Diagnostic Monitoring of High-dimensional Networked Systems (with Applications in Manufacturing and Healthcare System)

Yan Jin

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

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Shuai Huang, Chair

Wanpracha Art Chaovalitwongse

Shan Liu

Xiao-Hua (Andrew) Zhou

Program Authorized to Offer Degree:
Industrial & System Engineering

University of Washington

Abstract

Diagnostic Monitoring of
High-dimensional Networked Systems
(with Applications in Manufacturing and Healthcare System)

Yan Jin

Chair of the Supervisory Committee:
Assistant Professor Shuai Huang
Industrial & Systems Engineering

Rapid advances in sensor and information technology have resulted in both spatially and temporally data-rich environment, which creates a intensive need for us to develop novel statistical methods and computationally efficient algorithm to extract intelligent knowledge and informative patterns from these complicated system. For example, smart manufacturing, or it is also called advanced manufacturing technology take the benefit of information, technology and human intelligence to bring about a rapid revolution in the development and application of manufacturing. Another example is that, in health care field, such as Alzheimer's disease, the researchers have acquired a large number of biomarkers from various modalities including genotyping, neuroimaging and clinical assessment. These changes and development always produces a complicated high dimensional networked systems.

However, the statistical and computational challenges for addressing these complicated systems lay in their complex structures, such as high dimensionality, hierarchy, multi modality, heterogeneity and data uncertainty. On the other hand, a bunch of recent development in statistic, optimization and machine learning, such as graphical model, dimension reduction and feature screening technology provide more insights and angles to address the problems coming from these complex high dimensional networked system.

I depict the development of novel statistical model and computationally efficient algorithm to analysis the high dimensional networked system, and show how these proposed models are applied to real world applications, such as manufacturing system and Alzheimer's Disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I wishes to express sincere appreciation to my advisor, Professor, Shuai Huang, for his generous support, endless patience and encouragement, and insightful guidance throughout my phd study. I could not have imaged having a better advisor and mentor for my Ph.D. study.

I would like also to thank my committee members, Prof. W. Art Chaovalitwongse, Prof. Shan Liu, and Prof. Xiao-Hua (Andrew) Zhou. Prof. Art provided me precious comments to help me understand the knowledge of neuroimaging when I was working in IBIC. Prof. Liu gave me many valuable questions about diagnositc monitoring to help me think deeply and differently. My gratitude also goes to Prof. Zhou, who provided data, important comments, and suggestions for the multimodal biomarkers and research on Alzherimer's disease when I was working with his group. Without the help of them, this thesis is impossible.

Thank my collaborators, Special thanks to Dr. Yi Su from WUSTL and Dr. Guan Wang from NextEV company, it is my great preseasure to work with them and have their advises and suggestions.

I would like to thank all other people who helped me at UW.

Last but not least, I would like to thank my wife Chuchu Xue, daughter Elsa Jin, and parents Baozhong Jin and Shuiping Zhou, for their support and understanding throughout my life.

Thanks for all your encourgement!

DEDICATION

to my family

Chapter 1

INTRODUCTION

1.1 Motivation

Rapid advances in sensor and information technology have resulted in both spatially and temporally data-rich environment, which creates a intensive need for us to develop novel statistical methods and computationally efficient algorithm to extract intelligent knowledge and informative patterns from these complicated system. For example, smart manufacturing, or it is also called advanced manufacturing technology take the benefit of information, technology and human intelligence to bring about a rapid revolution in the development and application of manufacturing. Another example is that, in health care field, such as Alzheimer's disease, the researchers have acquired a large number of biomarkers from various modalities including genotyping, neuroimaging and clinical assessment. These changes and development always produces a complicated high dimensional networked systems.

To model the networked multivariate system, such as the sensor network in manufacturing system and brain connectivity in healthcare system, the graphical model like Bayesian network and Markov random field is one common and beneficial approach. After representing in graphical model, it is convenient to analyze its properties based on their statistical distributions and graph theory.

On the other hand, there are a bunch of issues emerging to tackle with in particular domains after representation of graphical model. For example, in fault detection and diagnosis problem, since the fault of predecessors could propagate to their successors, it is of interest to investigate the propagation mechanism. And it is necessary to develop more efficient algorithm to solve the graphical model problem theoretically, because most real world application requires real time results.

1.2 Research objectives

The objectives of my research would include:

- Develop novel statistical models for fault detection and fault diagnosis in large scale complex problem, including high-dimensional data, hierarchically-structured data, multi-modality data and uncertainty data, by drawing on recent theoretical developments in statistics, machine learning and quality control.

- Develop computationally efficient safe screening techniques to dramatically speed up the diagnostic monitoring procedure in large scale systems.

- Apply the developed novel statistical models for knowledge discovery and decision making from real-world dataset, including manufacturing and healthcare system.

1.3 Organizations

This thesis is presented in a multiple manuscript format. Each of the chapters, 2 to 5, is written as in individual research paper, with references listed in the end of the thesis. And chapter 6 would discuss the ongoing work and future work. The relationships among these chapters are depicted in Figure 1.1.

Chapter 2 presents a novel high-dimensional statistical model, called LASSO-BN, which is useful to do the fault detection and diagnosis simultaneously in large scale networked system that could be represented as BN. We did theoretical analysis to reveal the difference between the proposed LASSO-BN method with the state-of-art, VS-MSPC. And it has been evaluated by extensive numerical studies with benchmark methods.

Chapter 3 develops a computationally efficient algorithm to solve the sensor fault diagnosis problem, which is when hundreds or thousands of sensors are monitoring a high-dimensional process that generate abnormality signals that are actually interconnected, how could we perform fault diagnosis that is statistically accurate and computationally feasible? With applications on simulated and real-world processes, the proposed method is thoroughly

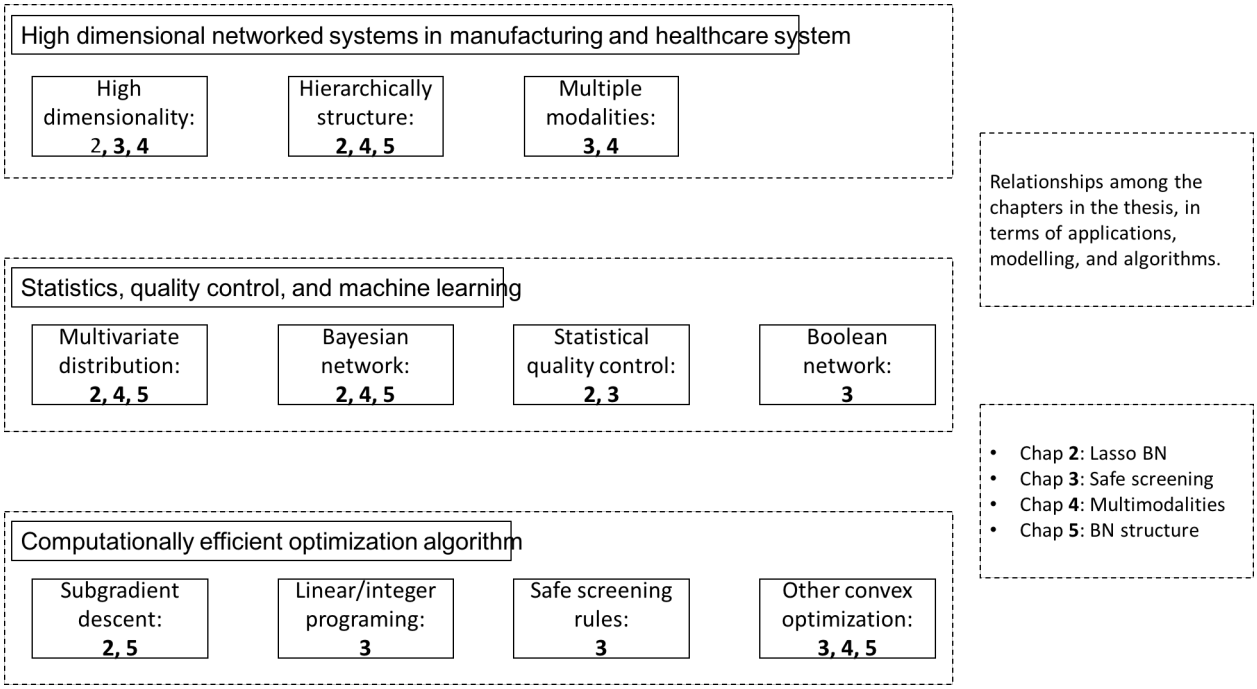


Figure 1.1: Relationships among the chapters in this thesis

evaluated which shows promising performance in terms of effectiveness, accuracy, and efficiency.

Chapter 4 takes a systematical perspective (mixed type Bayesian network) to study patterns of disease progression. We take into consideration of multimodal biomarkers such as APOE genotypes, SNP variants, demographics, FDG-PET, amyloid PET, MRI, and neuropsychological assessment. We conduct this study using ADNI baseline dataset, and find that the learned BN model provides findings that are in consistent with the AD literature.

Chapter 5 proposes a first of its kind method that can systematically elicit expert knowledge, optimally matched to observational data that has been collected, to identify the influential relationships between variables. Such a Bayesian learning framework will further lead to a systematical optimization formulation to automate the expert elicitation process. We conduct extensive numerical experiments on simulated data and real-world data to demonstrate the utility of our method and show its superior performance than baseline approaches.

Chapter 2

DIAGNOSTIC MONITORING OF MULTIVARIATE PROCESS VIA A LASSO-BN FORMULATION

Quality control of multivariate processes has been extensively studied in the past decades, however, fundamental challenges still remain due to the complexity of these multivariate processes and the decision-making challenges that require not only sensitive fault detection but also identification of the truly out-of-control variables. In many existing developments for multivariate process monitoring, the fault detection and the diagnosis are usually considered as two separate tasks. Recent developments have revealed insights that selective monitoring of the potentially out-of-control variables, identified by a variable selection procedure that is combined with the process monitoring method, could lead to promising performances of process monitoring. Following this line, we further propose the theme of diagnostic monitoring, that takes one step further than the selective monitoring idea and directs the monitoring effort on the potentially out-of-control variables, while the identification of the truly out-of-control variables can be achieved by integrating the process monitoring formulation with process cascade knowledge represented by a Bayesian Network. Computationally efficient algorithms are also developed for solving the optimization formulation with interesting connection to the LASSO problem being identified. Both theoretical analysis and extensive experiments on simulated dataset and real-world applications are conducted that show the superior performance of the proposed method.

2.1 Introduction

Recent rapid advances in sensor and information technologies have provided unprecedented opportunities for monitoring manufacturing systems that have many important process vari-

ables. Statistical monitoring of these multivariate processes has shown to be challenging, due to the complexity of these multivariate systems and the high anticipation of the decision-making capabilities. It is anticipated that the process monitoring methods can not only provide timely fault detection, but also identify the truly out-of-control variables that are responsible for the fault signal. In many existing research developments for multivariate process monitoring, the fault detection and the truly out-of-control diagnosis are usually considered as two separate tasks. For instance, the Hotellings chart, the Multivariate EWMA and Multivariate CUSUM charts were developed for fault detection but not much for diagnosis. Since these methods only provide overall evaluation of the process condition, ad-hoc diagnosis methods have been developed to identify the truly out-of-control variables that result in the out-of-control signals. Examples of these diagnosis methods include [36] that used Principal Component Analysis (PCA), the regression-based methods proposed in [62, 63, 64], and the MYT-decomposition method proposed in [110, 111].

Considering the fault detection and diagnosis as two separate tasks may not be the optimal solution for multivariate process monitoring. As pointed out in [160], it is very rare to see all the process variables experience shifts simultaneously in many real-world applications. As a result, by monitoring all the process variables to detect process changes could be very challenging, since the process change signal due to a small number of variables is rather weak when many in-control variables are included and treated equally. Therefore, to simultaneously conduct diagnosis and fault detection, some recent works such as [160, 171, 18], have proposed interesting frameworks that integrate variable selection and process monitoring simultaneously. For example, the basic idea of the VS-MSPC method proposed in [160] is, rather than monitoring all the variables, it is better to identify the potentially out-of-control variables by forward variable selection methods and then monitor these selected variables via a Shewhart-type chart. Similarly, [171] proposed to use the EWMA statistic to accumulate recent observations and then use the Least Absolute Shrinkage and Selection Operator (LASSO) for variable selection. Numerical studies in these recent developments have demonstrated that this simultaneously monitoring and diagnosis method performs well

in a number of applications.

The objective of our research is to extend [160, 171, 18]s developments to the multivariate processes where the cascade relationships between the process variables can be characterized via Bayesian Networks (BN). It has been found that the cascade relationships between the process variables hold great value for accurate determination of the truly out-of-control variables. Although the existing diagnostic methods such as the MTY approach can be used for identifying the variables that significantly contribute to an out-of-control signal, these identified variables may not truly be the out-of-control variables. This is because that the changes on the truly out-of-control variables will propagate to their downstream variables, misleading the diagnosis methods to include both the truly out-of-control variables and their downstream variables as identified out-of-control variables. For example, Figure 2.1 shows a hot forming process where the cascade relationships between the five process variables (X_1, blank holding force; X_2, temperature; X_3, tension in workpiece; X_4, material flow stress; X_5, final dimension of workpiece) can be represented as a Bayesian Network (BN). If is out-of-control, its effect will propagate to and further impact , resulting in out-of-control signals on all these variables. Thus, without incorporating the cascade relationships between process variables into the diagnostic procedure, it is very difficult to separate the truly out-of-control variables with their downstream variables that also exhibit out-of-control signals. Both our numerical experiments and theoretical analysis in later sections of this paper also confirmed that the methods proposed in [160, 171] could not identify the truly out-of-control variables due to the foregoing reasons.

Thus, in this paper, we propose a method called LASSO-BN for statistical monitoring of the multivariate processes that can be represented as BNs. We show that the LASSO-BN method can be formulated as a constrained likelihood estimation problem, which is later shown to be computationally equivalent to the LASSO formulation. We further provide theoretical comparison of the LASSO-BN with existing methods such as variable selection based multivariate statistical process control (MSPC) methods developed in [160, 171], which reveals theoretical reason why the LASSO-BN could outperform them.

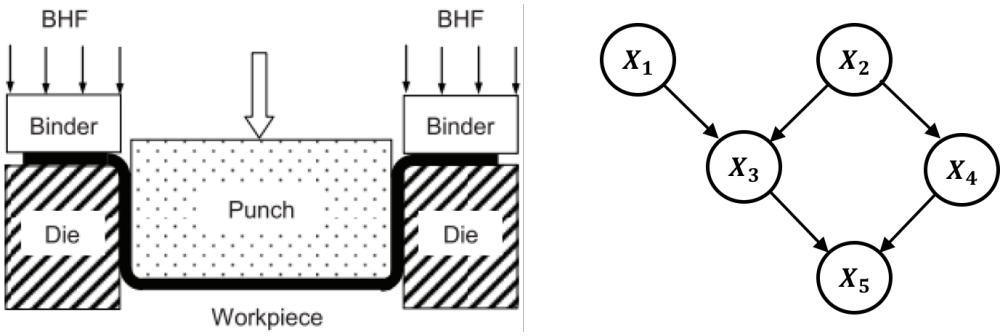


Figure 2.1: 2-D illustration of the hot forming process, and the corresponding BN structure

While we limit our current scope on the BNs, it is worthy of mentioning that the BN model is very capable and flexible that can model a wide range of problems. Comparing with other BN-based diagnosis methods such as [94, 165], LASSO-BN method integrates monitoring and diagnosis into a systematic formulation, and inherits the nice variable selection properties of sparse learning. Also, LASSO-BN is able to utilize the parameters of the BN for statistical inference, while in the other methods such as the causation-based T^2 method, only the qualitative cascade information was used. The article is structured in the following manner. Section 2.2 briefly reviews the basic concepts of Bayesian network, the existing multivariate monitoring methods, the generalized likelihood ratio test (GLRT), and the variable selection based MSPC methods such as the VS-MSPC that builds on GLRT. We then develop our LASSO-BN approach in Section 2.3, and investigate its application on both simulated dataset in Section 2.4 and a real-world dataset such as the Tennessee Eastman Chemical process in Section 2.5. A theoretical study is conducted in Section 2.6 to further reveal why the proposed method is superior to existing methods. Conclusions are drawn in Section 2.7.

2.2 Review of the BN, the GLRT control chart, and the VS-MSPC

In this paper, we concern the multivariate processes that can be characterized as a multivariate Gaussian distribution. Denote the p process variables as $\boldsymbol{X} = \{X_1, \ldots, X_p\}$, and $P(X_1, \ldots, X_p) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. With-

out loss of generality, we assume that $\boldsymbol{\mu} = \mathbf{0}$ when the process is in-control. In this section, we will briefly review the BN method that is useful for modeling the Gaussian multivariate process, and the GLRT control chart that is useful for monitoring the process.

2.2.1 The Gaussian BN

The BN has been found to be an effective tool for modeling complex manufacturing systems by a number of researchers, e.g., [94, 99, 100]. As modeling the joint distribution $P(X_1, \ldots, X_p)$ is the first step to establish the baseline of the process to be monitored, in what follows, we will show how the BN can be helpful to specify $P(X_1, \ldots, X_p)$. In particular, BN provides an intuitive yet powerful framework for encapsulating the complex relationships between the variables by the use of a directed acyclic graph (DAG) structure. The DAG consists of directed arcs between the variables. If there is a directed arc from X_i to X_j, i.e., $X_i \rightarrow X_j$, then X_i is called a parent of X_j. Let pa_j denote the set of variables that are all parents of X_j. A variable is independent of its non-descendants given its parents. Thus, the joint distribution $P(X_1, \ldots, X_p)$ can actually be decomposed into a product of p conditional probability distributions:

$$P(X_1, \ldots, X_p) = \prod_{i=1}^{p} P(X_i|pa_i) \tag{2.1}$$

where the conditional probability distributions $\{P(X_i|pa_i, i = 1, 2, \ldots, p\}$ are also called parameters of the BN. In other words, this decomposition property of BN implies that, to learn the multivariate joint probability distribution, we only need to estimate p conditional probability distributions which are usually with smaller dimensionality, making it much easier and applicable to handle multivariate manufacturing systems in real-world applications.

For multivariate Gaussian distributions, the conditional distribution $P(X_i|pa_i)$ in Equation 2.1 can be equivalently written as a linear regression model between any variable with its parent variables, as

$$X_i = \sum_{j \in pa_i} \omega_{ij} X_j + \epsilon_i \tag{2.2}$$

where ω_{ij} is the regression coefficient, $\epsilon_i \sim N(b_i, \sigma_i^2)$ represents the local variation of variable X_i with mean level as b_i and variance as σ_i^2. Based on Equation 2.2, the conditional distribution $P(X_i|pa_i)$ an be written as:

$$P(X_i|pa_i) = N(\sum_{j \in pa_i} \omega_{ij} X_j + b_i, \sigma_i^2) \tag{2.3}$$

For independent nodes whose parent variable set is null, i.e., $pa_i = null$, the density function is a univariate Gaussian distribution, $N(b_i, \sigma_i^2)$. Thus, based on Equation 2.1, the logarithm of the joint distribution is the logarithm of the product of these conditional distributions over all nodes in the DAG, leading to the following expression:

$$\ln P(X_1, \ldots, X_p) = \sum_{i=1}^{p} \ln P(X_i|pa_i) = -\sum_{i=1}^{p} \frac{(X_i - \sum_{j \in pa_i} \omega_{ij} X_j - b_i)^2}{2\sigma_i^2} + \text{const} \tag{2.4}$$

On the other hand, as we know that $P(X_1, \ldots, X_p) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it is of interest to convert the information encoded in Equation 2.4 to the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. This conversion is particularly useful for process monitoring purpose as many existing process control charts take $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as input information. A recursive procedure can be used. Based on the BN structure, it is always possible to rearrange the variables in a cascade order so earlier variables wont be children of later variables in this order. For instance, for the BN in Figure 2.1, the variables can be ordered as X_1, X_2, X_3, X_4, X_5 or X_1, X_2, X_4, X_3, X_5, while either order is a cascade order. Without loss of generality, assume that $X_1, \ldots, X_p$ is such a cascade order. With this order, the recursive procedure to derive the unconditional mean of each variable X_i is shown in below:

- Starts withX_1. From Equation 2.3, it is known that $\mu_1 = b_1$.

- Then, by following the order of the cascade, we can derive the mean of X_i as

$$\mu_i = \sum_{j \in pa_i} \omega_{ij}\mu_j + b_i \tag{2.5}$$

Similarly, we can obtain the covariance matrix for $P(X_1, \ldots, X_p)$ in the form of a recursion relation, i.e., for each $cov(X_i, X_j)$, we have

$$cov(X_i, X_j) = \sum_{k \in pa_j} \omega_{jk} cov(X_i, X_j) + \boldsymbol{I}_{ij}\sigma_j^2 \tag{2.6}$$

where $\boldsymbol{I}$ is the identity matrix.

