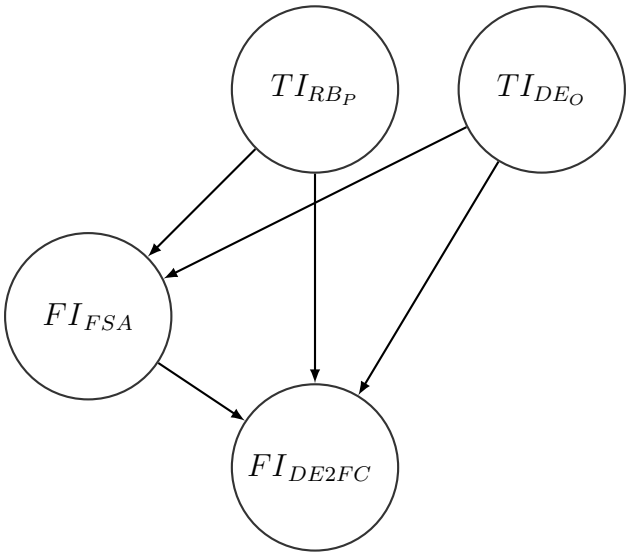

Figure 5.15: Causal graph extracted using the robust multi-model method

Chapter 6

Conclusions

6.1 Summary of the research

This thesis developed a new framework for causality analysis for multi-modal processes using Bayesian networks which are a special class of probabilistic graphical models. The two main methods developed and the findings from these methods can be summarized as follows:

1. In chapter 3, we developed a method for Granger causality analysis for multi-model processes. As the process has more than one mode of operation, a single VAR model is not enough to model the process. We developed a Granger causality technique with multi-mode VAR model using a variational Bayesian approach such that the causal structures extracted from different modes are consistent. It was achieved by assuming the corresponding elements of the coefficient matrices from all the modes to come from the same Gaussian distribution with precision parameter following a gamma distribution. This introduces same sparsity in coefficient matrices across modes. The proposed method also provides a simpler statistical test for checking the significance of the causal connections. The theoretical findings were confirmed using simulation process data.

2. In chapter 4, we extended our proposed work to handle data contaminated

with outliers. Data from real industrial process systems is often corrupted with outliers. These outliers, if not handled properly, can greatly reduce the performance of causal analysis. Robust extension of the earlier work is able to handle outliers efficiently. Outliers in the data are handled by assuming a t-distribution for the model residuals. Variational Bayesian approach is once again used for parameter estimation. Simulation case studies were used to verify the performance of the method.

3. In chapter 5, a comparative study of the two methods was done using data from a deethanizer column tower. The data consists of both flooding and normal operation data. However, the actual data did not have any outliers in it, hence we introduced outliers manually (20 percentage). The robust method was able to detect the events of flooding correctly in the presence of outliers unlike the multi-model case.

6.2 Directions for future work

1. The robust extension of the proposed method is assumed to be resistant only against outliers. The method can be extended for missing data as well.

2. In this method, we assume a BN initially. In its place, a method to propose the best possible BN structure can be developed.

3. In the proposed methods, we assumed that the causal structure did not change with time. However, this is not always true in real systems. The causal structures can be time dependent. In such cases, a time varying BN structure needs to be used.

Bibliography

[1] L. Baccalá and K. Sameshima, "Partial directed coherence:a new concept in neural structure determination," *Biol Cybern*, vol. 84, pp. 463–474, 2001.

[2] R.Alquist, L.Kilian, and R. J. Vigfusson, "Handbook of economic forecasting," in, A. T. G. Elliott, Ed. Elsevier, 2013, vol. 2, ch. Forecasting the Price of Oil, pp. 427–507.

[3] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[4] G. Wang and M. Takigawa, "Directed coherence as a measure of interhemispheric correlation of EEG," *International Journal of Psychophysiology*, vol. 13, pp. 119–128, 1992.

[5] S. Kullback, "Information theory and statistics," 1959.

[6] E. Naghoosi, B. Huang, E. Domlan, and R. Kadali, "Information transfer methods in causality analysis of process variables with an industrial application," *Journal of Process Control*, vol. 23, pp. 1296–1305, 2013.

[7] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, "Partial Granger causality—Eliminating exogenous inputs and latent variables," *Journal of Neuroscience Methods*, vol. 172, pp. 79–93, 2008.