2.2.2 The GLRT control chart

For monitoring multivariate Gaussian processes, the GLRT method has been a benchmark method. It encompasses a number of existing process monitoring methods including the Hotellings T^2 chart. Details of the GLRT control chart were provided in [80, 160]. Specifically, let $\boldsymbol{x}_t$ be the process measurement at time point t, i.e., $\boldsymbol{x}_t \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Process monitoring is to examine the following statistical hypothesis:

$$H_0 : \boldsymbol{\mu} \in \Omega_0 \text{ versus } H_1 : \boldsymbol{\mu} \in \Omega_1, \tag{2.7}$$

where $\Omega_0 = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{0}\}$ represents the parameter space when the process is in control; $\Omega_1 = \{\boldsymbol{\mu} : \boldsymbol{\mu} \neq \mathbf{0}\}$ is the parameter space when the process is out of control. To detect the potential mean shift, the GLRT statistics is defined as

$$\lambda(\boldsymbol{x}_t) = \frac{\max_{\boldsymbol{\mu} \in \Omega_0} L(\boldsymbol{x}_t, \boldsymbol{\mu})}{\max_{\boldsymbol{\mu} \in \Omega_1} L(\boldsymbol{x}_t, \boldsymbol{\mu})} \tag{2.8}$$

where $L(\boldsymbol{x}_t, \boldsymbol{\mu})$ is the likelihood of $\boldsymbol{x}_t$. The null hypothesis is rejected if $\lambda(\boldsymbol{x}_t) < c_1$, where $c_1 > 0$ is a constant that corresponds to a desired type-I error. With the details of the theoretical derivation omitted, it has been shown in [160] that the rejection region is equivalent to

$$\Lambda(\boldsymbol{x}_t) = \min_{\boldsymbol{\mu} \in \Omega_1} \{-\boldsymbol{x}_t^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t + (\boldsymbol{x}_t - \boldsymbol{\mu})\boldsymbol{\Sigma}^T (\boldsymbol{x}_t - \boldsymbol{\mu})\} \leq \log c_1 \tag{2.9}$$

We reject the null hypothesis if

$$\Lambda(\boldsymbol{x}_t) = \boldsymbol{x}_t^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t - \min_{\boldsymbol{\mu} \in \Omega_1} \{(\boldsymbol{x}_t - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu})\} \geq C \tag{2.10}$$

where $C = -\log c_1$. The value of C can be obtained by the simulation procedure proposed in [160]. As we have shown in previous section, with the knowledge of the BN, the joint distribution $P(X_1, \ldots, X_p) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be readily derived by the recursive procedure. The distributions parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be used as input of the GLRT method described here, for detecting possible process changes on the mean levels.

2.2.3 The control charts with variable selection capability

As mentioned before, it has been demonstrated that considering the fault detection and the diagnosis as two separate tasks may not be the optimal solution for multivariate process monitoring, particularly when there is only a small portion of the variables that are out-of-control. An intuitive explanation is that the majority of the in-control variables form a strong background noise that buries the out-of-control signal. Thus, both the VS-MSPC and LASSO-based Control Chart have been proposed in [160, 170] to simultaneously identify the out-of-control signal and perform diagnosis. Both methods follow a similar idea, which is to apply variable selection in conjunction with the GLRT statistics for process monitoring, i.e., VS-MSPC uses a stepwise variable selection method while the LASSO-based Control Chart uses LASSO. Specifically, the VS-MSPC employs the following formulation:

$$\Lambda(\boldsymbol{x}_t) = \boldsymbol{x}_t^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_t - S^2 \geq C \tag{2.11}$$

where $S^2 = \min_{\boldsymbol{\mu}} \{(\boldsymbol{x}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}) + \lambda M\}$, and $M = \sum_j I(|\mu_j| \neq 0)$, and λ is a parameter to control the penalty term. Both methods have shown superior performances on the process monitoring and diagnosis accuracy than the Hotellings T^2 chart. However, neither method is developed and examined in applications where the multivariate processes can be represented as BNs. Apparently, neither method can incorporate the cascade relationships between the variables encoded in a BN, which are critical information for understanding

and predicting how process changes propagate and accumulate from upstream variables to downstream variables.

2.3 LASSO-BN Formulation for the Diagnostic Monitoring Control Chart

2.3.1 The diagnostic monitoring concept

It is important to point out that the information encoded in $\{b_i = 0, i = 1, 2, \ldots, p\}$ is essential for determining the truly out-of-control variables since only the variables that have $b_i \neq 0$ are the truly out-of-control variables. Figure 2.2 gives an illustration of this fault propagation in the hot forming process. While X_1 is out-of-control, the variables X_3 and X_5 will also show out-of-control signals. Without knowledge of the cascade between these variables, it is difficult to identify the truly out-of-control variable. On the other hand, the predictive relationships between the variables (i.e., as characterized by Equation 2.5) could be very valuable information for enhancing the diagnosis.

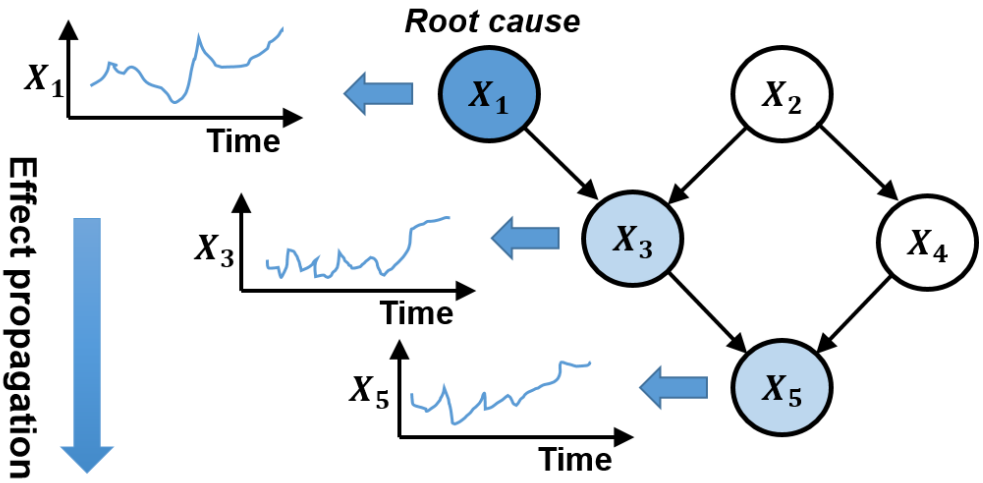


Figure 2.2: Illustration of the process fault propagation in the hot forming process: the fault in the variable X_1 will propagate to its descent variables, X_3 and X_5.

However, none of the existing methods (such as those developed in [160, 171]) can ef-

fectively incorporate the cascade relationship between the variables to recover the nonzero members in $b_i = 0, i = 1, 2, \ldots, p$. Instead, they aim to recover the nonzero members in $\boldsymbol{\mu}$, making them sub-optimal on the identification of the truly out-of-control variables. Although they can identify out-of-control variables out of a large number of variables, it is reasonable to suspect that both methods may not be able to identify the truly out-of-control variables. In other words, for the identified variables that exhibit out-of-control mean shifts, both methods cannot further identify which variables truly have mean shifts, and which variables dont have mean shifts but are just influenced by upstream variables with mean shifts. Thus, one major advantage of our proposed LASSO-BN chart is its capability in searching for the truly out-of-control variables by incorporating the cascade relationships between the variables as shown in Equation 2.2. Note that, a key feature of the proposed diagnostic monitoring concept is that we integrate diagnosis and monitoring. We are motivated by the consideration that if we allocate the monitoring resource on selected variables that are highly possible to be out-of-control, we could filter out many in-control variables and increase the statistical power for change detection. Indeed, later our numerical studies will show that this integrated framework will lead to better monitoring and diagnosis, and the advantage is particularly significant for high-dimensional applications.

2.3.2 The formulation of the LASSO-BN method

The BN model provides important knowledge for linking the mean shifts of the truly out-of-control variables with the mean shifts of the variables that exhibit out-of-control mean shift signals. Particularly, assume that the manufacturing system is undergoing a process shift that has a new mean level $\boldsymbol{\mu} \neq \mathbf{0}$. The LASSO-BN aims to recover the true mean shifts of the variables $\{b_i, i = 1, 2, \ldots, p\}$ that are underlying the exhibited mean shifts $\boldsymbol{\mu}$. According to the BN model, it is known from Equation 2.5 that $\mu_i = \sum_{j \in pa_i} \omega_{ij} \mu_i + b_i, i = 1, \ldots, p$. We propose the following constrained optimization formulation:

$$
\begin{aligned}
S^2 = & \min_{\boldsymbol{\mu} \in \Omega_1} \{(\boldsymbol{x}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu})\} \\
& \text{subject to } \mu_i = \sum_{j \in pa_i} \omega_{ij} \mu_i + b_i, i = 1, \ldots, p
\end{aligned}
\tag{2.12}
$$

In the same spirit of VS-MSPC and LASSO-based Control Chart, we impose the assumption that only a few variables are the truely out-of-control variables. Specifically, we impose the L1 norm penalty on $\{b_i, i = 1, 2, \ldots, p\}$ to identify the few nonzero elements. This leads to the following formulation:

$$
\begin{aligned}
S^2 = & \min_{\boldsymbol{\mu} \in \Omega_1} \{(\boldsymbol{x}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}) + \lambda \sum_{i=1}^{p} |b_i|\} \\
& \text{subject to } \mu_i = \sum_{j \in pa_i} \omega_{ij} \mu_i + b_i, i = 1, \ldots, p
\end{aligned}
\tag{2.13}
$$

This formulation can be further rewritten as an optimization problem of unknown variables $\{b_i, i = 1, 2, \ldots, p\}$ rather than $\{\mu_i, i = 1, 2, \ldots, p\}$. It is known that the relationship between $\boldsymbol{\mu}$ and $\boldsymbol{b}$ is linear, i.e.,

$$
\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} 1 & w_{21} & \ldots & w_{p1} \\ w_{12} & 1 & \ldots & w_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1p} & w_{2p} & \ldots & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} = \boldsymbol{W} \boldsymbol{b}
$$

The problem is how to identify $\boldsymbol{W}$ which is determined by the path coefficients $\{\omega_{ij}, i, j = 1, 2, \ldots, p\}$, which is solved by the following Lemma 1.

Lemma 1: The matrix $\boldsymbol{W}$ equals to the inverse of $\boldsymbol{I} - \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is the path coefficients matrix, i.e., $\omega_i j$ is the element in the row i and column j of $\boldsymbol{\Omega}$.

Proof: Since $\boldsymbol{\mu} = \boldsymbol{\Omega}\boldsymbol{\mu} + \boldsymbol{b}$, and $\boldsymbol{\mu} = \boldsymbol{W}\boldsymbol{b}$, we could derive that $\boldsymbol{W}\boldsymbol{b} = \boldsymbol{\Omega}\boldsymbol{\mu} + \boldsymbol{b}$. This leads to $(\boldsymbol{I} - \boldsymbol{\Omega})\boldsymbol{\mu} = \boldsymbol{b}$, and further, $\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{\Omega})^{-1}\boldsymbol{b} = \boldsymbol{W}\boldsymbol{b}$. Therefore, it is shown that $\boldsymbol{W} = (\boldsymbol{I} - \boldsymbol{\Omega})^{-1}$.

With $\boldsymbol{\mu} = \boldsymbol{W}\boldsymbol{b}$, the original constrained optimization problem can be rewritten as the following unconstrained problem:

$$\min_{\boldsymbol{b}\in\Omega_1}\{(\boldsymbol{x}_t - \boldsymbol{W}\boldsymbol{b})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_t - \boldsymbol{W}\boldsymbol{b}) + \lambda\sum_{i=1}^{p}|b_i|\} \tag{2.14}$$

While Equation 2.14 resembles the LASSO formulation, it actually has a simpler structure than a general LASSO problem and can be solved very efficiently. We can further reveal that this optimization problem can be simplified by using the following Lemma 2.

Lemma 2: $\boldsymbol{\Sigma}^{-1} = \boldsymbol{W}^{-T}\boldsymbol{D}^{-1}\boldsymbol{W}^{-1}$

Proof: The covariance matrix $\boldsymbol{\Sigma}$ is defined as $\boldsymbol{\Sigma} = cov(\boldsymbol{X})$. Specifically, $cov(X_i, X_j) = cov(\boldsymbol{W}_{i\cdot}\boldsymbol{b}, \boldsymbol{W}_{j\cdot}\boldsymbol{b}) = \sum_k w_{ik}w_{jk}\sigma_k^2$, where $\boldsymbol{W}_{i\cdot}$ denotes for the i^{th} row of matrix $\boldsymbol{W}$. Thus, $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{D}\boldsymbol{W}^T$ where $\boldsymbol{D}$ is a diagonal matrix with $d_{kk} = \frac{2}{k}$. Equivalently, this leads to that $\boldsymbol{\Sigma}^{-1} = \boldsymbol{W}^{-T}\boldsymbol{D}^{-1}\boldsymbol{W}^{-1}$.

Lemma 2 actually implies that $\boldsymbol{\Sigma}^{-1}$ can be rewritten as the product of two matrices, i.e., $\boldsymbol{\Sigma}^{-1} = \boldsymbol{R}^T\boldsymbol{R}$ where $\boldsymbol{R} = \boldsymbol{W}^{-1}\sqrt{\boldsymbol{D}}^{-1}$. Plug in this expression into Equation 2.14 we will have

$$\min_{\boldsymbol{b}\in\Omega_1}\{(\boldsymbol{R}\boldsymbol{x}_t - \sqrt{\boldsymbol{D}}^{-1}\boldsymbol{b})^T(\boldsymbol{R}\boldsymbol{x}_t - \sqrt{\boldsymbol{D}}^{-1}) + \lambda\sum_{i=1}^{p}|b_i|\} \tag{2.15}$$

Let $\boldsymbol{z}_t = \sqrt{\boldsymbol{D}}\boldsymbol{R}\boldsymbol{x}_t$ and $\lambda = tr(\boldsymbol{D})\cdot\lambda$, the above optimization problem can be written as:

$$S^2 = \min_{\boldsymbol{b}\in\Omega_1}((\boldsymbol{z}_t - \boldsymbol{b})^T(\boldsymbol{z}_t - \boldsymbol{b})) + \lambda\|\boldsymbol{b}\|_1 \tag{2.16}$$

The problem 2.16 is actually a saturated LASSO problem. The elements b_i can be solved independently and closed form solution can be derived as:

$$b_i = \begin{cases} z_{ti} - \lambda & \text{if} z_{ti} - \lambda > 0 \\ 0 & \text{if} |z_{ti}| < \lambda \\ z_{ti} + \lambda & \text{if} z_{ti} - \lambda < 0 \end{cases} \tag{2.17}$$

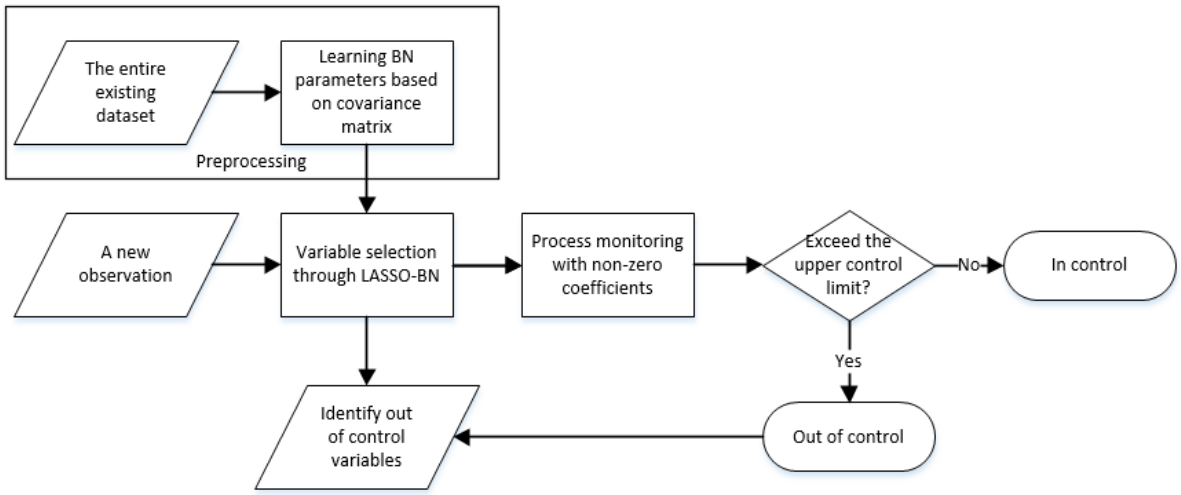


Figure 2.3: Flow chart of applying LASSO-BN to diagnostic monitoring

After we identify the potential out-of-control variables through the use of LASSO-BN, then monitor these variables by a multivariate control chart such as Hotellings T^2 chart.

2.3.3 Summary of the diagnostic monitoring procedure using LASSO-BN

As a summary, Figure 2.3 illustrates the procedure of implementing the LASSO-BN Chart for online SPC applications. It consists of the following major steps

Offline learning

The parameters of the LASSO-BN will be learned based on historical process data. Particularly, as analogical to the estimation of the parameters of the multivariate normal distribution for using Hotellings T^2 chart, the offline learning of LASSO-BN refers to the estimation of the parameters $\{\omega_{ij}, \sigma_i^2, i, j = 1, 2, \ldots, p\}$ to establish the BN. Learning of a BN generally has two steps: learning of the structure and learning of the parameters. Structure learning concerns how to recover the unknown BN structure when this structure is unknown.

Extensive research efforts have been developed for BN structure learning, which could be roughly divided into two categories, the constraint-based methods and the score-based methods. While structure learning has been investigated in the literature to a great extent, it is still a very difficult problem and is only needed when there is lack of knowledge of the cascade relationships between the variables. In this paper, we mainly concern the use of BN for diagnostic monitoring and limit our scope to the manufacturing applications where the cascade relationships between the variables could usually be obtained by the domain knowledge, such as the hot forming process shown in Figure 2.1 and the Tennessee Eastman Process that will be discussed in Section 2.5. Therefore, here, the main learning task of BN is to estimate the parameters $\{\omega_{ij}, \sigma_i^2, i, j = 1, 2, \ldots, p\}$. This estimation procedure consists of applying linear regression estimation p times, each time concerns one regression model as described in Equation 2.2. With the estimated parameters we could derive R and solve Equation 2.16 by using Equation 2.17.

Online monitoring

To use the LASSO-BN for online monitoring, for instance, consider the process monitoring with individual observation. For the new observation $\boldsymbol{x}_t$, we transform it to $\boldsymbol{z}_t = \boldsymbol{R}\boldsymbol{x}_t$, then plug it in Equation 2.16 and identify the potential M out-of-control variables. Here, following [160], M is a pre-specified parameter that reflects the prior belief of how many out-of-control variables could potentially emerge. Note that, in general formulation of LASSO problems, it has been difficult to directly derive the corresponding penalty parameter λ that can achieve the given number of selected variables. However, in our case, we have found a very interesting property of the formulation as we noted in Equation 2.16 and Equation 2.17, which admits simple closed-form solution. Based on this connection, deriving the penalty parameter λ is straightforward. A Hotellings T^2 chart is applied to monitor these identified M variables. Out-of-control alarm will be triggered if the T^2 statistics is out of the control limit and these M variables will be potential out-of-control variables responsible for the out-of-control alarm. Then actions should be taken to adjust the process, remove the faults in the truly

out-of-control variables, and bring the process back to normal state.

2.4 Simulation studies

In this section, we will conduct simulation studies to evaluate the effectiveness of the proposed diagnostic monitoring method. We consider BNs that are small, medium, large, i.e., corresponding to $p = 30, 50, 100$, respectively. With a given p, the BN structure and the associated parameters $\{\omega_{ij}, i, j = 1, 2, \ldots, p\}$ are randomly generated by using existing BN simulation tools such as R package pcalg. Without loss of generality, we always use $\sigma_i^2 = 1$ and $b_i = 0$ for $i = 1, 2, \ldots, p$. The generated BN establishes the baseline of the in-control process, and data generated from this BN are in-control process observations. To simulate out-of-control data, we randomly assign mean shifts to some variables by setting $b_i = \sigma$ for these variables. Different choices of $\sigma \in \{0.3, 0.5, 0.7, 1, 1.5\}$ are investigated to evaluate how sensitive the proposed method could be to different magnitudes of the mean shifts. We also investigate different number of out-of-control variables in our simulation. While more out-of-control variables present relatively easier problem for both monitoring and diagnosis, here we present our results when only 3 variables are truly out-of-control variables that have mean shifts. For each combination of p and σ, we simulate 1000 process observations as the training data to learn the BN parameters and another 1000 process observations as the out-of-control data to test the Average Run Length (ARL) of the control charts. The in-control ARL for all the charts used in the simulation is set to 200 and each ARL is obtained using at least 10,000 replications. Since our primary motivation is to enhance the diagnosibility, we also compare the proposed method with existing methods. Limited by the availability of the codes in some recent works and the difficult to reproduce the computational procedure and results, here, we focus our comparison with the VS-MSPC and Causation-based T^2 method on identifying the truly out-of-control variables in terms of sensitivity and specificity. Specifically, note that, in ground truth, each process variable is either in-control (class I) or out-of-control (class II), while the monitoring method could classify the process variables as either in-control or out-of-control. Then, if a truly in-control variable is classified as an

out-of-control variable, it is a false positive; if a truly out-of-control variable is classified as an in-control variable, then it is a false negative. Therefore, this resembles a binary classification problem and the ROC curve can be used to aggregate the diagnostic performance of a range of choices of M. The Area under the Curve (AUC) can be reported as an overview of the ROC curve: a more accurate diagnostic method, a larger AUC value is expected. The AUC values of the simulation results are shown in Table 2.1. Note that in Table 2.1, we further consider the sample size for the subgroup for online monitoring, e.g., $n_s = 2$ means that for each time point there are 2 observations that can be used for diagnosis. For the Causation-based T^2 method, $n_s = 10$.

It can be observed from Table 2.1 that the proposed LASSO-BN control chart is superior on identifying the truly out-of-control variables, particularly when δ is small or/and p is large. This is expected since the LASSO-BN control chart is capable of using the cascade information for enhancing the diagnosibility, while on the contrary, the VS-MSPC control chart can only identify the variables that exhibit mean shifts. The reason for the better performance of the proposed LASSO-BN method than the causation-based T^2 method is probably that the LASSO-BN method integrates monitoring and diagnosis into a systematic formulation, and inherits the nice variable selection properties of sparse learning. Also, the parameters of the BN are actually used for statistical inference in the LASSO-BN method, while in the causation-based T^2 method, only the qualitative cascade information was used. In addition, we found that with increasing magnitude of mean shift, i.e., larger δ, the diagnostic accuracy of all the three methods can be improved. Another general trend is that the larger the sample size of the subgroup of observations for online monitoring (i.e., n_s), the better diagnosis performance for all methods.

Besides the comparison of the diagnostic accuracy, the comparison of the methods on the ARL is also shown in Table 2.2. Because the exact number of truly out-of-control variables is usually unknown in practice, we present results for a reasonable range of M, i.e., $M \in \{2, 3, 4, 5\}$ for each of the simulated scenarios. Overall, it can be observed that the LASSO-BN method outperforms VS-MSPC and T^2 chart in all kinds of situations with

smaller ARL values when there is a process fault. Also, the performance of LASSO-BN seems to be robust on the mis-specification of the parameter M, i.e., recall that the true number of out-of-control variables is 3 in the simulation model. This robust property is also observed for VS-MSPC, which is consistent with the results observed in [160].

Table 2.1: Comparison of LASSO-BN with VS-MSPC on the diagnostic accuracy of out-of-control variables

		LASSO-BN				VS-MSPC				Causation
		n_s				n_s				Based
p	δ	1	2	5	10	1	2	5	10	T^2
30	0.3	0.55	0.70	0.74	0.76	0.49	0.55	0.58	0.58	0.48
	0.5	0.96	0.96	0.99	1.00	0.64	0.61	0.65	0.66	0.46
	0.7	0.98	0.99	1.00	1.00	0.64	0.68	0.70	0.72	0.51
	1	1.00	1.00	1.00	1.00	0.74	0.79	0.80	0.83	0.56
	1.5	1.00	1.00	1.00	1.00	0.78	0.80	0.83	0.84	0.58
50	0.3	0.68	0.63	0.66	0.72	0.43	0.52	0.57	0.60	0.45
	0.5	0.90	1.00	1.00	1.00	0.66	0.68	0.64	0.69	0.38
	0.7	0.90	0.99	1.00	1.00	0.69	0.70	0.67	0.75	0.48
	1	1.00	1.00	1.00	1.00	0.78	0.80	0.83	0.86	0.52
	1.5	1.00	1.00	1.00	1.00	0.74	0.77	0.87	0.89	0.55
100	0.3	0.62	0.69	0.79	0.82	0.54	0.63	0.70	0.72	0.38
	0.5	0.69	0.91	1.00	1.00	0.58	0.70	0.81	0.85	0.39
	0.7	0.97	0.99	1.00	1.00	0.56	0.74	0.83	0.88	0.46
	1	1.00	1.00	1.00	1.00	0.95	0.94	0.98	1.00	0.45
	1.5	1.00	1.00	1.00	1.00	0.93	0.98	0.99	1.00	0.51

Table 2.2: Comparison of LASSO-BN with VS-MSPC and T^2 on the ARL

p	δ	T^2	LASSO-BN				VS-MSPC			
			M				M			
			2	3	4	5	2	3	4	5
30	0.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0
	0.3	184.8	192.4	186.2	182.7	181.1	188.8	186.2	184.3	184.2
	0.5	145.3	155.2	143.2	142.8	142.9	152.2	153.2	148.9	148.2
	0.7	100.8	110.1	103.2	99.5	92.8	128.5	105.8	102.9	99.8
	1	86.8	90.9	94.3	85.5	71.4	118.7	99.0	92.4	90.5
	1.5	32.3	31.7	31.5	28.2	25.9	47.7	43.4	40.9	39.9
50	0.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0
	0.3	188.2	193.2	190.8	185.5	184.6	194.2	192.8	189.9	190.9
	0.5	172.8	180.0	175.4	172.2	175.2	178.8	179.2	171.2	172.3
	0.7	123.4	123.1	121.5	118.0	116.4	125.5	122.8	119.9	119.0
	1	101.9	117.6	115.1	92.0	84.9	120.5	119.0	103.1	90.1
	1.5	45.2	43.9	40.8	39.1	38.2	74.1	64.5	57.5	43.1
100	0.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0	200.0
	0.3	190.2	191.9	188.9	187.0	186.5	192.0	190.1	189.4	188.0
	0.5	180.2	182.5	175.6	176.2	173.4	181.3	180.0	178.8	178.2
	0.7	145.3	150.1	148.2	140.8	139.9	161.1	168.5	155.2	150.3
	1	128.2	133.3	144.9	114.9	126.6	153.8	156.1	145.2	142.3
	1.5	59.5	62.1	65.8	55.2	55.2	79.9	75.0	76.6	64.9

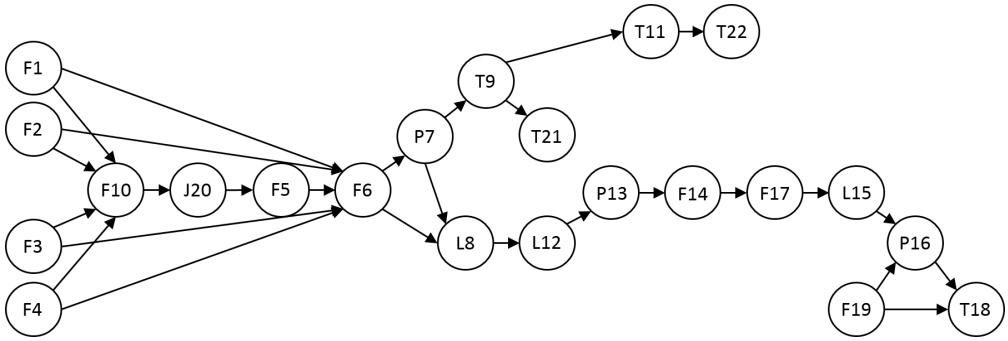


Figure 2.4: The BN of the TEP constructed by engineering knowledge of the process; the name for each node represents a specific process variable defined in the original TEP problem and can be found in [165]

2.5 Real-world example: the Tennessee Eastman Process (TEP)

The Tennessee Eastman Process (TEP) has been a benchmark process for evaluating process monitoring and fault diagnosis methods ever since the Eastman Chemical Company created this process simulator. The TEP is a chemical process that is composed of 12 input variables (manipulated variables) and 41 output variables (measurement variables). A BN of the TEP has been built in [165] that focused on 22 selected variables among the 41 measurement variables. The BN structure was identified by prior process knowledge and process flow sheet. As described in [165], the cascade of the 22 variables was known, so these 22 variables were sorted in terms of process flow order from upstream to downstream units and then placed into network hierarchy as nodes without any arcs. Then, the interactions among the variables are analyzed based on the prior process knowledge and used to determine where to place the arcs, leading to the completion of the BN structure as shown in Figure 2.4.

With knowledge of BN structure, in-control process data from TEP archive at University of Washington `http://depts.washington.edu/control/LARRY/TE/download.html` can be used to estimate the parameters of the BN. The same procedure used in the simulation studies can be applied here to generate out-of-control data, and then, compare the performance of

LASSO-BN with VS-MSPC and the causation-based T^2 method using this data. Specifically, we simulate mean shift in randomly chosen process variables, e.g., the number of out-of-control variables are set as $p_{oc} \in \{1, 3, 5\}$, and investigate a range of magnitudes of the mean shifts, e.g., $\delta \in \{0.3, 0.5, 0.7, 1, 1.5\}$.

The performances of the LASSO-BN, the VS-MSPC, and the causation-based T^2 method are shown in the Table 2.3. Overall the conclusion is consistent with the results in the simulated dataset reported in Table 2.1, showing superior performance of LASSO-BN on all cases considered.

2.6 Theoretical analysis of LASSO-BN and variable selection based control charts

It has been demonstrated in Section 2.4 and 2.5 that the proposed LASSO-BN method outperforms VS-MSPC and causation-based T^2 method across different situations. In this section we aim to identify some theoretical insights of why the LASSO-BN method could lead to better performance. Our result is partially intuitive, also surprising to a certain degree, since it reveals that, as long as the underlying system can be represented as a BN, the formulation of VS-MSPC is actually biased.

Specifically, our theoretical study uses a simple BN that has only two variables, as shown in Figure 2.5. The path coefficient $\omega_{21} = a$ and the inherent variances of the two variables are 1, e.g., $\sigma_1^2 = 1$ and $\sigma_2^2 = 1$. Then, given a mean shift vector $\boldsymbol{b} = (b_1, b_2)^T$, the mean vector $\boldsymbol{\mu}$ is $\boldsymbol{\mu} = (\mu_1, \mu_2)^T = \boldsymbol{W}\boldsymbol{b} = \begin{bmatrix} 1 & 0 \\ a & 1 \end{bmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = (b_1, ab_1 + b_2)^T$.

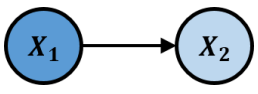


Figure 2.5: A BN with only two variables

From Equation 2.17 we have known the closed form solutions for estimating b_i. In what follows we show how to theoretically evaluate the probability of proposed method to correctly

Table 2.3: Comparison of LASSO-BN with VS-MSPC on the diagnostic accuracy of out-of-control variables for the TEP case

		LASSO-BN				VS-MSPC				Causation
		n_s				n_s				Based
p	δ	1	2	5	10	1	2	5	10	T^2
1	0.3	0.64	0.86	1.00	1.00	0.55	0.50	0.71	0.73	0.55
	0.5	0.93	1.00	1.00	1.00	0.74	0.76	0.81	0.76	0.56
	0.7	1.00	1.00	1.00	1.00	0.81	0.76	0.86	0.86	0.55
	1	1.00	1.00	1.00	1.00	0.81	0.81	0.86	0.91	0.57
	1.5	1.00	1.00	1.00	1.00	0.83	0.81	0.86	0.91	0.61
3	0.3	0.73	0.80	0.82	1.00	0.74	0.71	0.77	0.71	0.53
	0.5	0.98	0.95	1.00	1.00	0.74	0.67	0.76	0.74	0.58
	0.7	1.00	1.00	1.00	1.00	0.76	0.76	0.76	0.76	0.54
	1	1.00	1.00	1.00	1.00	0.76	0.81	0.81	0.81	0.62
	1.5	1.00	1.00	1.00	1.00	0.79	0.81	0.86	0.86	0.63
5	0.3	0.67	0.95	1.00	1.00	0.62	0.64	0.60	0.61	0.58
	0.5	0.95	1.00	1.00	1.00	0.71	0.74	0.71	0.74	0.59
	0.7	0.98	1.00	1.00	1.00	0.71	0.71	0.74	0.72	0.64
	1	1.00	1.00	1.00	1.00	0.76	0.76	0.76	0.76	0.65
	1.5	1.00	1.00	1.00	1.00	0.81	0.81	0.86	0.88	0.68

identify the truly out-of-control variables. Specifically, recall that $\boldsymbol{z}_t = \boldsymbol{R}\boldsymbol{x}_t$. Since $\boldsymbol{x}_t$ is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{z}_t$ follows a multivariate normal distribution with mean $\boldsymbol{R\mu} = \boldsymbol{W}^{-1}\boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{b} = \boldsymbol{b}$ and variance $\boldsymbol{R\Sigma R}^{-T} = \boldsymbol{I}$. In other words, $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{I})$. Examples of the distribution of $\boldsymbol{z}_t$ under different situations of $\boldsymbol{b}$ are shown in Figure 2.6.

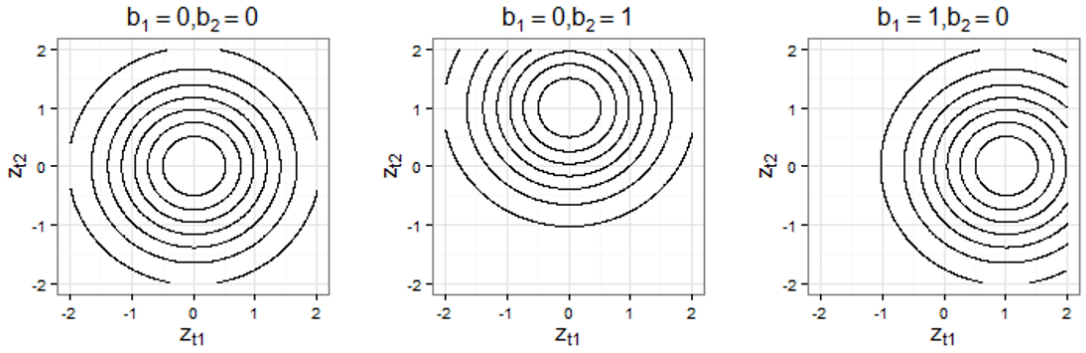


Figure 2.6: The distribution of $\boldsymbol{z}_t$ under different situations of $\boldsymbol{b}$

From the Figure 2.6 we could observe that the distribution of $\boldsymbol{z}_t$ faithfully captures the underlying ground truth of $\boldsymbol{b}$. E.g., consider the situation that the process is in-control, $b_1 = b_2 = 0$. Recall that from Equation 2.17 we know that $b_i = 0$ as long as $|z_{ti}| < \lambda$, then, Figure 2.6 (a) implies that we could have a limited chance of making false alarms since the rectangle region (corresponds to $\lambda = 2$) captures the majority of the probability mass, leading to the solution $b_1 = b_2 = 0$ which is correct. Similarly, consider the situation that the process is out-of-control and $b_1 = 0, b_2 = 1$ (Figure 2.6 (b)), then it is clear that the rectangle region no longer captures the majority of the probability mass (particularly, for z_{t2}) and tends to give a solution that b_2 is nonzero since the value of z_{t2} is very likely to be outside of the rectangle region. Similar observation can be obtained by examining Figure 2.6 (c).

While Figure 2.6 illustrates that the proposed LASSO-BN method has the capability of

identifying the truly out-of-control variable, similar figures can be generated to show why the other methods such as VS-MSPC will not be as efficient as the proposed method to capture the underlying mean shifted variables. For the sake of space limit, here, instead of presenting the figures, we could further quantify the probability of detecting the correct mean shifted variables for different situations for both methods. For instance, consider a case that $\lambda = 1$ and the path coefficient $a = 0.5$. Then the probabilities of identifying the correct mean shifted variables for different situations by the proposed method and the VS-MSPC are shown in Table 2.4.

Table 2.4: Probability of identifying truly out-of-control variables for the BN in Figure 2.5

		LASSO-BN	**VS-MSPC**
$b_1 = 1$	$b_2 = 0$	0.3413	0.0747
$b_1 = -1$	$b_2 = 0$	0.3413	0.0747
$b_1 = 0$	$b_2 = 1$	0.3413	0.3123
$b_1 = 0$	$b_2 = -1$	0.3413	0.3123

Apparently, it shows that the proposed method has a significant gain in the diagnosis performance than the VS-MSPC. Although our analysis is done on a simple BN with only two nodes, the general insight can be applied to more complex BN structures, evidenced by the numeric results reported in Sections 2.4 and 2.5, which actually imply that the utilization of the cascade structure of the process variables plays an essential role for diagnosis. The reason for the disadvantage of VS-MSPC and other variable selection based MSPC method could be that they have biased design for the penalty function. For example, consider that the LASSO-based control chart that has the following optimization formulation:

$$S^2 = \min_{\boldsymbol{\mu}} \left\{ (\boldsymbol{x}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}) + \lambda \sum_{i=1}^{p} |\mu_i| \right\}$$

If the underlying process is represented as a BN structure, then, following the same line

as in Section 2.3, it can be rewritten as

$$
\begin{aligned}
S^2 &= \min_{\boldsymbol{b}\in\Omega_1} \left\{ \frac{1}{2}(\boldsymbol{z}_t - \boldsymbol{b})^T(\boldsymbol{z}_t - \boldsymbol{b}) + \lambda\|\boldsymbol{W}\boldsymbol{b}\|_1 \right\} \\
&= \min_{\boldsymbol{b}\in\Omega_1} \left\{ \frac{1}{2}(z_{t1} - b_1)^2 + \frac{1}{2}(z_{t2} - b_2)^2 + \lambda|b_1| + \lambda|ab_1 + b_2| \right\}
\end{aligned}
$$

Apparently, $|ab_1 + b_2|$ is a biased penalty term which is not as straightforward as $|b_2|$ that is used in LASSO-BN. This will lead to a biased solution region as well, i.e., the closed-form solution is shown in below:

$$
\begin{cases} b_1 = 0 \\ b_2 = z_{t2} - \lambda \end{cases} \qquad z_{t1} \in (\lambda - a\lambda, \lambda + a\lambda); z_{t2} > \lambda
$$

$$
\begin{cases} b_1 = 0 \\ b_2 = z_{t2} + \lambda \end{cases} \qquad z_{t1} \in (-\lambda - a\lambda, -\lambda + a\lambda); z_{t2} < -\lambda
$$

$$
\begin{cases} b_1 = z_{t1} - \lambda - a\lambda \\ b_2 = 0 \end{cases} \qquad z_{t1} \in (\lambda + a\lambda, +\infty); z_{t2} = \lambda
$$

$$
\begin{cases} b_1 = z_{t1} - \lambda - a\lambda \\ b_2 = 0 \end{cases} \qquad z_{t1} \in (-\infty, -\lambda - a\lambda); z_{t2} = -\lambda
$$

Since $\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{b}, \boldsymbol{I})$, from the solutions shown above, it can be derived what is the probability of identifying the truly out-of-control variables for different situations. E.g., suppose that $b_1 \neq 0$ and $b_2 = 0$, then, to recover this true solution, it is desired that $z_{t1} \in (\lambda + a\lambda, +\infty); z_{t2} = \lambda$ or $z_{t1} \in (-\infty, -\lambda - a\lambda); z_{t2} = -\lambda$. However, it seems that the probability of $z_{t1} \in (\lambda + a\lambda, +\infty)$ or $z_{t1} \in (-\infty, -\lambda - a\lambda)$ depends on the values of a, and the probability could become even smaller when a is larger. Not like this, the LASSO-BN can achieve universal good performance on the fault diagnosis as shown in Figure 2.6, due to its incorporation of the cascade relationships between the variables that can be translated into

unbiased penalty term in its optimization formulation as shown in 2.14 and 2.16. While the theoretical analysis is conducted on a small BN, it reveals some essential differences between the proposed LASSO-BN method with the existing methods such as VS-MSPC.

2.7 Conclusion

The contributions of this paper can be summarized as follows:

- To the best of our knowledge, we made the first effort to develop a diagnostic monitoring method that conducts fault detection and diagnosis simultaneously for complex multivariate processes represented as BNs. Computationally efficient algorithm is developed, with interesting connection with the LASSO formulation also being identified.