[8] W. Hesse, E. Möller, M. Arnold, and B. Schack, "The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies," *Journal of neuroscience methods,*, vol. 124, 27–44, 1 2003.

[9] J. Geweke, "Measures of conditional linear dependence and feedback between time series," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 907–915, 1984.

[10] Y. Saito and H. Harashima, "Tracking of information within multichannel EEG record," 1981.

[11] M. Kamiński and K. Blinowska, "A new method of the description of the information flow in the brain structures," *Biol Cybern*, vol. 65(3), pp. 203–210, 1991.

[12] M. Ding, Y. Chen, and S. L. Bressler, "Granger Causality: Basic Theory and Application to Neuroscience," *Handbook of Time Series Analysis*, 2006.

[13] W. A. Freiwald, P. Valdes, J. Bosch, R. Biscay, and J. Jimenezand L.M. Rodriguezand V. Rodriguezand A.K. Kreiterand W. Singer, "Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex," *Journal of neuroscience methods,*, vol. 94, 105–119, 1999.

[14] T. Schreiber, "Measuring Information Transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, 2 2000. DOI: 10.1103/PhysRevLett.85.461. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.85.461.

[15] K. Hlaváckova-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, pp. 1–46, 1 2007.

[16] M. Palus, V. Komárek, Z. Hrncír, and K. Sterbová, "Synchronization as adjustment of information rates: Detection from bivariate time series," *Phys. Rev. E*, vol. 63, p. 046 211, 4 2001. DOI: 10.1103/PhysRevE.63.046211. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.63.046211.

[17] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear Granger causality," *Phys. Rev. Lett.*, vol. 100, p. 144 103, 14 2008. DOI: 10.1103/PhysRevLett.100.144103. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.100.144103.

[18] R. Kannan and A. K. Tangirala, "Correntropy-based partial directed coherence for testing multivariate Granger causality in nonlinear processes," *Phys. Rev. E*, vol. 89, p. 062 144, 6 2014. DOI: 10.1103/PhysRevE.89.062144. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.89.062144.

[19] N. Ancona, D. Marinazzo, and S. Stramaglia, "Radial basis function approach to nonlinear Granger causality of time series," *Phys. Rev. E*, vol. 70, p. 056 221, 5 2004. DOI: 10.1103/PhysRevE.70.056221. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.70.056221.

[20] X. Sun, W. Daelemans, B. Goethals, and K. Morik, "Assessing nonlinear Granger causality from multivariate time series," Sep. 2008. DOI: 10.1007/978-3-540-87481-2_29.

[21] L. Barnett and A. K. Seth, "The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference," *Journal of Neuroscience Methods*, vol. 223, pp. 50–68, 2014.

[22] M.J.Beal, "Variational algorithms for approximate bayesian inference," PhD thesis, University of London, 2003.

[23] M. B. Christopher, "Pattern recognition," in. New York, NY, USA: Springer-Verlag, ch. Approximate inference.

[24] R. Raveendran and B. Huang, "Variational Bayesian approach for causality and contemporaneous correlation features inference in industrial process data," *IEEE Transactions on Cybernetics*, vol. 49, no. 7, pp. 2580–2590, 2019.

[25] L. Kalliovirta, M. Meitz, and P. Saikkonen, "Gaussian mixture vector autoregression," English, *HECER discussion papers*, no. 386, pp. 1–37, Nov. 2014, ISSN: 1795-0562.

[26] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[27] P. J. Huber, *Robust Statistics*. John Wiley Sons,NY, 1981.

[28] S. Khatibisepehr and B. Huang, "A Bayesian approach to robust process identification with ARX models," *AIChE Journal*, vol. 59, no. 3, pp. 845–859, 2013.

[29] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.

[30] H.Kodamana, B. Huang, R. Ranjan, Y. Zhao, R. Tan, and N. Sammaknejad, "Approaches to robust process identification: A review and tutorial of probabilistic methods," *Journal of Process Control*, vol. 66, pp. 68–83, 2018.

[31] K. Lange, R. Little, and J. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989, cited By 802. DOI: 10.1080/01621459.1989.10478852. [Online]. Available: https://www2.scopus.com/inward/record.uri?eid=2-s2.0-84950441032&doi=10.1080%2f01621459.1989.10478852&partnerID=40&md5=26fea885716fd41904cb58f923a586e6.