- Theoretical analysis is performed to reveal the difference between the proposed LASSO-BN method with methods such as VS-MSPC. The implication of the theoretical analysis is major considering that the VS-MSPC is shown to be biased regarding the design of the penalty function.

- The proposed LASSO-BN method is evaluated by extensive numerical studies in comparison with existing benchmark methods such as the VS-MSPC method. It shows that with the incorporation of the BN, the LASSO-BN method can significantly improve the diagnosis performance on identifying truly out-of-control variables.

While we limit our current scope on the BNs, it is worthy of mentioning that the BN model is very capable and flexible that can model a wide range of problems. There are a number of limitations of the proposed method that need to be investigated in future research. For example, in many manufacturing applications, the processes may have feedback loops that cannot be sufficiently represented as a BN. Rather, a more flexible network model such as those models with cyclic interactions as described in [36] can be used and further incorporated in the diagnostic monitoring formulation. It is also of interest to investigate how to conduct diagnostic monitoring when the cascade relationships between the process

variables are unknown or only partially known, then, structural learning of the BN structure will need to be integrated with the domain knowledge and the diagnostic monitoring method. Right now, the proposed method can only be applied to manufacturing processes where the cascade structure can be obtained from domain knowledge. Another future work is to theoretically investigate the effect of sample size for estimation of BN parameters, e.g., an interesting question to ask is how robust the proposed LASSO-BN formulation could be against small sample sizes. Also, it is of interest to develop a data-driven approach for determining the parameter M for applications where there is a lack of prior knowledge of the number of potential out-of-control variables. In addition, since high-dimensional processes are becoming more and more ubiquitous in many areas, how to extend the LASSO-BN formulation and scale it up for high-dimensional processes will be a very important research problem. Last but not least, it is worthy of mentioning that the identification of the truly out-of-control variables is not equivalent with the concept root-cause diagnosis, while the later one is more complicated and challenging, i.e., a hidden variable may exist in the system that lead to shifts in multiple variables, but this hidden root-cause variable could not be identified via our approach. While the root-cause diagnosis usually requires more efforts that are probably outside of the scope of statistical inference and decision-makings, our method could still provide valuable clues for locating the root-cause variables by narrowing down the search area or identifying the out-of-control variables that are the nearest variables of the hidden root-cause variables.

Chapter 3

FAULT DIAGNOSIS FOR LARGE-SCALE NETWORKED SYSTEMS VIA BOOLEAN COMPRESSIVE SENSING AND SAFE SCREENING

There are a few strong assumptions including Gaussian distribution and the known structure of the manufacturing system that could bring limits to real world diagnostic monitoring in Chapter 2. For example, in a large scale system with unknown structure, not all the signals from sensors could be realized as a continuous distribution, is there any feasible approach to tackle with it? More specifically, modern manufacturing highlights abundance of sensors for real-time monitoring. One challenge is fault diagnosis since in such an interconnected environment, the observed abnormalities are also interconnected. To tackle this challenge, we propose an integrated framework that unifies multivariate process monitoring, boolean compressive sensing, and convex optimization.

3.1 Introduction

In this paper, we investigate such a problem: when hundreds or thousands of sensors are monitoring a high-dimensional process that generate abnormality signals that are actually interconnected, how could we perform fault diagnosis that is statistically accurate and computationally feasible? It is common that thousands of control charts are operating simultaneously in nowadays' manufacturing processes. E.g., in a semiconductor manufacturing process, thousands of critical process variables should be monitored [88]. With the emerging framework such as smart manufacturing [28, 25] or advanced manufacturing that highlights unprecedented connectivity of manufacturing infrastructure and abundance of sensors for real-time monitoring of many system entities, such a large-scale process monitoring oper-

ation will be further expanded. While how to effectively monitor the variations of system entities and synthesize decentralized information into global knowledge for many system-level decision-makings have been challenging issues, we focus on one particular challenge which is fault diagnosis in an interconnected high-dimensional environment.

Besides the dimensionality issue, we recognize that the statistical distribution of variables could be complicated. It results in a fundamental difficult for conventional statistical process control methods since they rely on statistical distributions to characterize the baseline of the process. Also, in an interconnected environment, the interconnection is a challenge since the observed abnormalities are interconnected as well. But it is also a blessing in statistical sense, as it reflects inherent low-dimensionality underlying the high-dimensional signal which could be exploited to tackle the curse of dimensionality. In the literature, there has been awareness that the structure of a networked multivariate process could help process monitoring and fault diagnosis, e.g., the benefit of using cascade relation among variables (a certain kind of interconnection between variables) to help multivariate process monitoring is studied in [94]. However, existing works in this line cannot handle large-scale networked system and they often assume detailed knowledge of cascade relationship among variables and accurate statistical characterization of underlying process.

To tackle these challenges, we propose an integrated framework that unifies multivariate process monitoring, boolean compressive sensing, and convex optimization. The main purpose of this paper is to develop a proof-of-the-concept framework using the boolean formulation. The advantage of the boolean formulation is that it imposes less assumptions on the process models, e.g., it doesn't assume any kind of multivariate distribution to characterize the baseline of the process. Thus, it can be applied to a range of multivariate processes. Also, within this formulation, we could identify interesting connection with compressive sensing and convex optimization, which provides opportunities for methodological development for high-dimensional applications. Compressive sensing or sparse sampling [4] is a signal processing technique for efficiently acquiring and reconstructing a signal, by finding solutions to underdetermined linear systems where there are more unknown variables

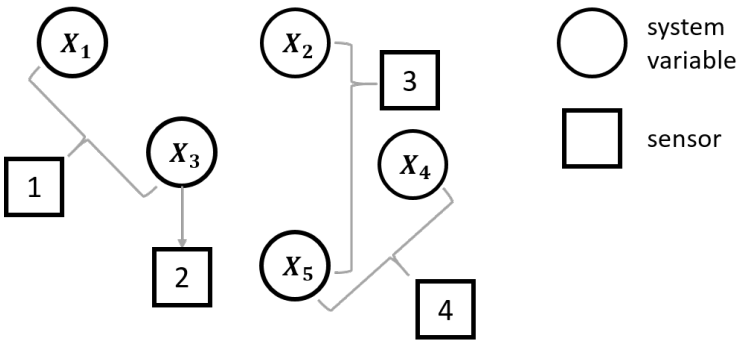


Figure 3.1: An illustration of the sensor network

than equations. Compressive sensing theory has been used to optimize sensor allocation and system monitoring recently, such as [6]. On the other hand, compressive sensing and sparse signal recovery result in challenging computational formulations which could be NP-hard [120]. To tackle computational problem, we are inspired by the recent discovery that for some optimization problems, it is plausible to conduct problem-specific "screening" analysis via exploitation of its duality and optimality condition to safely remove certain variables that will be "inactive" in the optimal solution before actually solving the problem, i.e., in fault diagnosis context, "inactive" means the variable is not a fault. Such removal of variables is known as screening in the sparse signal representation literature [39]. Thus, in this paper, our contribution is to formulate the fault diagnosis in an interconnected high-dimensional environment as a boolean compressive sensing problem, and further utilize the duality theory to develop the screening framework that will greatly reduce the computational load of real-time fault diagnosis operations.

The article is structured in the following manner. We will introduce our formulation of the problem in Section 3.2. We then develop our fault diagnosis approach in Section 3.3, and investigate its application on both simulated data sets in Section 3.4 and a real-world manufacturing process in Section 3.5. We will present conclusions and future research directions in Section 3.6.

3.2 Problem Description

In this paper, we concern an interconnected process that includes p variables. Since our primary interest is fault diagnosis, we denote the faulty status of these variables as $\mathbf{x} = [x_1, x_2, \ldots, x_p]^T$, where $x_i = 0$ means the variable is normal, otherwise $x_i = 1$. To monitor these variables, rather than assuming that each variable is monitored by a sensor, we consider a more generalized framework that sensors could be placed in the process anywhere so that a sensor could monitor multiple variables. This generalized framework has been common in many real-world applications and used in some studies as well [72, 168]. Note that this scenario includes the former scenario as a special case. Thus we denote the sensor measurements as $Z = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q]^T$ and fault detection status of the q sensors as $\mathbf{y} = [y_1, y_2, \ldots, y_q]^T$, where $y_i = 0$ means the sensor shows normal, otherwise $y_i = 1$. Note that, based on each sensor, we assume that an appropriate control chart (or in a more general sense, a data-driven anomaly detector) is built based on the sensor measurement that will return result as $\mathbf{y} = [y_1, y_2, \ldots, y_q]^T$ with $y_i \in \{0, 1\}$. Further, the relationship between sensor status with underlying faulty status of variables can be obtained from the physical layout of sensor network, characterized in boolean formulation as:

$$\mathbf{y} = (A \vee \mathbf{x}) \otimes \boldsymbol{\xi} \tag{3.1}$$

where A is a design matrix that reflects the information regarding which sensor measures which variables, $\otimes$ is boolean XOR operation, and $\boldsymbol{\xi}$ is boolean vector of errors, i.e., to account for sensor errors in fault detection. The objective of fault diagnosis is to recover $\mathbf{x}$ from $\mathbf{y}$ based on the design matrix A. For example, given a multivariate process with 5 variables and 4 sensors in Figure 3.1, the boolean formulation is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \vee \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \otimes \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix}$$

While this is a toy example to illustrate the problem formulation, it is usually an underdetermined system since the number of sensors is not necessary to be larger than the number of process variables. Also, with the disruption of sensor errors as characterized by $\boldsymbol{\xi}$, the formulation presents a challenging problem due to its combinatorial nature, which will be further exacerbated in high-dimensional applications. In Section 3.3, we will develop computational method to solve this problem.

3.3 Methodology

3.3.1 The Optimization Formulation

The essential idea of compressive sensing to mitigate underdetermined system could be borrowed here for solving Equation 3.1. This leads to the following formulation:

$$\min \|\mathbf{x}\|_0 \quad \text{subject to } \mathbf{y} = (A \vee \mathbf{x}) \otimes \boldsymbol{\xi},$$

The basic idea is to pursue sparse solution (i.e., as the use of the l_0 norm in the objective function) that could sufficiently represent the signal (i.e., as the solution has to fit the constraint). As this formulation presents a l_0 problem that is computationally challenging, we learn from the existing literature of sparse learning [39, 132] that advocates the use of convex relaxation to approximate the original l_0 problem. Following [39], this leads to the relaxed formulation 3.2. The complete derivation could be found in the Appendix.

$$
\begin{aligned}
\mathcal{P}(\mathbf{x}, \boldsymbol{\xi}) = \min \quad & \sum_{i=1}^{p} x_i + \lambda \sum_{j=1}^{q} \boldsymbol{\xi}_j \\
\text{subject to} \quad & A_{\mathbb{T}}\mathbf{x} + \boldsymbol{\xi}_{\mathbb{T}} \geq \mathbf{1} \\
& A_{\mathbb{F}}\mathbf{x} = \boldsymbol{\xi}_{\mathbb{F}} \\
& 0 \leq x_i \leq 1, i = 1, \ldots, p \\
& 0 \leq \boldsymbol{\xi}_i \leq 1, i \in \mathbb{T} \\
& 0 \leq \boldsymbol{\xi}_i, i \in \mathbb{F}
\end{aligned}
\tag{3.2}
$$

where $\mathbb{T} = \{i | y_i = 1\}$ is the set of out of control sensor signals, $\mathbb{F} = \{i | y_i = 0\}$ is the set of in control sensor signals, and $A_{\mathbb{T}}$ and $A_{\mathbb{F}}$ are the corresponding subsets of rows of A. Let $\mathbf{a}_i$ stand for the ith row of A and let $\mathbf{a}^j$ stand for the jth column of A for $i, j \in \{1, \ldots, p\}$. Furthermore, let $\mathbf{a}^j_{\mathbb{T}}$ and $\mathbf{a}^j_{\mathbb{F}}$ consist of the components of $\mathbf{a}^j$ that correspond to the index sets $\mathbb{T}$ and $\mathbb{F}$, respectively.

3.3.2 Safe Screening for Fault Diagnosis

The formulation in equation 3.2 produces a large Linear Programming (LP) problem for high-dimensional applications. As process fault diagnosis is inherently real-time online operation, solving high-dimensional LP problems in real-time is computationally demanding and probably unfeasible, undermining its potential applicability in real-world problems. Thus, we develop safe screening approaches by exploiting the convexity of the LP formulation. Our aim is to provide computationally inexpensive pre-computation which allows us to eliminate as many columns in matrix A as possible that will not affect the optimal solution.

Safe screening rules derivation

We assume that strong duality holds such that primal and dual optimal points are attained. The dual problem $\mathcal{D}(\boldsymbol{\mu})$ of the primal formulation 3.2 could be written as:

$$\begin{aligned} \mathcal{D}(\boldsymbol{\mu}) = \max \quad & \sum_{i=1}^{t} \mu_i \\ \text{subject to} \quad & \boldsymbol{\mu} A_{\mathbb{T}} \leq \frac{1}{\lambda} \mathbf{1}_p + \mathbf{1}^T A_{\mathbb{F}} \\ & 0 \leq \mu_i \leq 1, i = 1, \ldots, t \end{aligned} \tag{3.3}$$

where $\boldsymbol{\mu}$ is the dual vector and t is equal to size of $\mathbb{T}$. The complete derivation of the dual formulation could be found in the Appendix.

To derive a screening method, let's first assume that a lower bound γ on the optimal value of the dual problem is given, i.e., $\gamma \leq \mathcal{D}(\boldsymbol{\mu})$. In what follows, we will show how such a knowledge can greatly help us to perform safe screening. First, since γ is a lower bound on the dual objective function, we can safely add the corresponding lower bound constraint in the dual problem:

$$\begin{aligned} \max \quad & \sum_{i=1}^{t} \mu_i \\ \text{subject to} \quad & \mu A_{\mathbb{T}} \leq \frac{1}{\lambda} \mathbf{1}_p + \mathbf{1}^T A_{\mathbb{F}} \\ & 0 \leq \mu_i \leq 1, i = 1, \ldots, t \\ & \sum_{i=1}^{t} \mu_i \geq \gamma \end{aligned} \tag{3.4}$$

where the first constraint could be written in column vector format as

$$\boldsymbol{\mu}[\mathbf{a}_{\mathbb{T}}^1, \ldots, \mathbf{a}_{\mathbb{T}}^p] - \mathbf{1}^T[\mathbf{a}_{\mathbb{F}}^1, \ldots, \mathbf{a}_{\mathbb{F}}^p] \leq \frac{1}{\lambda} \mathbf{1}_p \tag{3.5}$$

which turns out to be $\boldsymbol{\mu}\mathbf{a}_{\mathbb{T}}^j - \mathbf{1}^T \mathbf{a}_{\mathbb{F}}^j \leq \frac{1}{\lambda}, j = 1, \ldots, p$. Due to the optimality conditions for the problem 3.3, the constraint that satisfies $\boldsymbol{\mu}\mathbf{a}_{\mathbb{T}}^j - \mathbf{1}^T \mathbf{a}_{\mathbb{F}}^j < \frac{1}{\lambda}$ at optimum means that the corresponding primal variable, i.e., x_j, is inactive, which means $x_j = 0$. This insight suggests a great opportunity for safe screening: if we could efficiently estimate the upper bound of

the constraint $\boldsymbol{\mu}\mathbf{a}_{\mathbb{T}}^{j} - \mathbf{1}^{T}\mathbf{a}_{\mathbb{F}}^{j}$ and find it is smaller than $\frac{1}{\lambda}$, we could safely remove x_j from the primal formulation. This suggests to solve the following simple LP problem:

$$\begin{aligned} T(\gamma, x_j) := \max_{\boldsymbol{\mu}} \quad & \sum_{i=1}^{t} \mu_i a_{\mathbb{T}}^{j} - \sum_{i=1}^{f} a_{\mathbb{F}}^{j} \\ \text{subject to} \quad & \sum_{i=1}^{t} \mu_i \geq \gamma \\ & 0 \leq \mu_i \leq 1, i = 1, \ldots, t \end{aligned} \tag{3.6}$$

where f is equal to size of $\mathbb{F}$. If it turns out that $\frac{1}{\lambda} > T(\gamma, x_j)$ (also named as a screening test here), we could safely eliminate the jth variable. Thus, if we could obtain closer lower bound γ to approximate $\mathcal{D}(\lambda)$ (e.g., larger γ), $T(\gamma, x_j)$ could only be smaller or stay the same. If the jth variable is indeed inactive, the ability to estimate a smaller $T(\gamma, x_j)$ will increase the likelihood of passing the screening test $\frac{1}{\lambda} > T(\gamma, x_j)$. Thus, we could conclude that the closer the lower bound γ to approximate $\mathcal{D}(\lambda)$, the more powerful screening test to identify more inactive variables.

Lower bound obtained by dual scaling

In order to get a tight lower bound γ to enhance the screening power, one way is to find a dual feasible point $\boldsymbol{\mu}$ for $\mathcal{D}(\boldsymbol{\mu})$, and then, set $\gamma = \sum_{i=1}^{t} \mu_i$, since the duality gap is always greater than or equal to 0 according to the weak duality for primal minimization linear program.

To get a dual feasible point $\boldsymbol{\mu}$, we can use a simple greedy heuristic where every nonzero component of $\mu^{\mathbf{0}}$ is 1. In other words, $\mu^{\mathbf{0}}$ corresponds to a subset s of the row indices $\{1, \ldots, t\}$ of $\mathbf{A}_{\mathbb{T}}$ such that $\sum_{i \in s} (A_{\mathbb{T}})_i \leq 1^T A_{\mathbb{F}}$; after all, $\mu^0 A_{\mathbb{T}} \leq 1^T A_{\mathbb{F}}$ with μ^0 as a binary vector implies that μ^0 is feasible for equation 3.3. We initialize s to $\emptyset$, and then, simply go through the rows of $A_{\mathbb{T}}$ in some fixed order (increasing from 1 to t), and for a row k, if

$$\sum_{i \in s} (A_{\mathbb{T}})_i + (A_{\mathbb{T}})_k \leq 1^T A_{\mathbb{F}}$$

then we set s to be $s \cup \{k\}$.

3.4 Simulation Studies

In this section, we conduct experiments to evaluate performance of the proposed method using a range of multivariate processes whose structural information are publicly available. These processes could be downloaded from the Bayesian Network Repository, `http://www.bnlearn.com/bnrepository`: Earthquake ($p = 5$), Pigs ($p = 441$), and MUNIN ($p = 1041$). We further randomly create two larger processes (DAG1 ($p = 4164$) and DAG2 ($p = 11451$)) by combining multiple MUNIN networks. These networked processes have been widely used in the literature for performance evaluation since they provide a high quality representation of the diverse processes that we may encounter in real-world applications. We further randomly assign root faults to some process variables to create the binary vector $\mathbf{x}$. To account for monitoring uncertainties, we also simulate sensor errors in the binary vector $\boldsymbol{\xi}$ with different noise level (i.e., which refers to the percentage of sensors showing error results). With a randomly generated sensor layout matrix A (when the number of sensors is given), the outcome $\mathbf{y}$ could also be generated according to Equation 3.1. We conducted experiments across different combinations of the number of sensors, number of root faults, and sensor noise levels, on all the five processes. Due to page limit, in what follows we present some representative results.

3.4.1 Evaluation of Diagnosis Accuracy

We first investigate the performance of the fault diagnosis method. Table 3.1 shows the results on the five processes under different settings. Note that, in ground truth, each process variable is either in-control (class I) or out-of-control (class II), while the diagnosis method could classify the process variables as either in-control or out-of-control. Then, if a truly in-control variable is classified as an out-of-control variable, it is a false positive; if a truly out-of-control variable is classified as an in-control variable, then it is a false negative. Therefore, this resembles a binary classification problem and the ROC curve can be used to aggregate the diagnostic performance of a range of choices of λ. The Area Under the

Curve (AUC) can be reported as an overview of the ROC curve: a more accurate diagnostic method, a larger AUC value is expected. The AUC values of the simulation results (when noise level is 0.02) are shown in Table 3.1, implying that the proposed method could lead to accurate fault diagnosis. Also, it shows that even with a relatively small number of sensors such as 20% of the process variables, a reasonable diagnosis accuracy could be achieved. The more sensors, the better diagnosis accuracy.

3.4.2 Effectiveness of the Screening Method

We then evaluate the computational savings of the safe screening method developed in Section 3, which will remove many fault-free variables before solving the LP problem. Table 3.2 gives the results. The exact number and the ratio of variables screened by the proposed safe screening method are shown in the last two columns. It shows that our safe screening method is very effective to remove many fault-free variables without even solving the LP problem.

Table 3.1: Diagnosis Accuracy (AUC) of Our Method

Process	Root Faults	Number of Sensors			
		$0.2p$	$0.5p$	$0.75p$	p
Earthquake	1	0.38	0.92	1.00	1.00
Pigs	10	0.61	0.85	0.94	1.00
MUNIN	20	0.60	0.78	0.92	0.98
DAG1	20	0.61	0.81	0.90	1.00
DAG2	20	0.56	0.81	0.91	1.00

Table 3.2: Variable Screening Results (amount)

Process	Root Faults	Variables Screened	Fraction Screened
Earthquake	1	3	0.60
Pigs	10	415	0.94
MUNIN	20	998	0.96
DAG1	20	4013	0.96
DAG2	20	11258	0.98

Table 3.3: Real World Application Results

Root Faults	Variables Screened	AUC	Time (seconds)	
			LP	Screening + LP
1	830	0.72	1.281	0.342
5	832	0.94	1.292	0.318
10	827	0.91	1.542	0.412
20	817	0.86	2.146	0.420

3.4.3 Evaluation of Running Time

We further report the computational time of the fault diagnosis method using the LP formulation as well as the two step formulation that applies screening first then solves a reduced LP problem. Our results show that the screening method is computationally efficient, which can greatly reduce the computational demand for our fault diagnosis framework. All experiments were ran in the R programming environment on an OS X system with 1.6GHz Intel Core i5, and 4GB 1600 MHz DDR3. The LP problem is solved by the R package "lpSolve". Our results are reported in Table 3.4 (when noise level is 0.02, number of sensor is $0.5p$). Here we only focus on large processes (i.e., MUNIN, DAG1 and DAG2) due to page limit. The second column shows computational time for solving the full LP problem without any screening. The last column shows computational time for screening + solving the reduced LP problem. We can see that our screening approach dramatically reduces the total time for large processes.

Table 3.4: Performance Comparison (seconds) (when noise level is 0.02 and number of sensor is $0.5p$)

Process	Solve LP	Safe Screening			
		Lower Bound	Screening Test	Reduced LP	Total Time
Earthquake	0.010	0.005	0.125	0.005	0.135
Pigs	0.398	0.005	0.341	0.005	0.351
MUNIN	2.961	0.025	0.490	0.004	0.519
DAG1	90.734	0.535	3.331	0.009	3.875
DAG2	38.521(minutes)	14.182	78.314	0.013	92.509

3.5 A Real World Application

We further study our method's performance on a real-world manufacturing process (the company's name is not disclosed here for confidentiality reason) that consists of 837 process variables and 409 sensors. The sensor layout matrix A has been given. The same procedure as in Section 4 for randomly generating root faults on the process is used here. Similarly, the diagnosis accuracy, effectiveness of the screening method, and computational time, are evaluated on this real-world process example while the results are reported in Table 3.3 (when sensor noise level is 0.02), which is consistent with results on the simulated processes.

3.6 Conclusion

We develop a boolean formulation of fault diagnosis method for high-dimensional networked systems which is robust and imposes no assumption on the statistical distribution of the process variables. While the cost is that the boolean formulation will result in a large LP problem, we further exploit the structure of the problem and conduct "screening" analysis via exploitation of its duality and optimality condition to safely remove certain variables that will be "inactive" in the optimal solution before actually solving the problem. With applications on simulated and real-world processes, the proposed method is thoroughly evaluated which shows promising performance in terms of effectiveness, accuracy, and efficiency.

As we mentioned, the main purpose of this paper is to develop a proof-of-the-concept framework using the boolean formulation. There are many future directions that could be exploited. One direction is to further equip the boolean formulation with probabilistic characterization of the underlying process and the relationship between the process variables with sensors. As we only aggregate uncertainty and errors using $\boldsymbol{\xi}$ in our current formulation without any characterization, this direction will further enhance the power of our fault diagnosis approach. We will also investigate how this approach could be useful for guiding sensor allocation for optimal fault diagnosis.

Chapter 4

HETEROGENEOUS MULTIMODAL BIOMARKERS ANALYSIS FOR ALZHEIMER'S DISEASE VIA BAYESIAN NETWORK

The Bayesian network not only could be used to formulate the cascade relation for diagnostic monitoring, but also could be applied to reflect the influences among different biomarkers in health care field, such as diagnosis of Alzheimer's disease. By 2050 it is estimated that the number of worldwide Alzheimer disease (AD) patients will quadruple from the current number of 36 million, and currently no proven disease-modifying treatments are available. Currently the underlying disease mechanisms remain under investigation, and recent studies suggest that the disease involves multiple etiological pathways. To better understand the disease and develop intervention, prevention, and treatment strategies, a number of ongoing studies including the Alzheimers Disease Neuroimaging Initiative (ADNI) enroll a large number of study participants and acquire a large number of biomarkers from various modalities including genotyping, fluid biomarkers, neuroimaging, neuropsychometric test, and clinical assessments. However, a systematic approach that can integrate all the data collected is lacking. The overarching goal of our study is to use machine learning techniques to understand the relationships between different biomarkers and to establish a system-level model that can better describe the interactions among biomarkers and provide superior diagnostic and prognostic information. In this pilot study, we use probabilistic Bayesian network (BN) to analyze multimodal data from ADNI, including demographics, MRI, PET, genotypes and neuropsychometric measurements and demonstrate our approach to have superior prediction accuracy.

4.1 Introduction

Alzheimer's disease (AD) is a highly prevalent neurodegenrative disease, and is widely recognized as a major, escalating epidemic and a world-wide challenge to global health care systems. Considerable research efforts have been devoted to establish a disease model of AD that could lead to greater understanding of the events that occur in AD. One major development is the discovery of the Aβ hypothesis that assumes AD begins with abnormal processing of transmembrane Aβ precursor protein. Such a malfunction of the metabolism will trigger a series of pathological events, resulting in the toxic beta-amyloid plaque in human brain which is characteristics of AD.

This disease model has been articulated in Jack et al [76] who presented a hypothetical model for biomarker dynamics in AD pathogenesis. The model begins with the abnormal deposition of Aβ fibrils, as evidenced by a corresponding drop in the levels of soluble Aβ42 in cerebrospinal fluid (CSF) and increased retention of the positron emission tomography (PET) radioactive tracer [11C]-labeled Pittsburgh compound B (11C-PiB) in the cortex. This will result in subsequent neuronal damage that can be captured by increased levels of CSF tau protein and synaptic dysfunction follows that can be evidenced by decreased [18F]-fluorodeoxyglucose (FDG) uptake measured by PET. As neuronal degeneration progresses, atrophy in certain areas typical of AD such as hippocampus regions becomes detectable by magnetic resonance imaging (MRI). So far, Jack's model has been widely studied, confirmed, refined, and enriched. While many details in the disease model are still unknown, investigators from academia and the pharmaceutical industry have been actively developing biomarkers to gain better and more accurate knowledge of the mechanisms of AD pathology to facilitate a range of clinical tasks such as early diagnosis, treatment effect evaluation, treatment planning, better clinical trial design and drug developments.

While most of the existing efforts mentioned above focus on single modality of biomarker analysis, recently, there have been a few studies that proposed to study many biomarkers of heterogeneous nature jointly. For instance, Ye et al [163] integrated multiple complementary

data and initiated the work to use the multiple kernel learning method for multimodal integration for AD research. Shen et al did a sequence of work on multimodal classification [167] and regression [166] based on multimodality data, and achieved better prediction accuracy than those model with single biomarker. However, most of these works focus on prediction. Less effort has been devoted to study the interactions of these multimodal biomakers for better understanding of the disease as a whole.

Thus, in our study, we take a systematical perspective to study patterns of disease progression. We take into consideration of multimodal biomarkers such as APOE types, SNP variants, demographics, FDG-PET, AV45-PET, MRI, and neuropsychological assessment. We adopt a powerful machine learning model, the Bayesian Network (BN), as the major tool for studying the influential relationships among the variables. A main premise of using BN model for multimodal biomarker integration is that it could provide more details regarding the potential mechanism of the disease progression, than those black-box prediction models [163, 167, 166]. Specifically, while the existing black-box prediction models throw in all the multimodal biomarkers as predictors parallel in the prediction equation regardless of their heterogeneous clinical nature, their clinical roles are not revealed since each biomarker is assigned with a quantitative weight in the prediction equation that only determines whether or not the biomarker is important. Moreover, this weight is not an absolute presentation of evidence, as it is essentially a multivariate concept that depends on the existence of other biomarkers in the equation. This results in the risk of excluding important biomarkers which hold significant clinical value but not significant statistical prediction value due to redundancy with other biomarkers. Also, from these black-box prediction models, there is no indication of how the biomarkers influence each other, whether or not some biomarkers mediate the effects from other biomarkers to disease outcomes. Presumably, the relationships between the multimodal biomarkers could be very complex, and our study is motivated by the lack of capacity of existing multimodal biomarker integration methods to discover and model these relationships. On the other hand, although not a causal model, BN models have been found very effective in a range of applications to study the "layers" of influence among

variables. It could lead to very useful knowledge regarding the "chain reaction" of a sequence of events captured by the biomarkers' measurements. BN is a powerful data-driven model that seeks the best mechanism model that is consistent with a set of measurements from a cohort of patients. Thus, it translates naturally into a semantic description of the disease similar to a clinician's intuitive description of its progression.

The reminder of the paper is structured as follows: In Section 4.2, we will provide description of the dataset that will be used in this study and the BN, particularly, the mixed type probabilistic Bayesian network due to the heterogeneous nature of the biomarkers; In Section 4.3, we will present the learning results and validation efforts; We then conclude our study in Section 4.4.

4.2 Methods

4.2.1 Data

The data used in this paper were obtained from ADNI database `www.loni.ucla.edu/ADNI`. The primary goal of ADNI has been to test whether the serial MRI, PET, other biological markers, and clinical and neuro-psychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, aged 55 to 90, to participate in the research approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years.

Table 4.1: Subject Information at Baseline

	AD(n=114; 47M/67F)			MCI(n=283; 128M/155F)			HC(n=120; 52M/68F)		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Age	74.6	8.1	56.5-89.6	73.9	6.7	58.5-90.6	73.4	7.3	55.0-89.6
Edu	16.0	2.6	8.0-20.0	16.4	2.7	9.0-20.0	16.8	2.6	9.0-20.0
MMSE	23.8	1.6	20.0-26.0	27.0	2.1	24.0-30.0	28.8	1.9	24.0-30.0
ADAS	15.5	7.8	4.0-51.0	14.6	9.5	0.0-51.0	10.8	8.8	3.0-31.0

4.2.2 Subjects

The ADNI general eligibility criteria are described at `www.adni-info.org`. Briefly, subjects are between 55 and 90 years of age, having a study partner able to provide an independent evaluation of functioning. Specific psychoactive medications will be excluded. General inclusion/exclusion criteria are as follows: 1) healthy subjects: MMSE scores between 24 and 30, a Clinical Dementia Rating (CDR) of 0, non depressed, non MCI, and non demented; 2) MCI subjects: MMSE scores between 24 and 30, a memory complaint, having objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II, a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia; and 3) Mild AD: MMSE scores between 20 and 26, CDR of 0.5 or 1.0, and meets the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD.

Our study includes the baseline measurements of 517 ADNI subjects. The cohort contains 114 AD patients, 283 MCI patients, and 120 healthy controls. Table 4.1 lists the demographics of these subjects.

4.2.3 Biomarkers

The description about biomarkers to be analyzed is listed in Table 4.2. These biomarkers are heterogeneous in terms of both clinical nature and statistical characteristics. While this list is still limited, it provides a good presentation of the genetic, demographic, neuroimaging, and clinical aspects of the disease. Among these markers, some are categorical biomarkers, such as sex (male or female) and SNPs (carrier or non carrier), while some are numeric biomarkers such as some clinical measurements. Note that, we also include some SNPs variants which are the top genetic risk factors for AD reported at `http://www.alzgene.org/TopResults.asp`.

4.2.4 Probabilistic Bayesian Network

A BN is a graphical model that characterizes the influential relationships among variables $X = \{X_v; v \in V\}$. Let $D = (V, E)$ be a Directed Acyclic Graph (DAG), where V is a finite set of nodes and E is a finite set of directed edges between the nodes. The DAG defines the structure of the BN. Each node $v \in V$ in the graph corresponds to a random variable X_v, i.e., in our study, a biomarker is a variable. In the DAG, the relationship between each variable X_v with its parents variables denoted as $pa(v)$ can be characterized as a conditional probability distribution, $p(x_v|x_{pa(v)})$. Then, the joint probability distribution of a BN could be deduced as

$$p(x) = \prod_{v \in V} p(x_v|x_{pa(v)}) \tag{4.1}$$

For this reason, the set of conditional probability distributions for all variables in the network, denoted as $\mathcal{P}$, is called the parameter of the BN. A Bayesian network for a set of random variables X is then the pair $(D, \mathcal{P})$.

Table 4.2: Description of Heterogeneous Multimodal Biomarkers

Biomarker	Description
Age	Age
Sex	Gender
Edu	Years of education
FDG	Average FDG-PET
AV45	Average AV45 SUVR
HippoNV	the normalized hippocampus volume
APOE4	Apolipoprotein E4 polymorphism
rs3818361	CR1 gene rs3818361 polymorphism
rs744373	BIN1 gene rs744373 polymorphism
rs11136000	Clusterin CLU gene rs11136000 polymorphism
rs610932	MS4A6A gene rs610932 polymorphism
rs3851179	PICALM gene rs3851179 polymorphism
rs3764650	ABCA7 gene rs3764650 polymorphism
rs3865444	CD33 gene rs3865444 polymorphism
MMSE	Mini-Mental State Examination
ADAS-cog	Alzheimer's Disease Assessment Scale

4.2.5 Mixed Type Bayesian Network

In this paper, we adopt the mixed type Bayesian network model that handles both discrete and continuous variables, which is developed in [92]. For mixed type BNs, the set of nodes V can be further specified as $V = \Delta \cup \mathrm{T}$, where Δ and T are the sets of discrete and continuous nodes, respectively. The set of variables X can then be denoted as $X = \{X_v; v \in V\} = (I, Y) = \{(I_\delta, Y_\tau); \delta \in \Delta, \tau \in \mathrm{T}\}$, where I and Y are the sets of discrete and continuous variables, respectively. For a discrete variable δ, we let $\mathcal{I}_\sigma$ denote the set of levels.

It has been a challenge to model the mixed type Bayesian network. As mentioned earlier, a BN consists of the structure D and the parameter $\mathcal{P}$. The central challenge for modeling mixed type Bayesian network is the development of appropriate models for characterizing $\mathcal{P}$. In our study, we follow the seminar works in [92] that models the joint probability distribution by factorizing it into a discrete part and a mixed part, so

$$p(x) = p(i, y) = \prod_{\delta \in \Delta} p(i_\delta | i_{pa(\delta)}) \cdot \prod_{\tau \in \mathrm{T}} p(y_\tau | i_{pa(\tau)}, y_{pa(\tau)}) \tag{4.2}$$

where the first part of products of conditional probabilities is for discrete nodes, and the second part is for continuous nodes.

For discrete nodes, conditional probabilities are parameterized as

$$\theta_{i_\sigma | i_{pa(\sigma)}} = p(i_\sigma | i_{pa(\sigma)}, \theta_{\sigma | i_{pa(\sigma)}}), \tag{4.3}$$

where $\theta_{\sigma | i_{pa(\sigma)}} = (\theta_{i_\sigma | i_{pa(\sigma)}})_{i_\sigma \in \mathcal{I}_\sigma}$. The parameters are subject to the constraints that $\sum_{i_\sigma \in \mathcal{I}_\sigma} \theta_{i_\sigma | i_{pa(\sigma)}} = 1$ and $0 \leq \theta_{i_\sigma | i_{pa(\sigma)}} \leq 1$.

For continuous nodes, the local probability distributions are Gaussian linear regressions on the continuous parents with parameters depending on the configuration of the discrete parents, as shown in below:

$$\theta_{\tau | i_{pa(\tau)}} = (b_{\tau | i_{pa(\tau)}}, w_{\tau | i_{pa(\tau)}}, \sigma^2_{\tau | i_{pa(\tau)}}), \tag{4.4}$$

so that

$$Y_\tau | i_{pa(\tau)}, y_{pa(\tau)}, \theta_{\tau|i_{pa(\tau)}}$$
$$\simeq \mathcal{N}(b_{\tau|i_{pa(\tau)}} + y_{pa(\tau)} w_{\tau|i_{pa(\tau)}}, \sigma^2_{\tau|i_{pa(\tau)}}).$$

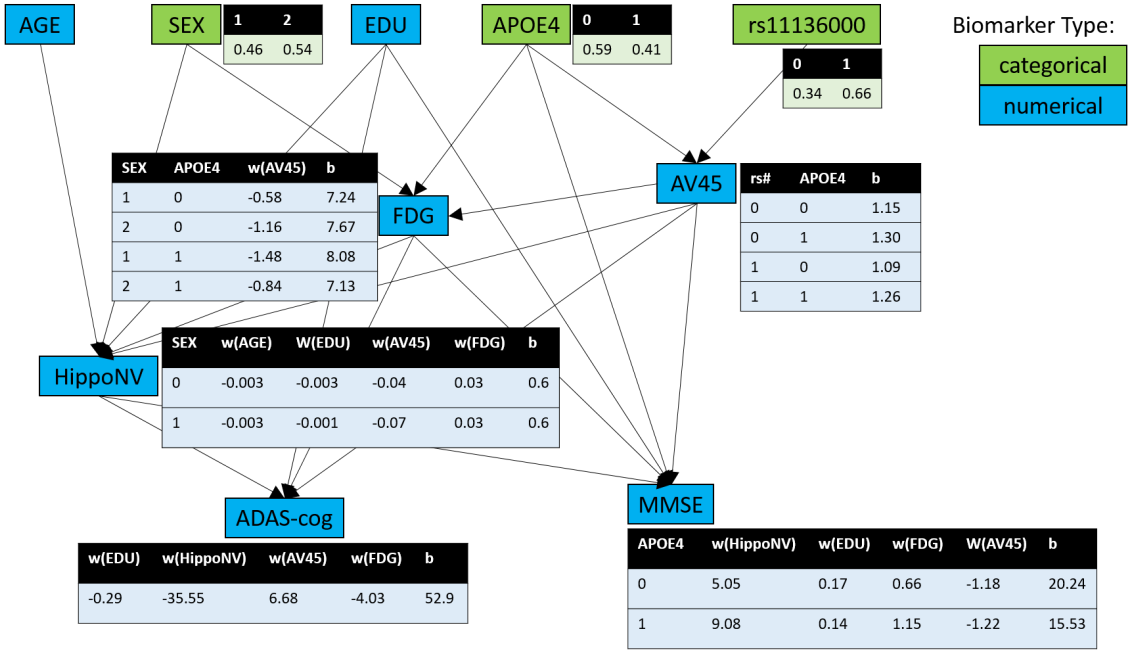


Figure 4.1: Learn Mixed Type Bayesian Network using Heterogeneous Multimodality Data at Baseline.

4.2.6 Learning of Mixed Type BN from Data

With the BN model specified for mixed type variables, the next task is to identify a structure learning algorithm that can find the optimal DAG structure. This is called the structure learning problem in the BN literature. The basic formulation of this problem, according to the score-based method, starts with a dataset T and a scoring function ϕ. Then, the task is to find a Bayesian network $B \in \mathcal{B}_n$ that maximizes the values $\phi(B, T)$. The standard

methodology is to use search algorithms, such as heuristic search, greedy hill-climbing, genetic algorithms, tabu search and more, conducted over eligible search space $\mathcal{B}_n$ to search the DAG structure that maximizes the score. In this study, we use the score function developed in [142] for mixed type BN, which can be readily implemented in the R package "bnlearn" [142]. After having identified the optimal DAG structure, parameter estimation could be conducted via maximum likelihood estimation according to (4.2).

4.3 Results

We then apply the mixted type BN on the heterogeneous biomarkers of the ADNI cohort we have collected. Missing values in the dataset are imputed by the median value of corresponding biomarkers. In order to identify a stable DAG structure, first, we use Bootstrap method to generates 100 new training sets by sampling the original data set with replacement, then, learn the optimal DAG structure on each bootstrapped dataset. We then derive the final DAG structure by keeping those arcs which appear at least in half of these DAG structures learned from bootstrapped datasets. This strategy has been suggested in previous works for BN applications [43] that has been found effective to robustify the learning result. Note that, here, we also utilize the prior knowledge in the learning of the DAG structure, i.e., the genetic factors could be parents of other factors not the other way around, while the disease outcome variables such as ADAS-cog and MMSE score could only be in the bottom of the BN model. This prior knowledge is used in the BN learning and greatly reduces the search space of the eligible DAG structures.