[32] J. Zhu, Z. Ge, and Z. Song, "Robust semi-supervised mixture probabilistic principal component regression model development and application to soft sensors," *Journal of Process Control*, vol. 32, pp. 25 –37, 2015, ISSN: 0959-1524. DOI: https://doi.org/10.1016/j.jprocont.2015.04.015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959152415000906.

[33] D. G.MacLachlan, *Finite mixture models Wiley series in Probability and Statistics*. 2000.

[34] P. Li, X. Huang, F. Li, X. Wang, W. Zhou, H. Liu, T. Ma, T. Zhang, D. Guo, D. Yao, and P. Xu, "Robust Granger analysis in Lp norm space for directed EEG network analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, pp. 1959–1969, 2017.

[35] T. Schäckand, M. Muma, M. Feng, C.Guan, and A.Zoubir, "Robust nonlinear causality analysis of non-stationary multivariate physiological time series," *IEEE Transactions on Biomedical Engineering*, vol. PP, pp. 1–1, May 2017. DOI: 10.1109/TBME.2017.2708609.

[36] A. Fujita, K. Kojima, A. Patriota, J. Sato, P. Severino, and S. Miyano, "A fast and robust statistical test based on likelihood ratio with bartlett correction to identify granger causality between gene sets," *Bioinformatics (Oxford, England)*, vol. 26, pp. 2349–51, Sep. 2010. DOI: 10.1093/bioinformatics/btq427.

Appendix A: First Appendix

The expressions for lower bound is given in Table A.1. The posterior distributions with their updated parameters are given in A.2.

Table A.1: Lower Bound Expression

$$
\begin{aligned}
L \geq \quad & \frac{1}{2}\sum_{s=1}^{ub}\sum_{d=1}^{D}\ln|\Sigma_{\hat{W}_d{}^s}| + \frac{ub}{2}\sum_{d=1}^{D}\sum_{m=1}^{M}(\psi(a)-\ln b_{dm}) - \frac{1}{2}\sum_{s=1}^{ub}\sum_{d=1}^{D} tr[\lambda_d(\Sigma_{\hat{W}_d{}^s} + (\hat{W}_d^{s\,T}\hat{W}_d^s)] \\
& + 0.5(ubDM) - a\sum_{d=1}^{D}\sum_{m=1}^{M}\ln b_{dm} + DMa^*\ln b^* + DM\ln\frac{\Gamma(a)}{\Gamma(a^*)} \\
& - (a-a^*)\sum_{d=1}^{D}\sum_{m=1}^{M}(\Psi(a)-\ln b_{dm}) + a\sum_{d=1}^{D}\sum_{m=1}^{M}(1-\frac{b^*}{b_{dm}}) \\
& + \sum_{t=1}^{N}\sum_{s}^{ub}\alpha_{new}^{S(t)=s}[\Psi(\alpha_{new}^* m_S) - \Psi(\alpha_{0,new}^*) - \ln\alpha_{new}^{S(t)=s}] + \ln\frac{\alpha_0^*}{\alpha_{0,new}^*} \\
& + \sum_{s=1}^{ub}\left[\ln\frac{\alpha_{new}^* m_S}{\alpha^* m_S^*} + (\alpha_{new}^* m_S - \alpha^* m_S^*)(\Psi(\alpha_{new}^* m_S) - \Psi(\alpha_{0,new}^*))\right] - \frac{DN}{2}\ln 2\pi \\
& + \frac{D}{2}\sum_{s=1}^{ub}\sum_{t=1}^{N}\alpha_{new}^{S(t)=s}\ln\delta^s - \sum_{s=1}^{ub}\sum_{t=1}^{N}\alpha_{new}^{S(t)=s}\frac{\delta^s}{2}tr[y(t)y(t)^T] \\
& + \sum_{s=1}^{ub}\sum_{t=1}^{N}\alpha_{new}^{S(t)=s}\sum_{d=1}^{D}\delta^s y(t)_d\hat{W}_d^s y(t-1) \\
& - \sum_{s=1}^{ub}\sum_{t=1}^{N}\alpha_{new}^{S(t)=s}\sum_{d=1}^{D}\frac{\delta^s}{2}tr[(\Sigma_{\hat{W}_d^s} + (\hat{W}_d^s)^T\hat{W}_d^s)(y(t-1)y(t-1)^T)]
\end{aligned}
$$

where $\Sigma_{W_d} = diag([{\beta_{d1}}^{-1}, {\beta_{d2}}^{-1}, ..., {\beta_{dM}}^{-1}])$, $\lambda_d = diag([\frac{a}{b_{d1}}, \frac{a}{b_{d2}}, ..., \frac{a}{b_{dM}}])$ and $\alpha_{0,new}^* = \sum_{s=1}^{ub}\alpha_{new}^* m_S$

Table A.2: Update equations

Distribution	**Parameters**
$q(W_d^s \mid \hat{W}_d^s, \Sigma_{\hat{W}_d^s})$	$\Sigma_{\hat{W}_d{}^s} = [\lambda_d + \delta^s \sum_{t=1}^N \alpha_{new}^{S(t)=s} y(t-1)y(t-1)^T]^{-1}$,
	$\hat{W}_d^s = \Sigma_{\hat{W}_d{}^s}[\sum_{t=1}^N \alpha_{new}^{S(t)=s} \delta^s y_d(t) y(t-1)]$
$q(\beta_{dm} \mid a, b_{dm})$	$a = a^* + \frac{ub}{2}, b_{dm} = b^* + \frac{1}{2}\sum_{s=1}^{ub}[\hat{W}_{dm}^s{}^2 + \Sigma_{\hat{W}_{dm}^s}]$
$q(\alpha^s \mid \alpha_{new}^* m_s)$	$\alpha_{new}^* m_s = \alpha^* m_s^* + \sum_{t=1}^N q(S(t)=s)$
$q(S(t)=s) = \alpha_{new}^{S(t)=s}/z$	Z is normalizing constant

where $\alpha_{new}^{S(t)=s} = e^P$, such that

$$\mathrm{P}=\Psi(\alpha_{new}^* m_s) - \Psi(\alpha_{0,new}^*) - \frac{D}{2}\ln(2\pi) + \frac{D}{2}\ln\delta^s - \frac{\delta^s}{2} tr[y(t)y(t)^T] + \sum_{d=1}^D \delta^s y(t)_d \hat{W}_d^s y(t-1)$$
$$\text{-}\sum_{d=1}^D \frac{\delta^s}{2} tr[(\Sigma_{\hat{W}_d^s} + (\hat{W}_d^s)^T \hat{W}_d^s)(y(t-1)y(t-1)^T)]$$

$$Z = \sum_{s=1}^{ub} \alpha_{new}^{S(t)=s}, \quad \Sigma_{W_d} = diag\left(\left[{\beta_{d1}}^{-1}, {\beta_{d2}}^{-1}, ..., {\beta_{dM}}^{-1}\right]\right), \lambda_d = diag\left(\left[\frac{a}{b_{d1}}, \frac{a}{b_{d2}}, ..., \frac{a}{b_{dM}}\right]\right)$$

$$\delta^s = \frac{D * \sum_{t=1}^N \alpha_{new}^{S(t)=s}}{2} \Bigg(\sum_{t=1}^N \frac{\alpha_{new}^{S(t)=s}}{2} tr(y(t)y(t)^T - \sum_{t=1}^N \alpha_{new}^{S(t)=s} \sum_{d=1}^D y_d(t) \hat{W}_d^s y(t-1)$$
$$+ \sum_{t=1}^N \alpha_{new}^{S(t)=s} \sum_{d=1}^D \frac{1}{2} tr\left[(\Sigma_{\hat{W}_d^s} + (\hat{W}_d^s)^T \hat{W}_d^s)(y(t-1)y(t-1)^T)\right]\Bigg)^{-1}$$

Appendix B: Second Appendix

The expressions for lower bound is given in Table B.1. The posterior distributions with their updated parameters are given in B.2.