The final BN model is shown in Figure 4.1. We use green to represent categorical variables while using blue to represent numerical variables. The probability tables of categorical variables, and the parameters of the conditional Gaussian distribution w, b for continuous variables, are shown along the DAG structure as well. For example, node HippoNV in Figure 4.1 has five parents: sex is binary when other four are numerical. The relationship between the HippoNV with other continuous variables such as AGE, EDU, AV45 and FDG is characterized as a regression model, while parameters of this regression model vary according

to the categorical variable SEX.

Overall, this network structure is consistent with existing knowledge in AD literature. As expected [139, 20, 115, 35, 164], the APOE e4 was associated with higher amyloid burden (as measured by AV45 PET imaging) and lower cerebral glucose metabolism (as measured by FDG PET). A direct impact of e4 with MMSE score was also identified in our results in agreement with previous reports [9, 32], although its underlying mechanism warrants further investigation. An association of the SNP rs11136000 with amyloid burden was also identified, in agreement with the potential role of clusterin (CLU, the gene that SNP rs11136000 is associated with) in Aβ clearance [84, 140]. Based on this study, it is also identified that there were direct relationships between amyloid burden and cognitive performance which may reflect the direct neurotoxic effect of Aβ and its derivatives or indirect impact through pathways that were not represented in the biomarkers we included in this study [69, 70, 60]. The direct interaction between cerebral glucose metabolism and cognitive function as identified in this study was also in agreement with prior knowledge [96, 101, 116, 119]. The identified relationship between years of education and the cognitive performance be a cognitive reserve effect as reported by a number of studies [148, 146, 147]. In summary, using probabilistic Bayesian network, we identified inter-biomarker relationships that are in good agreement with existing knowledge about AD.

4.3.1 Evaluation of the Prediction Accuracy with BN

Besides comparing our results with AD literature, we further pursue numerical validation. Specifically, as "MMSE" and "ADAS-cog" are two important clinical outcomes, it is of interest to see if the learned BN owns significant prediction capability of the two outcomes. Thus, in this section, we compare the prediction capability of BN with three common regression techniques (implemented in R environment), such as linear regression (lm()), decision tree (rpart()), and random forest (randomForest()). The target metric we would like to measure and compare is mean square error (MSE), which serves as the goodness of fit in regression problem. We use 10-fold cross validation to obtain unbiased estimates of MSE. To set up

cross validation procedure, we randomly divide the original dataset into training set (70% observations) and testing set (30% observations) in each round.

Table 4.3: 10 fold cross validation MSE result

	Mean (SD)	
	MMSE	ADAS-cog
Bayesian network	2.810 (0.441)	35.380 (3.244)
Linear regression	3.125 (0.439)	38.748 (4.364)
Decision tree	3.758 (0.552)	42.195 (4.306)
Random forest	2.914 (0.330)	35.218 (4.932)

Table 4.3 lists the mean and standard deviation of MSE of two regression models. In terms of the average of the MSE, the BN achieves a better accuracy than the linear regression and decision tree in both MMSE and ADAS-cog prediction, while its performance is close to the random forest which has been known to be a very powerful prediction model despite its black-box nature. Similar observation could also be made in terms of the variance of the MSE.

4.3.2 Validation of the Identified BN via the Covariance Patterns

We also innovate here to analyze the covariance patterns to help validate the learned BN model. The covariance patterns essentially characterize the undirected associations among variables. Thus, a BN model that aims to explain the influential relationships between the variables is expected to be able to explain the associations that are observed in data. Specifically, to derive the associations among variables, we use Pearson correlation for continuous variables, polychoric correlation for categorical variables, and polyserial correlation for a categorical variable and a continuous variable. The heterogeneous correlation matrix is computed using R package "polycor". Figure 4.2 shows the associations we have observed

from the biomarkers. Each row/column represents one biomarker. The color intensity shows the strength of an association. Note here we only present the magnitude of the associations to focus the purpose on validation with the BN model. Overall, the association patterns revealed in Figure 4.2 is quite consistent with our learned BN model. For instance, from Figure 4.2 it is clear that the ADAS-cog is strongly associated with the variables FDG, AV45, HippoNV, and APOE4. While this is consistent with the BN as shown in Figure 4.1, we also notice that in Figure 4.2 we could not detect that the association between APOE4 with ADAS-cog could be mediated by the variable FDG. Thus, by learning the BN model, we could identify more layers in the relationships between the variables and could shed light to useful discoveries of the underlying mechanism of the disease progression.

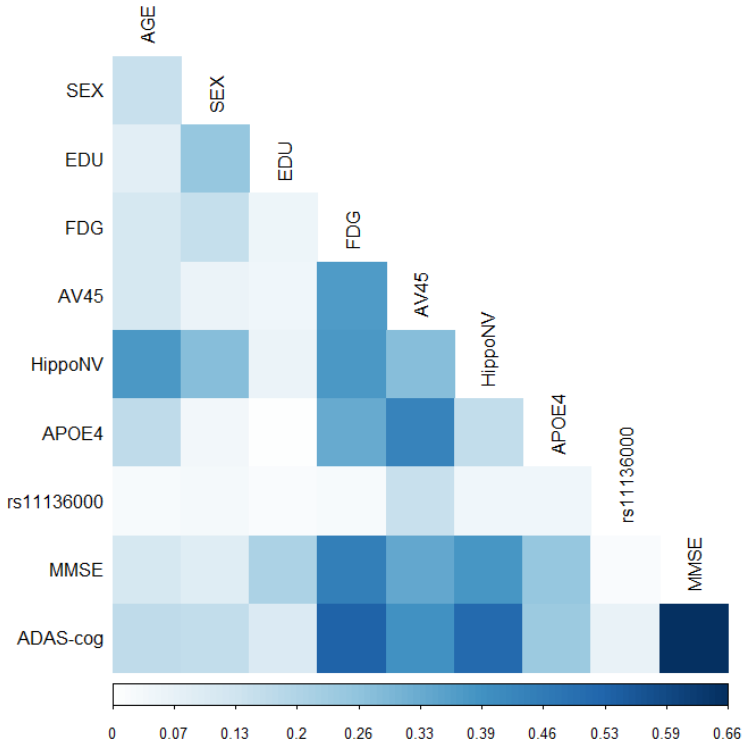


Figure 4.2: Visualization for Heterogeneous Correlation Matrix

4.3.3 Validation of the Identified BN via RuleFit

In order to validate the structure of the learned BN, another approach we propose to use is the RuleFit [43] method. RuleFit is a powerful method to discover complex interactions among variables. Again, it is a predictive model, so it lacks the capability of the BN to provide possible explainations of the relationships among the variables. But in the same spirit as the use of the association patterns to validate the BN model, we hope to see consistence between the BN structure with the interaction patterns the Rulefit could identify.

Thus, we apply the Rulefit on our data to identify the interactions among the biomarkers that can predict the two outcomes, MMSE and ADAS-cog. Table 4.4 lists the five rules we have identified. Column 1 gives the scaled importance for each rule. Column 2 (support) refers to the fraction of the samples in the dataset to which the rule applies. Apparently, it seems that there is great consistence between the two methods, while the BN model can provide more details of the underlying relationships among the variables.

4.4 Conclusion

In this paper, we propose to use the mixed type Bayesian network to model the interactions among heterogeneous multimodal biomarkers. We conduct this study using ADNI baseline dataset, and find that the learned BN model provides findings that are consistent with the AD literature. We further validate the learned BN structure via the prediction accuracy of clinical outcomes, capability to explain association patterns among variables, and comparison with powerful feature selection method. In future work, we would like to investigate the use of dynamic BN models to incorporate the temporal data that is available in ADNI dataset. Critical changes of the biomarkers that may indicate disease progression may be discovered and how these significant clinical events could be synthesized to be a systematical disease model is a very interesting and exciting research direction.

Table 4.4: RuleFit: 10 Most Important Rules

Impo.	Supp.	Rule
		y: MMSE
100	0.78	61.85<AGE<86.85 & HippoNV>0.38
91.3	0.81	AGE<85.75 & FDG>5.78
74.6	0.15	FDG<5.85 & AV45>1.11
68.2	0.06	EDU<19.5 & HippoNV<0.38 & APOE4=1
46.9	0.75	5.76<FDG<7.25
		y: ADAS-cog
100	0.73	FDG>5.75 & HippoNV>0.39
62.5	0.65	FDG>4.9 & 1.02<AV45<1.51
41.2	0.72	EDU<19.5 & HippoNV<0.55
41	0.44	FDG>6.34 & rs3764650=0
35.9	0.17	1.23<AV45<1.63 & rs744373=0

Chapter 5

OPTIMAL EXPERT KNOWLEDGE ELICITATION FOR BAYESIAN NETWORK STRUCTURE IDENTIFICATION

Bayesian network (BN) has been a popular tool for gaining mechanistic understanding of variables by revealing how the variables influence each other. It has been found very effective in a few studies in quality control and process monitoring. However, for complex problems where the structure of a BN is unknown, a common approach is to learn the BN structure from observational data. A fundamental bottleneck of this approach is that observational data can only be used to discover part of the influential relationships among variables. To overcome this problem, we propose to combine observational data and expert knowledge. To the best of our knowledge, our approach is the first of its kind that formulates an experimental design framework to automate the expert elicitation process and collect the most informative expert knowledge, optimally matched to the observational data, to learn the BN structure.

5.1 Introduction

Bayesian network (BN) is a graphical model for representing influential relationships among variables. It has a directed acyclic graph (DAG) structure, which is a directed graph with no cycles and thus can encode topological orderings. Such topological orderings could be used to model the influential relationships among variables. It is also interpreted as a causal model in some applications where some strong assumptions can be imposed to establish the equivalence between the statistical dependency among variables as causality [126]. No matter whether or not the causality can be derived, BN models have been a popular tool for gaining mechanistic understanding of variables by revealing how the variables influence

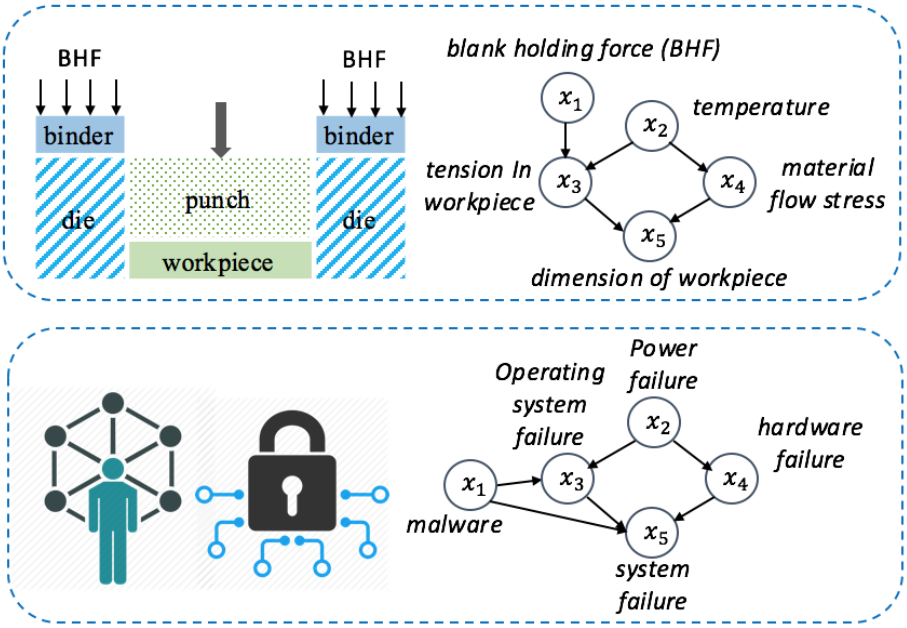


Figure 5.1: A BN derived from a manufacturing hot-forming process (upper) and a BN derived from a computer system security monitoring diagram (lower).

each other. Its popularity has been evidenced by its wide applications in many fields such as genetics in [47, 48, 172], ecology in [130], social sciences in [54, 65], biomedical informatics in [122], brain sciences in [73, 74], manufacturing in [59], and quality control and monitoring in [94, 158, 99, 100, 96].

Figure 5.1 provides an illustration of two real world applications where BNs are derived from data and then used to facilitate decision makings. The upper example shows the influential relationships among system variables in the hot forming process, where the final dimension of workpieces could be influenced by a variety of direct and indirect causes. The lower example shows a computer system security monitoring diagram, which serves as a diagnosis map to identify root causes of system failure. For both cases, representing influential relationships using BNs not only facilitates the fault diagnosis procedure, but also simplifies the statistical modeling of a joint distribution of system variables.

However, learning the DAG structure of variables appears to be a very challenging task. As a BN essentially embodies a joint distribution, statistical estimation approaches have been developed to learn the DAG structure based on observational data, which are commonly assumed to be randomly sampled from the underlying joint distribution. This approach has been extensively studied in [67, 48, 118, 23, 21, 20]. However, it has been found that the theoretical bottleneck of these methods is that observational data can be used to only discover part of the influential relationships, encoded in the so-called "essential graph" (or "equivalent class"). The essential graph of a BN is a mixed graph that includes both directed arcs and undirected arcs. A deeper reason of this limitation is that, merely from observational data, we could only identify statistical dependency relations between variables. Thus, the DAG structures that imply the same set of dependency relations between variables are not distinguished by observational data alone. For instance, the DAGs (b) and (c) in Fig. 5.2 cannot be distinguished by observational data alone, since all of them encode the same independence relations. Thus, to augment the observational data and identify all influential relationships among the variables, a common philosophy is to pursue experimental design strategies for intervention data collection. For instance, considering the two variables in Fig. 5.2, x_3 and x_4. If intervention can be imposed on x_3 and it turns out that the distribution over x_4 does not change, while intervention on x_4 does change the distribution over x_3, then it implies that the edge connecting x_3 and x_4 should be $x_4 \to x_3$. Motivated by this observation, a line of research works have been spurred to develop optimal experimental design methods to maximize the likelihood of learning the DAG structure with minimized number of interventions. Exemplary works include [137, 38, 37].

In our study, we pursue another line of philosophy to augment the observational data for DAG learning, which is the expert knowledge elicitation. We focus on a particular type of elicitation operation that acquires pairwise comparison between variables, i.e., ask an expert if a variable is likely to be upstream of another variable. This is the most common form of expert knowledge regarding the influential relationships between variables. While pairwise comparison owns the advantages such as ease to implement, on the other hand, the number

of potential pairwise comparison grows exponentially with the number of variables. Thus, we aim to develop a computational framework that can generate an optimal set of operations for expert knowledge elicitation regarding the ordering of the variables. Note that, expert knowledge elicitation has been studied in the BN literature, mostly for parameter learning rather than structure learning. There is a few works such as [40, 91] that studied the use of expert knowledge elicitation for structure learning, however, these are heuristic procedures that are not scalable, neither automatically optimized. Also, due to their qualitative nature, how they can be optimally integrated with learning algorithms based on observational data is also lacking.

Comparing with the approaches that use interventions to perturb the system in order to learn the influential relationships [137, 38, 37], our approach is more cost-effective and can be applied to some applications where intervention is physically hard to conduct. The following application provides such an example that is common in real-world applications, but is outside of the scope of existing methods. For instance, the Key Performance Indicator (KPI) has been a very important concept in business analytics and performance management, drawing increasingly attentions from many corporations to measure and monitor many KPIs of their interest on daily basis if not hourly or minutely. It has been pointed out in the literature such as [143, 112, 15] that the key to analyze the KPIs, and to further convert them into valuable business decision-makings, is to study the influential relationships between the KPIs. In other words, it is important to understand which KPIs drive which KPIs, so management or investment strategies can be better informed and implemented. However, although an abundance of KPI measurements can be obtained, whether or not we could learn this "mechanistic understanding" of the KPIs from observational data is up to debate, and how to learn it is still an open question. On the other hand, expert knowledge has been found very useful to identify the influential relationships among the KPIs in [141]. Apparently, knowledge-based practice has the difficulty of being scaled up. Also, it lacks the flexibility to incorporate the objective information encoded in the observational data. Thus, there is a need to develop a computational framework that can learn the influential

relationships among the KPIs by using both the observational data and expert knowledge automatically and cost-effectively elicited to augment the observational data.

Our general approach is to first develop a Bayesian learning framework that can combine the two types of data. This is plausible, since given a specific DAG structure, the likelihood of the observational data can be analytically derived based on the corresponding joint distribution the DAG encodes ([50, 26, 67, 45, 23]). On the other hand, we further develop another probabilistic framework to model the expert pairwise comparison data. Then, within a Bayesian learning framework, both sources of data can be combined to obtain the probability distribution of the possible BN models. Based on this probability distribution, uncertainty of the ordering of the variables can be evaluated which will provide critical evidence for us to better collect new data via expert knowledge elicitation, that can maximally reduce the uncertainty of our estimation of the ordering of the variables.

The rest of the paper is organized as follows: We will introduce the basic concepts and background of BN, and some BN structure learning algorithms from observational data in Section 5.2. Then, we will present our proposed method in Section 5.4. We will conduct extensive numerical experiments in Section 5.4 to show that the proposed approach outperforms baseline approaches across a number of benchmark BN models and different levels of expert knowledge accuracy. We further implement the proposed method on two real-world applications, one in healthcare and another one in business analytics in Section 5.5. Finally we conclude our work and discuss future directions in Section 5.6.

5.2 Background and related works

A Bayesian network (BN) over a set of random variables $\boldsymbol{x} = \{x_1, ..., x_p\}$ is a set (G, θ_G) that represents a distribution over the joint space of x via chain rule: $P(x_1, ..., x_p) = \prod_i P(x_i|U_i)$, where U_i is the parent set of x_i according to the DAG structure of the BN, and θ_G is the corresponding parameter. The DAG structure that encodes the parent-child relation of the variables is commonly denoted as $G = \{V, E\}$, where V denotes for the p nodes (each node is a variable) and E is the edge set. Learning the BN structure from observational data

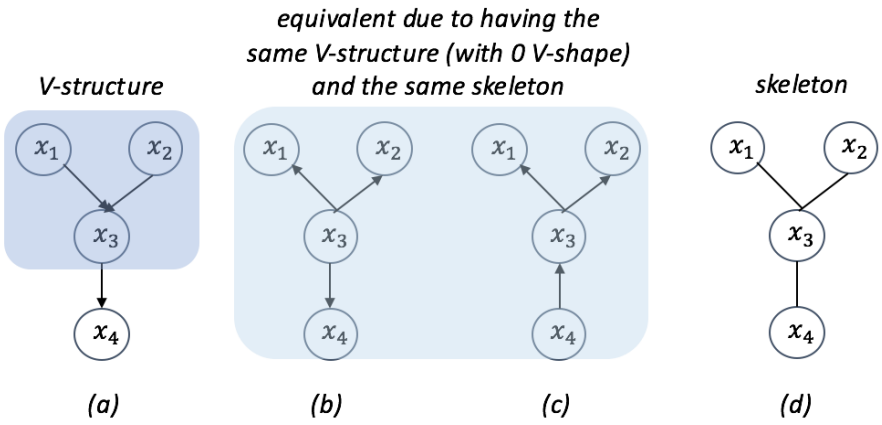


Figure 5.2: Illustration of Markov Equivalence. (a) is a V-structure with V-shape showing common effect of x_1 and x_2 on x_3. (b) and (c) are Markov equivalent. (d) is the skeleton of (a)-(c).

refers to the challenge to find the optimal DAG structure G that maximizes a certain score which evaluates the goodness-of-fit of the DAG structure to the observed data. For instance, a Bayesian Score was developed in [26] for discrete BNs while another score was developed in [67] for Gaussian BNs. It is commonly assumed that the observational data are randomly sampled from the joint distribution of the variables specified by the BN model, so most of the score functions are developed based on this probabilistic framework ([26, 67]). There are some other information-theoretic score functions developed as well in [30, 133], but still the fact that a BN is a structured representation of a complex joint distribution of the variables lays the foundation of these score functions. With a score function, a search procedure is commonly used to search through the eligible DAG structures to identify the optimal DAG. There is another line of BN learning algorithms, named as constrained-based algorithms ([104, 30]), which use hypothesis testing methods to identify the dependency structure of the variables. We omit the details of these algorithms here since our method is mostly related to the score-based algorithms.

5.2.1 Observational Equivalence

Despite the success of the BN structure learning algorithms in many applications, it has been found that the theoretical bottleneck of these methods is that observational data can be used to only discover part of the directional relationships, encoded in the so-called "essential graph" (or "equivalent class"). As a matter of fact, the validity of learning the DAG structure of a BN from observational data is established on two assumptions, namely that the BN and its encoded joint distribution obey the faithfulness assumption and causal sufficiency. The faithful assumption is to ensure that the independence relationships in both the graph structure and the underlying joint distribution are equivalent. Causal sufficiency means that there are no latent variables. While these assumptions establish the theoretical foundation for learning the DAG structure from observational data, it has also been found that, only the essential graph can be discovered with observational data alone. A brief elaboration of the essential graph involves the concepts *skeleton* and *v-structures*. As illustrated in Figure 5.2(d), the skeleton of G is the undirected graph on V where every arc in E has been rendered as undirected. A *v-structure* describes the structure that two variables both have effects on a third variable, e.g., both x_1 and x_2 have directed arcs to node x_3. In the illustration, Figure 2(a) shows a simple v-structure. It is known that two DAGs, if having the same *skeleton* and the same sets of *v-structures*, will be indistinguishable by observational data alone([157]). These DAGs that are indistinguishable by observational data are also said to belong to the same Markov equivalence class. Based on this result, the concept, "essential graph", was developed. The essential graph is a mixed graph, where an edge is oriented if and only if it has the same orientation in every DAGs in the equivalence class. The essential graph essentially encodes the maximum set of directed arcs that we could learn from the observational data under the faithfulness assumption and causal sufficiency.

5.2.2 Intervention Data and Expert Elicitation

Motivated by the limitation of learning the structure from data, taking experiments to collect intervention data has been held as a promising means for learning the full DAG structure. The common setting for intervention data collection assumed in the literature is, as illustrated in Section 5.1, to intervene on some variables and observe the influence on other variables. A few studies have been developed to optimize the intervention strategies, i.e., such as to minimize the number of interventions. It has also been studied regarding how many experiments are required to discover G. This line of research was initiated in a series of works by Eberhardt, Glymour, and Scheines [137, 38, 37], while most of them focused on single-variable interventions, i.e., each time, an intervention is imposed on one variable only. Eberhardt considered multi-variable interventions to be far more complicated to analyze [38]. [153] proposed another methodology that is not based on intervention, rather, it is more like a query operation to selectively sample from the BN. Obviously, our proposed expert elicitation method is fundamentally different from this line of works. We have pointed out in Section 5.1 that there are a few works such as [40, 91, 17] that studied the use of expert knowledge elicitation for structure learning, however, these are heuristic procedures that are not scalable for large-scale applications. Also, due to their qualitative nature, the interface with the learning algorithms based on observational data is also lacking, and there is lack of systematic optimization formulation to automate the expert elicitation process.

5.2.3 Existing Works of BN in Quality Control

The literature of BN applications on engineering problems is vast. Here, we illustrate the relevance of BN for engineering problems using quality control applications. Over the last decade, the BN has become a popular representation for encoding uncertain knowledge and has been used to improve a number of quality control and system monitoring tasks, particularly for improving the root-cause diagnosis and sensor allocation if the underlying process model can be represented by a BN [95, 158, 99, 100, 94]. As illustrated in Figure.

1, if x_1 is out-of-control, its effect will propagate to x_3 and further impact x_5, resulting in out-of-control signals on all these variables. Modeling the cascade relationships between the five control variables during the manufacturing hot-forming process would expedite the root-cause diagnosis procedure to identify x_1 as the root of this chain reaction. More examples of how BN can be useful for quality control and monitoring can be found in [95, 158, 99, 100, 94].

However, for the aforementioned applications, a crucial assumption underlying these works is the availability of an accurate BN model to represent the process, which is actually very hard to obtain in many applications.

5.3 Methodology

Our proposed optimal expert elicitation framework, as illustrated in Figure 5.3, can be decomposed into the learning and sensing modules. In the learning module, we develop a Bayesian framework to combine both observational data and expert comparison data to obtain a posterior distribution over ordering of the variables. In the sensing module, the optimization model will identify the most informative new comparisons data that should be collected from the expert, which can maximally reduce the posterior uncertainty of the ordering of the variables.

In this paper, we use lower-case letters, e.g., x, to represent scalars, bold-face lower-case letters, e.g., $\boldsymbol{v}$, to represent vectors, and bold-face upper-case letters, e.g., $\boldsymbol{W}$, to represent matrices.

5.3.1 A Bayesian learning framework for combining observational data and expert elicitation data

Denote observational data and expert comparison data as D^{obs} and D^{ex} respectively. D^{obs} consists of n random samples of variables $\boldsymbol{x} = \{x_1, \ldots, x_p\}$, which are i.i.d observations from the joint distribution of variables represented by the underlying BN. D^{ex} consists of two parts. First, note that, for a networked system with p variables, there are $N = \binom{p}{2}$ possible pairwise comparisons. For each comparison, e.g., considering the k^{th} comparison involving

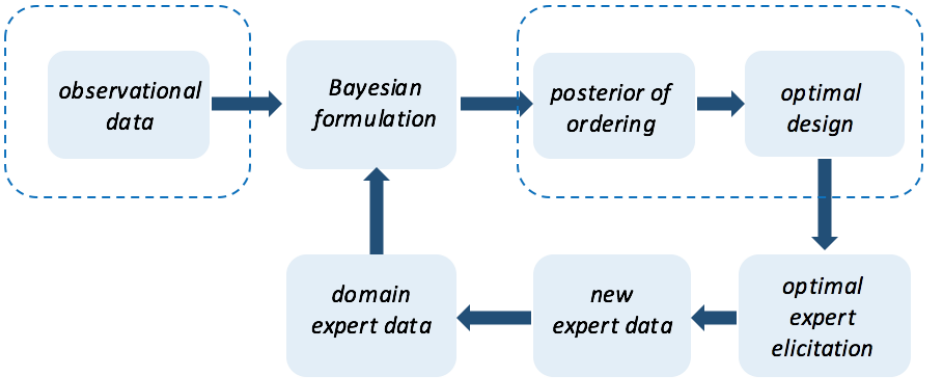


Figure 5.3: Flowchart of the Bayesian learning and sensing framework for optimal expert elicitation.

variables x_i and x_j, the expert may be asked whether or not x_i is upstream of x_j (i.e., denoted as $x_i > x_j$). The expert's response will be denoted as y_k, i.e., a positive y_k indicates that the expert knowledge more tends to support that variable x_i is upstream of variable x_j, while a negative y_k indicates the opposite. Note that the larger the y_k, stronger the knowledge. Thus, the expert data D^{ex} consists of the set of pairwise comparisons that have been queried (denoted as a set S) and the corresponding expert response data (denoted as a vector $\mathbf{y}$). Particularly, considering the sequential nature of our proposed method in data collection, we use subscript to distinguish data collected in different stages. For instance, following this line, we denote the initial observational data set as D_0^{obs}, and denote the initial expert comparison data as $D_0^{ex} = (S_0, \mathbf{y}_0)$. Then, the proposed Bayesian learning and sensing framework will learn the posterior distribution of the ordering of variables, building on which, an optimal expert knowledge elicitation plan could be derived to further collect new expert comparison data which will be denoted as $D_1^{ex} = (S_1, \mathbf{y}_1)$. This new data will help update the posterior distribution of ordering of variables, and data collection process will continue if needed. We will provide more detailed discussion of this sequential process and the stopping criteria.

In what follows, we introduce the details of how we could combine both the observational

data and expert comparison data to derive the posterior distribution of the ordering of variables. Instead of deriving the posterior distribution of the DAG structure of the BN, here we focus on the learning of the ordering of variables because this is the essential task for BN structure learning, and it has been found that learning the ordering can significantly reduce the computational complexity since the search space for ordering of variables is much smaller than the search space for DAGs [48, 151].

Developing the probabilistic formulation for expert data

It is reasonable to consider that the expert knowledge is always with uncertainty, and different experts may have different accuracy levels. Thus, a probabilistic model is needed to characterize not only the correspondence between the underlying ordering of variables with the expert comparison data, but also the expert's accuracy level. To develop this model, first, we invent a numerical vector to be a surrogate of the ordering of variables. This numerical vector, denoted as $\boldsymbol{\phi} \in R^p$, encodes the same ordering information between variables, since an upstream variable will have larger value in $\boldsymbol{\phi}$ than its downstream variables. Then, we could establish a probabilistic relationship between $\boldsymbol{\phi}$ and the observed D^{ex}, i.e., for the k^{th} comparison that involves variables x_i and x_j, we could assume that $y_k \sim N(\phi_i - \phi_j, \sigma^2/w_k)$. This essentially assumes that if the variable x_i is upstream (or downstream) of the variable x_j, we will expect to see positive (or negative) values of y_k. This is consistent with the nature of the expert comparison data we are adopting in this study. Note that, σ^2 encodes the overall accuracy level of the expert knowledge, as more knowledgeable expert will tend to have smaller σ^2. Also, w_k encodes uncertainty in this particular comparison, acting as the local accuracy level of the expert knowledge. In practice, experts could also provide their confidence level, i.e., w_k, along with y_k. Alternatively, when this information is lacking, we could simply assume $w_k = 1$ for all the comparison data. Following this line, we could further illustrate how we could represent the expert comparison data in a more compact matrix form. First, we invent a Boolean matrix, denoted as $\boldsymbol{B}$, where $\boldsymbol{B}_{k,j}$ is defined as:

$$\boldsymbol{B}_{k,j} = \begin{cases} 1 & \text{if } j = head(k) \\ -1 & \text{if } j = tail(k) \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

Here, $j = tail(k)$ if the k^{th} pairwise comparison is asked in the form as "$x_i > x_j$"; otherwise, $j = head(k)$ if "$x_j > x_i$". Then, it can be shown that

$$\mathbf{y} \sim N(\boldsymbol{B}\boldsymbol{\phi}, \sigma^2 \boldsymbol{W}^{-1}) \tag{5.2}$$

where $\boldsymbol{W}$ is the diagonal matrix of $\boldsymbol{w}$. Thus, for the initial expert comparison data, we could derive that $\boldsymbol{y}_0 \sim N(\boldsymbol{B}_0\boldsymbol{\phi}, \sigma^2 \boldsymbol{W}_0^{-1})$ where $\boldsymbol{B}_0$ is defined on the set $\boldsymbol{S}_0$.

Characterization of the prior distribution of $\boldsymbol{\phi}$ based on observational data

In what follows, we propose our method to derive the prior distribution of $\boldsymbol{\phi}$ by exploiting the information encoded in D^{obs}. It is an analytically intractable task, so we propose a computational procedure that is inspired by existing literature of BN structural learning. Some authors such as [47] suggested to use data perturbation methods such as Bootstrap in [47] to repeatedly bootstrap D_0^{obs} and apply an appropriate BN structure learning method on the perturbed dataset to learn the optimal DAG structure or the ordering of variables. It has been found in [47] that such a bootstrap procedure is especially robust. Thus, we propose the computational procedure that is depicted in Algorithm S2 as a sampling procedure of the orderings of variables based on D^{obs}:

After we draw the samples of ordering of variables $\{\hat{\boldsymbol{\phi}}_0^i, i = 1, 2, \ldots, m\}$ from observational data, the next step is to translate this knowledge and create the prior distribution of $\boldsymbol{\phi}$. Assuming that the prior distribution takes the form as $\boldsymbol{\phi} \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Lambda}_0^{-1})$. Then, the $\{\hat{\boldsymbol{\phi}}_0^i\}$ could be treated as random samples from the prior distribution and can be readily used to estimate the unknown parameters $\boldsymbol{\mu_0}$ and $\boldsymbol{\Lambda}_0^{-1}$ by maximum likelihood estimation.

The Bayesian learning framework

As a summary, based on the initial expert data D_0^{ex}, it has been known that we could derive that $\mathbf{y_0} \sim N(\boldsymbol{B}_0\boldsymbol{\phi}, \sigma^2\boldsymbol{W}_0^{-1})$ based on (5.2). And the prior distribution can be obtained from observational data as $\boldsymbol{\phi} \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Lambda}_0^{-1})$. Here, for the ease of derivation, in what follows we rewrite the prior as $\boldsymbol{\phi} \sim N(\boldsymbol{\mu_0}, \sigma^2\boldsymbol{\Lambda}_0^{-1})$, which is numerically equivalent if we change the scale of $\boldsymbol{\Lambda}_0$). Then, we could derive the posterior distribution of $\boldsymbol{\phi}$ with posterior mean $\boldsymbol{\mu_1}$ and posterior variance $\boldsymbol{\Lambda}_1^{-1}$, by learning from the Bayesian linear model literature such as [52]. While what we will derive in the following is largely borrowed from the literature, we present critical details in the derivation for completeness of our development. Specifically, the posterior distribution of $\boldsymbol{\phi}$ can be derived as:

$$
\begin{aligned}
& p(\boldsymbol{\phi}, \sigma^2 \mid \mathbf{y_0}, \boldsymbol{B}_0) \\
& \quad \propto p(\mathbf{y_0} \mid \boldsymbol{B}_0, \boldsymbol{\phi}, \sigma^2) p(\boldsymbol{\phi} \mid \sigma^2) p(\sigma^2) \\
& \quad \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y_0} - \boldsymbol{B}_0\boldsymbol{\phi})^T \boldsymbol{W}_0 (\mathbf{y_0} - \boldsymbol{B}_0\boldsymbol{\phi})\right) \\
& \qquad \times (\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\phi} - \boldsymbol{\mu_0})^T \boldsymbol{\Lambda}_0 (\boldsymbol{\phi} - \boldsymbol{\mu_0})\right) \\
& \qquad \times (\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2}\right)
\end{aligned}
\tag{5.3}
$$

Note that, here, we follow the Bayesian linear model literature ([52]) and use the inverse-Gamma distribution as the prior distribution for σ^2, where a_0 and b_0 are two parameters that can be specified by prior knowledge. We follow the suggestions made in ([52]) to specify a_0 and b_0 based on the non-informative prior principle, since it could provide robust performance for many applications. Also, our main interest is on the learning of the ordering of variables. The exponential parts in equation (5.3) could be combined as

$$
\begin{aligned}
& (\mathbf{y_0} - \boldsymbol{B}_0\boldsymbol{\phi})^T \boldsymbol{W}_0 (\mathbf{y_0} - \boldsymbol{B}_0\boldsymbol{\phi}) + (\boldsymbol{\phi} - \boldsymbol{\mu_0})^T \boldsymbol{\Lambda}_0 (\boldsymbol{\phi} - \boldsymbol{\mu_0}) \\
= & (\boldsymbol{\phi} - \boldsymbol{\mu_1})^T \left(\boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0\right) (\boldsymbol{\phi} - \boldsymbol{\mu_1}) + \mathbf{y_0}^T \mathbf{y_0} \\
& - \boldsymbol{\mu_1}^T (\boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0) \boldsymbol{\mu_1} + \boldsymbol{\mu_0}^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu_0}
\end{aligned}
\tag{5.4}
$$

where $\boldsymbol{\mu_1} = (\boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0 + \boldsymbol{\Lambda}_0)^{-1}(\boldsymbol{B}_0^T\boldsymbol{W}_0\mathbf{y_0} + \boldsymbol{\Lambda}_0\boldsymbol{\mu_0})$. Then, we could see that, the joint posterior distribution of $\boldsymbol{\phi}$ and σ^2 is actually a product of a normal distribution and an inverse-gamma distribution,

$$\begin{aligned} & p(\boldsymbol{\phi}, \sigma^2 \mid \mathbf{y_0}, \boldsymbol{B}_0) \\ & \quad \propto (\sigma^2)^{-k/2} \exp(z_1) \times (\sigma^2)^{\frac{-n-2a_0+2}{2}} \exp(z_2) \\ & \quad \propto p(\boldsymbol{\phi} \mid \sigma^2, \mathbf{y_0}, \boldsymbol{B}_0) p(\sigma^2 \mid \mathbf{y_0}, \boldsymbol{B}_0) \end{aligned}$$

where $z_1 = -\frac{1}{2\sigma^2}(\boldsymbol{\phi}-\boldsymbol{\mu_1})^T(\boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0+\boldsymbol{\Lambda}_0)(\boldsymbol{\phi}-\boldsymbol{\mu_1})$, and $z_2 = -\frac{2\boldsymbol{B}_0+\mathbf{y_0}^T\mathbf{y_0}-\boldsymbol{\mu_1}^T(\boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0+\boldsymbol{\Lambda}_0)\boldsymbol{\mu_1}+\boldsymbol{\mu_0}^T\boldsymbol{\Lambda}_0\boldsymbol{\mu_0}}{2\sigma^2}$. Essentially this suggests that the posterior distribution of $\boldsymbol{\phi}$ and σ^2 are $N(\boldsymbol{\mu_1}, \sigma^2\boldsymbol{\Lambda}_1^{-1})$ and Inv-Gamma, respectively. In summary, by combining the initial observational data D_0^{obs} and D_0^{ex}, we could derive the posterior mean of $\boldsymbol{\phi}$ as $\boldsymbol{\mu_1} = (\boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0+\boldsymbol{\Lambda}_0)^{-1}(\boldsymbol{B}_0^T\boldsymbol{W}_0\mathbf{y_0}+\boldsymbol{\Lambda}_0\boldsymbol{\mu_0})$, and the posterior variance as $\boldsymbol{\Lambda}_1^{-1} = (\boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0 + \boldsymbol{\Lambda}_0)^{-1}$.

5.3.2 A semidefinite programming (SDP) formulation for optimal expert comparison elicitation

In this section, we will develop an automated process for expert knowledge elicitation with a systematic optimization formulation. The central question to ask is, given the available data D_0^{obs} and D_0^{ex}, what is the optimal set of new expert comparison data we should further collect? As we have been able to derive the posterior distribution of $\boldsymbol{\phi}$ based on D_0^{obs} and D_0^{ex}. To see how this could be formulated, it is worthy of analyzing the structure of the posterior variance of $\boldsymbol{\phi}$, the $\boldsymbol{\Lambda}_1^{-1} = (\boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0 + \boldsymbol{\Lambda}_0)^{-1}$. Obviously, it can be derived that, given new expert comparison data $D_1^{ex} = (S_1, \mathbf{y_1})$, the posterior variance will be $\boldsymbol{\Lambda}_2^{-1} = (\boldsymbol{B}_1^T\boldsymbol{W}_1\boldsymbol{B}_1 + \boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0 + \boldsymbol{\Lambda}_0)^{-1}$, where $\boldsymbol{B}_1$ is defined on the set S_1. In order to more clearly identify the relationship between the candidate expert comparisons with $\boldsymbol{\Lambda}_2^{-1}$, we denote a matrix $\boldsymbol{B}^*$ that is defined on the set S^*. Obviously, S^* includes all the candidate comparisons that have not been included in S_0. Then, we could further rewrite $\boldsymbol{B}_1^T\boldsymbol{W}_1\boldsymbol{B}_1$ as $\boldsymbol{B}_1^T\boldsymbol{W}_1\boldsymbol{B}_1 = \sum_{k=1}^{|\boldsymbol{B}^*|} v_k\boldsymbol{a}_k^T\boldsymbol{a}_k$, where $\boldsymbol{a}_k$ is the k-th row of $\boldsymbol{B}^*$ and $\boldsymbol{v}$ is a Boolean vector $\in \{0,1\}^{|\boldsymbol{B}^*|}$, while $v_k = 1$ if the k^{th} comparison is included in S_1. With this, we have

almost formulated the optimization problem for new expert data elicitation, i.e., the goal is to identify the optimal solution of the decision variables v that can maximize the decrease of the posterior variance $(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0)^{-1}$, under the constraint that only a certain number of new expert comparisons, i.e., denoted as $\xi \in [0, |\boldsymbol{B}^*|]$, can be elicited.