Table B.1: Lower Bound Expression

$$
\begin{aligned}
L \geq \quad & \frac{1}{2}\sum_{s=1}^{ub}\sum_{d=1}^{D}\ln|\Sigma_{\hat{W_d}^s}| + \frac{ub}{2}\sum_{d=1}^{D}\sum_{m=1}^{M}(\psi(a)-\ln b_{dm}) + 0.5(ubDM) \\
& - \frac{1}{2}\sum_{s=1}^{ub}\sum_{d=1}^{D} tr[\lambda_d(\Sigma_{\hat{W_d}^s} + (\hat{W}_d^{sT}\hat{W}_d^s)] - a\sum_{d=1}^{D}\sum_{m=1}^{M}\ln b_{dm} + DMa^*\ln b^* + DM\ln\frac{\Gamma(a)}{\Gamma(a^*)} \\
& - (a-a^*)\sum_{d=1}^{D}\sum_{m=1}^{M}(\Psi(a)-\ln b_{dm}) + a\sum_{d=1}^{D}\sum_{m=1}^{M}(1-\frac{b^*}{b_{dm}}) + \ln\frac{\alpha_0^*}{\alpha_{0,new}^*} \\
& + \sum_{t=1}^{N}\sum_{s=1}^{ub}\alpha_{new}^{S(t)=s}[\Psi(\alpha_{new}^* m_S) - \Psi(\alpha_{0,new}^*) - \ln\alpha_{new}^{S(t)=s}] \\
& + \sum_{s=1}^{ub}\left[\ln\frac{\alpha_{new}^* m_s}{\alpha^* m_s^*} + (\alpha_{new}^* m_s - \alpha^* m_s^*)(\Psi(\alpha_{new}^* m_s) - \Psi(\alpha_{0,new}^*))\right] \\
& - \sum_{s=1}^{ub}\sum_{t=1}^{N}\nu_\alpha^{S(t)=s}\ln\nu_\beta^{S(t)=s} + \sum_{s=1}^{ub}N\frac{\nu^S}{2}\ln\frac{\nu^S}{2} + \sum_{s=1}^{ub}\sum_{t=1}^{N}\ln\frac{\Gamma(\nu_\alpha^{S(t)=s})}{\Gamma(\frac{\nu^s}{2})} \\
& - \sum_{s=1}^{ub}\sum_{t=1}^{N}(\nu_\alpha^{S(t)=s} - \frac{\nu^s}{2})(\Psi(\nu_\alpha^{S(t)=s}) - \ln\nu_\beta^{S(t)=s}) \\
& + \sum_{s=1}^{ub}\sum_{t=1}^{N}\nu_\alpha^{S(t)=s}(1 - \frac{\frac{\nu^s}{2}}{\nu_\beta^{S(t)=s}}) - \frac{DN}{2}\ln 2\pi + \frac{D}{2}\sum_{s=1}^{ub}\sum_{t=1}^{N}\alpha_{new}^{S(t)=s}\ln\delta^{S(t)=s} \\
& - \sum_{s=1}^{ub}\sum_{t=1}^{N}\left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right)\alpha_{new}^{S(t)=s}\frac{\delta^{S(t)=s}}{2}tr[y(t)y(t)^T] \\
& + \sum_{s=1}^{ub}\sum_{t=1}^{N}\left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right)\alpha_{new}^{S(t)=s}\sum_{d=1}^{D}\delta^{S(t)=s}y(t)_d\hat{W}_d^s y(t-1) \\
& - \sum_{s=1}^{ub}\sum_{t=1}^{N}\left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right)\alpha_{new}^{S(t)=s}\sum_{d=1}^{D}\frac{\delta^{S(t)=s}}{2}tr[(\Sigma_{\hat{W}_d^s} + (\hat{W}_d^s)^T\hat{W}_d^s)(y(t-1)y(t-1)^T)] \\
& + \sum_{s=1}^{ub}\sum_{t=1}^{N}\alpha_{new}^{S(t)=s}\times \\
& \left[-\ln\Gamma(\frac{\nu^s}{2}) + \frac{\nu^s}{2}\ln\frac{\nu^s}{2} + (\frac{\nu^s}{2}-1)\left(\Psi(\nu_\alpha^{S(t)=s}) - \ln\nu_\beta^{S(t)=s}\right) - \frac{\nu^s}{2}\left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right)\right]
\end{aligned}
$$

where $\Sigma_{W_d} = diag([{\beta_{d1}}^{-1}, {\beta_{d2}}^{-1}, ..., {\beta_{dM}}^{-1}])$, $\lambda_d = diag([\frac{a}{b_{d1}}, \frac{a}{b_{d2}}, ..., \frac{a}{b_{dM}}])$ and $\alpha_{0,new}^* = \sum_{S=1}^{ub}\alpha_{new}^* m_s$

Table B.2: Update Equations

Distribution	**Parameters**
$q(W_d^s \mid \hat{W}_d^s, \Sigma_{\hat{W}_d^s})$	$\Sigma_{\hat{W}_d{}^s} = [\lambda_d + \delta^s \sum_{t=1}^N \left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right) \alpha_{new}^{S(t)=s} y(t-1)y(t-1)^T]^{-1}$,
	$(\hat{W}_d{}^s)^T = \Sigma_{\hat{W}_d{}^s}[\sum_{t=1}^N \alpha_{new}^{S(t)=s} \left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right) \delta^s y_d(t) y(t-1)]$
$q(\beta_{dm} \mid a, b_{dm})$	$a = a^* + \frac{ub}{2}, b_{dm} = b^* + \frac{1}{2}\sum_{s=1}^{ub}[\hat{W}_{dm}^s{}^2 + \Sigma_{\hat{W}_{dm}^s}]$
$q(\alpha^s \mid \alpha_{new}^* m_s)$	$\alpha_{new}^* m_s = \alpha^* m_s^* + \sum_{t=1}^N q(S(t)=s)$
$q(R^{S(t)=s} \mid \nu_\alpha^{S(t)=s}, \nu_\beta^{S(t)=s})$	$\nu_\alpha^{S(t)=s} = \left[1 + \alpha_{new}^{S(t)=s}\right] \frac{\nu^s}{2} + (\frac{D}{2} - 1)\alpha_{new}^{S(t)=s}$
,	$\nu_\beta^{S(t)=s} = \left[1 + \alpha_{new}^{S(t)=s}\right] \frac{\nu^s}{2} + \alpha_{new}^{S(t)=s} \frac{\delta^s}{2} tr[y(t)y(t)^T]$
	$-\sum_{d=1}^D \alpha_{new}^{S(t)=s} \delta^s y(t)_d \hat{W}_d^s x(t)$
	$+\sum_{d=1}^D \alpha_{new}^{S(t)=s} \frac{\delta^{S(t)=s}}{2} tr[(\Sigma_{\hat{W}_d^s} + (\hat{W}_d^s)^T \hat{W}_d^s) y(t-1)y(t-1)^T]$
$q(S(t)=s) = \alpha_{new}^{S(t)=s}/z$	Z is normalizing constant

where $\alpha_{new}^{S(t)=s} = e^P$, such that P=$\Psi(\alpha_{new}^* m_s) - \Psi(\alpha_{0,new}^*) - \frac{D}{2}\ln(2\pi) + \frac{D}{2}\ln\delta^s - \frac{\delta^s}{2} tr[y(t)y(t)^T]$
$+\sum_{d=1}^D \delta^s y(t)_d \hat{W}_d^s y(t-1) - \sum_{d=1}^D \frac{\delta^s}{2} tr[(\Sigma_{\hat{W}_d^s} + (\hat{W}_d^s)^T \hat{W}_d^s)(y(t-1)y(t-1)^T)]$

$Z = \sum_{s=1}^{ub} \alpha_{new}^{S(t)=s}$, $\quad \Sigma_{W_d} = diag\left(\left[{\beta_{d1}}^{-1}, {\beta_{d2}}^{-1}, ..., {\beta_{dM}}^{-1}\right]\right), \lambda_d = diag\left(\left[\frac{a}{b_{d1}}, \frac{a}{b_{d2}}, ..., \frac{a}{b_{dM}}\right]\right)$

$$\delta^s = \frac{D * \sum_{t=1}^N \alpha_{new}^{S(t)=s}}{2} \left(\sum_{t=1}^N \left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right) \frac{\alpha_{new}^{S(t)=s}}{2} tr(y(t)y(t)^T - \sum_{t=1}^N \left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right) \alpha_{new}^{S(t)=s} \times \right.$$
$$\sum_{d=1}^D y_d(t) \hat{W}_d^s y(t-1) + \sum_{t=1}^N \left(\frac{\nu_\alpha^{S(t)=s}}{\nu_\beta^{S(t)=s}}\right) \alpha_{new}^{S(t)=s} \times$$
$$\left. \sum_{d=1}^D \frac{1}{2} tr\left[(\Sigma_{\hat{W}_d^s} + (\hat{W}_d^s)^T \hat{W}_d^s)(y(t-1)y(t-1)^T)\right] \right)^{-1}$$