This resembles many of the optimal design problems that have been discussed in the context of linear models. While many existing optimal design methods assume general structure for the design matrix and thus are limited by optimization options, we recognize that in our problem there is a special structure that can be exploited, which will lead to more powerful optimization formulations such as the Semi-definite Programming (SDP) formulation. On the other hand, while a few optimality criterion have been developed in optimal design, here, we propose to study the E-optimal design criteria first due to its robust nature and the subsequent computational benefit on optimization. Particularly, the E-optimal design criteria proposes to identify the subset of "design points" (in our case, the candidate comparisons) that can maximize the smallest nonzero eigenvalue of the information matrix, i.e., the information matrix is the inverse of the variance matrix, which is $(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0)$ in our case. This leads to the following optimization framework:

$$\begin{aligned} \max_{x} \quad & \boldsymbol{\lambda}_1(\boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0 + \sum_{l=1}^{|\boldsymbol{B}^*|} v_l \boldsymbol{a}_l^T \boldsymbol{a}_l) \\ \text{subject to} \quad & \mathbf{1}^T \boldsymbol{v} \leq \xi \\ & \boldsymbol{v} \in \{0,1\}^{|\boldsymbol{B}^*|} \end{aligned}$$

where $\boldsymbol{\lambda}_1(\boldsymbol{A})$ denotes the smallest nonzero eigenvalue of matrix $\boldsymbol{A}$. Since this problem is difficult to solve exactly, we propose to replace the Boolean constraint $\boldsymbol{v} \in \{0,1\}^{|\boldsymbol{B}^*|}$ by a relaxation, i.e., $\boldsymbol{v} \in [0,1]^{|\boldsymbol{B}^*|}$. Also the constraint $\mathbf{1}^T \boldsymbol{v} \leq \xi$ is binding in Formula 5.5 as $\xi \leq |\boldsymbol{B}^*|$, since $\boldsymbol{v} \leq \mathbf{1}$, the optimal value given by $\xi > |\boldsymbol{B}^*|$ is the same as that of $\xi = |\boldsymbol{B}^*|$. Therefore we have the following relaxation form:

$$\begin{aligned} \max_{x} \quad & \boldsymbol{\lambda}_1(\boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0 + \sum_{l=1}^{|\boldsymbol{B}^*|} v_l \boldsymbol{a}_l^T \boldsymbol{a}_l) \\ \text{subject to} \quad & \mathbf{1}^T \boldsymbol{v} = \xi \\ & \mathbf{0} \leq \boldsymbol{v} \leq \mathbf{1} \end{aligned} \tag{5.5}$$

Obviously, the optimal value from this relaxation form is an upper bound on the optimal value of the original form. Such a relaxation is a convex optimization problem since the constraints are linear functions of $\boldsymbol{v}$, and we only need to show that the objective function is a concave function of $\boldsymbol{v}$, which is shown in Lemma 1.

Lemma 1 *The optimization in* (5.5) *is a convex optimization problem.*

We could further show that the convex relaxation above leads to a semidefinite programming (SDP) problem. To see that, note that the objective in optimization in (5.5) can be equivalently stated as maximizing a new variable s where it is required that $s \leq \boldsymbol{\lambda}_1(\boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0 + \sum_{l=1}^{|\boldsymbol{B}^*|} v_l \boldsymbol{a}_l^T \boldsymbol{a}_l)$. This could be further restated as $\boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0 + \sum_{l=1}^{|\boldsymbol{B}^*|} v_l \boldsymbol{a}_l^T \boldsymbol{a}_l - s\boldsymbol{I}$ should be positive semi-definite. This leads to the following SDP formulation:

$$\begin{aligned} \max \quad & s \\ \text{subject to} \quad & s\boldsymbol{I} \preceq (\boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \boldsymbol{\Lambda}_0 + \textstyle\sum_{l=1}^{|\boldsymbol{B}^*|} v_l \boldsymbol{a}_l^T \boldsymbol{a}_l) \\ & \mathbf{1}^T \boldsymbol{v} = \xi \\ & \mathbf{0} \leq \boldsymbol{v} \leq \mathbf{1} \end{aligned} \tag{5.6}$$

This SDP can be solved using any standard SDP solver. In our experiments, we use the package "cvx" ([57]) to solve our problem which is shown to be fairly effective.

We could further show a useful property of the proposed formulation that indicates there is diminishing marginal effectiveness of the number of expert comparisons ξ on the objective

function. This property is shown in Figure S1 that is according to our empirical study of the proposed SDP formulation.

From this Figure S1, it is clear that, due to the concave shape of the objective function on ξ, the objective function has a much greater increase at the beginning than later on. This indicates that the first few expert comparisons could lead to a much greater marginal effect in reduction of posterior variance. And since this effect will gradually diminish, it indicates that probably only a few expert comparisons are needed for accurate learning of ordering of variables. While this is an empirical observation, to show that, denote $f(\boldsymbol{v}) = -\boldsymbol{\lambda}_1(L(\boldsymbol{v}))$ which has been known to be a convex function of v by Lemma 1. Denote $\omega(\xi) = \inf\{f(\boldsymbol{v})|\mathbf{1}^T\boldsymbol{v} \leq \xi, \mathbf{0} \leq \boldsymbol{v} \leq \mathbf{1}\}$, which can also be shown to be convex in Lemma 2 (by learning the proof from page 216 in [102]).

Lemma 2 *The function $\omega(\xi)$ is convex.*

On the other hand, it is easy to show that $-\omega(\xi)$ is monotonically increasing with the value of ξ, and is bounded on the domain of ξ. Since it is not likely that $-\omega(\xi)$ is linear due to its complicated functional form, only a function that takes the shape as shown in Figure S1 with diminishing marginal effectiveness can satisfy all these properties simultaneously.

5.3.3 Extension to Applications without Observational Data

Note that our proposed method can be easily extended to applications where observational data is not available. One approach is to assume non-informative prior distribution for $\boldsymbol{\phi}$, i.e., by assuming that $\boldsymbol{\Lambda}_0^{-1} = \sigma_0^2\boldsymbol{I}$ where σ_0^2 is a very large number and $\boldsymbol{I}$ is the identify matrix. Then, we could still apply our method to these applications based on the Bayesian learning framework and the optimal expert elicitation formulation. That said, there is still a particular problem that we need to address. Note that the posterior variance of the ordering of variables is $(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k\boldsymbol{a}_k^T\boldsymbol{a}_k + \boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0 + \sigma_0^{-2}\boldsymbol{I})^{-1}$. It can actually be shown in Lemma 3 that, the smallest eigenvalue of the corresponding information matrix, $(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k\boldsymbol{a}_k^T\boldsymbol{a}_k + \boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0 + \sigma_0^{-2}\boldsymbol{I})$, is σ_0^{-2}.

Lemma 3 *The smallest eigenvalue of the matrix,* $(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \sigma_0^{-2} \boldsymbol{I})$*, is* σ_0^{-2}*, and the corresponding eigenvector is* $\mathbf{1}$*.*

As the Lemma 3 indicates, we could not maximize the smallest eigenvalue of the information matrix. Thus, we propose to maximize the second smallest eigenvalue of $(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \sigma_0^{-2} \boldsymbol{I})$. Particularly, this will result in the following SDP formulation for the expert knowledge elicitation process

$$
\begin{aligned}
\max \quad & s \\
\text{s.t.} \quad & s(\boldsymbol{I} - \mathbf{1}\mathbf{1}^T/n) \preceq (\boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \sum_{l=1}^{|\boldsymbol{B}^*|} v_l \boldsymbol{a}_l^T \boldsymbol{a}_l + \sigma_0^{-2} \boldsymbol{I}) \\
& \mathbf{1}^T \boldsymbol{v} = \xi \\
& \mathbf{0} \leq \boldsymbol{v} \leq \mathbf{1}
\end{aligned}
$$

5.3.4 A validation procedure for the utility of the expert data

Theoretically, the expert comparison data is useful to help the learning of influential relationships between variables. However, in practice, it is reasonable to have the concern that whether or not the expert comparison data could be trusted, since expert data is subjective and varies among experts. It is worthy of highlighting that our proposed method actually takes into consideration of this concern. We use a probabilistic framework to characterize the relationship between the expert data with the underlying ordering of variables, and the parameters, σ^2 and $\boldsymbol{W}$, actually evaluate how accurate the expert data could be. For extreme cases where the expert is providing random guess information, our Bayesian learning framework will be able to estimate a large value of σ^2, and therefore, automatically assign more weight on the observational data for the final learning result of the BN. Besides the use of the parameters σ^2 and $\boldsymbol{W}$ to evaluate the utility of expert data, here, we also propose another approach, as a pseudo hypothesis testing procedure as a validation procedure. The basic rationale is that, if the expert data could be trusted, then it should be significantly

different from the random guess. Thus, our pseudo hypothesis testing procedure consists of three steps: 1) In the first step, a certain number of DAG structures will be randomly generated, and each of the DAG structures will be used to obtain a likelihood value on the observational data; 2) In the second step, we will sample the same number of DAG structures from the distribution of the ordering of the variables, that is learned solely from expert data (this can be done in our proposed Bayesian framework as a special case where no observational data is used), and again, each of these DAG structures will be used to obtain a likelihood value on the observational data; 3) Then, in the final step, we could compare the two distributions of the likelihood values and see if they are significantly different, i.e., by the use of a t-test. It is expected to see that there is significant difference between the expert knowledge with random guess which could justify the use of expert knowledge. We will demonstrate the use of this validation procedure in our numerical studies.

5.4 Experiments on simulated data

5.4.1 Methodology

In this section, we conduct experiments to evaluate the performance of the proposed method using a range of benchmark BN models. Specifically, we select 6 benchmark networks from the Bayesian Network Repository (BNR). These BNs have been widely used in the BN literature for performance evaluation since it provides a high quality representation of the diverse BN structures that we may encounter in real-world applications. As shown in Table S3, this cohort of BNs has the network sizes ranging from small to moderately large.

We compare our proposed method with the random sampling method that elicits expert knowledge on a random basis. To implement the random sampling method, we could follow the framework as shown in Figure 5.3 and just replace the use of the SDP method with the random sampling method.

In each simulation study, first, we select a BN network from Table S3, then, randomly generate parameters by following conventions that have been defined in the BN structure learning

literature such as [138, 66, 22]. As a BN model is essentially a joint distribution of the variables, we simulate observation data with a sample size 1000, which is our D_0^{obs}. We then apply a benchmark BN learning method on D_0^{obs} to learn the orderings of variables, i.e., we use the L1MB method, proposed in [138], that can be implemented in the Matlab DAGLearn Toolbox (`http://www.cs.ubc.ca/~murphyk/Software/DAGlearn/index.html`). Expert knowledge will be generated by the probabilistic model as mentioned in Section 3.1. Different accuracy levels of expert knowledge (i.e., as encoded in the parameter σ^2) will be used in our study.

We use two metrics for performance evaluation and comparison. The first metric is the correlation between the learned ordering of variables with the true ordering to evaluate the learning accuracy of both methods. The second metric is to compare the variance reduction of both methods. Specifically, we define the variance reduction ratio as follows:

$$\text{Variance Reduction Ratio} = \frac{Var(RS)_i - Var(SDP)_i}{Var(RS)_{i-1} - Var(RS)_i}$$

where $(\cdot)_i$ refers to results from the i-th iteration during the sequential learning process, and $var(SDP)$ and $var(RS)$ indicate the posterior variance of $\boldsymbol{\phi}$ obtained by the proposed method and random selection method, respectively. The variance reduction ratio essentially evaluates how much extra variance reduction the SDP method could provide on top of the random selection method. This metric facilitates the performance comparison since it is scale-invariant, and the value could be interpreted across different settings of other parameters. As a comparison, variance itself varies from case to case, which is hence not a good metric for performance evaluation and comparison.

5.4.2 Evaluation of Estimation Accuracy of the Ordering of Variables

In this section, we investigate the learning accuracy of our proposed Bayesian learning and sensing framework. The experimental results are presented in Figure 5.4, while all the 6 BN networks are investigated. A sequential procedure for expert knowledge elicitation is used, that has 10 iterations in total, while in each iteration, 4 comparison tasks are queried. Also, the accuracy level of the expert knowledge is set to be $\sigma^2 = 2$. Note that, experimental

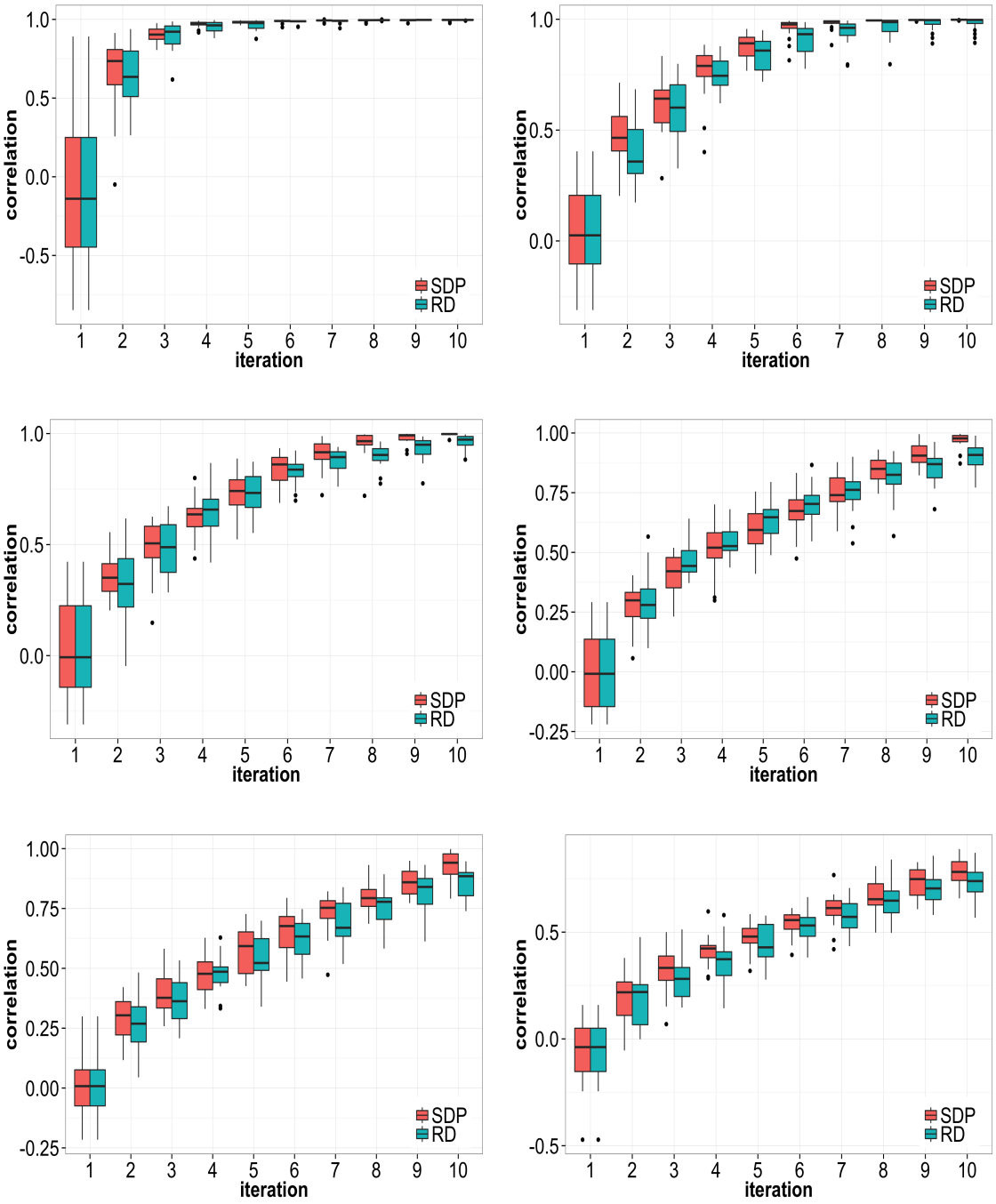


Figure 5.4: Evaluation of the proposed Bayesian learning and sensing framework, with either the SDP (red) method or the random selection method (blue), for estimating the ordering of the variables. Each figure corresponds to a network, which are asia, child, insurance, mildew, alarm, barley, from left to right and top to down.

results for other settings of these parameters lead to similar conclusion, so we only present the results as shown in Figure 5.4. Since for each simulated scenario, we repeat the experiments 100 times, so boxplot is used to present the overall results. The results clearly show that: 1) the proposed framework, either being integrated with the SDP method or the random selection method, could lead to effective learning of the underlying ordering of the variables. This indicates that the proposed Bayesian learning and sensing framework is effective in general. 2) The SDP method is better than the random selection method as it can lead to quicker detection of the underlying ordering of the variables. Also, the performance of the SDP method is more robust as it often generates results with tighter error bounds across the 6 BN networks.

5.4.3 Evaluation of Variance Reduction Performance

In this section, we further evaluate the efficiency of the proposed Bayesian learning and sensing framework in generating the candidate expert comparison tasks via the SDP formulation. Using the same setting of the parameters (such as $\sigma^2 = 2$ and the number of expert comparison tasks queried in each iteration is set to be 4), the experimental results are shown in Figure S6. Similar conclusions can be drawn as: 1) the proposed framework, either being integrated with the SDP method or the random selection method, could lead to effective reduction of the estimation variance. 2) The SDP method is better than the random selection method as it can lead to quicker reduction in the estimation variance of the orderings. For small network, we could observe that the estimation variance of the ordering by the proposed method quickly converges to very low (~ 0), while the estimation variance of the ordering by the random selection method still stays at a significant nonzero level. For networks with larger sizes, such as the four networks in the middle (i.e. the "child","insurance","mildew" and "alarm"), the estimation variance of the ordering by the proposed method usually approaches zero after $6 \sim 10$ iterations, while the random selection method needs more iterations.

Dataset	Scenario under $\sigma^2 = 1$	**Variance Reduction Ratio** $\pm s.d.$		
		@3th Iteration	**@6th Iteration**	**@9th Iteration**
Alarm	$nQuery = 1$	0.14 ± 0.10	0.15 ± 0.10	0.16 ± 0.11
	$nQuery = 2$	0.13 ± 0.11	0.11 ± 0.19	0.08 ± 0.20
	$nQuery = 4$	0.07 ± 0.08	0.08 ± 0.09	0.23 ± 0.32
Asia	$nQuery = 1$	0.29 ± 0.47	0.41 ± 0.63	1.92 ± 3.35
	$nQuery = 2$	0.20 ± 0.19	0.23 ± 0.40	0.50 ± 1.70
	$nQuery = 4$	0.11 ± 0.13	0.01 ± 0.02	-0.08 ± 0.09
Child	$nQuery = 1$	0.12 ± 0.09	0.16 ± 0.12	0.31 ± 0.53
	$nQuery = 2$	0.11 ± 0.09	0.19 ± 0.19	0.51 ± 0.50
	$nQuery = 4$	0.06 ± 0.07	0.21 ± 0.12	0.99 ± 2.06
Insurance	$nQuery = 1$	0.14 ± 0.09	0.15 ± 0.09	0.23 ± 0.18
	$nQuery = 2$	0.13 ± 0.09	0.07 ± 0.09	0.21 ± 0.26
	$nQuery = 4$	0.11 ± 0.07	0.19 ± 0.10	0.43 ± 0.38
Mildew	$nQuery = 1$	0.11 ± 0.11	0.16 ± 0.11	0.24 ± 0.23
	$nQuery = 2$	0.10 ± 0.06	0.09 ± 0.10	0.10 ± 0.15
	$nQuery = 4$	0.08 ± 0.07	0.08 ± 0.12	0.39 ± 0.29
Barley	$nQuery = 1$	0.09 ± 0.10	0.12 ± 0.09	0.18 ± 0.19
	$nQuery = 2$	0.06 ± 0.06	0.03 ± 0.08	0.01 ± 0.11
	$nQuery = 4$	0.05 ± 0.07	0.02 ± 0.07	0.11 ± 0.16

Table 5.1: Summary of experimental results when $\sigma^2 = 1$.

5.5 Experiments on real world applications

In what follows, we present our experimental results on two real-world applications, one is to learn the influential relationships among some critical KPIs for better understanding of human resource management in some manufacturing companies, and another one is to model the cascade of hypermetabolism reduction events for better understanding and staging of the Alzheimer's Disease (AD).

5.5.1 Identification of Influential Relationships of Key Performance Indicators (KPI) for Human Resource Management

There has been considerate interests in both academia and industry to develop methods to analyze the KPIs measurements, in order to gain a mechanistic understanding of the KPIs. It is because that a mechanistic understanding that can identify which KPIs drive which KPIs will be of great value for facilitating decision-makings. Thus, to demonstrate the utility of our method for KPI data analysis, we use a database that was collected from 31 KPIs of Human Resource Management from 197 manufacturing companies. Those KPIs cover a range of critical quality dimensions of the human resource management practices, which include the Employee Trait such as technical and problem solving skills, the reward policy in the company, the salaried or hourly employee turnover rates, the supply lead time and stability of demand, to name a few. There has been some prior knowledge regarding the influential relationships among some of the KPIs as mentioned in [75, 7], but a systematical study that can effectively combine both the observational data with expert knowledge has not been done yet to the best of our knowledge. Thus, we implement our proposed Bayesian learning and sensing framework on this dataset and interact with our expert (one of our co-authors) who is knowledgeable on this dataset to obtain the inquired expert comparison data. Note that, here, we limit our total number of expert comparison as 20 due to the limited capacity of the expert knowledge.

We show our results in Figure 5.5. Particularly, in Figure 5.5, the three figures from top

to bottom show the uncertainty of ordering of the KPIs when only observational data is used, observation data and 10 expert comparisons are used, and observation data and 20 expert comparisons are used, respectively. Apparently, it can be observed that, with observational data alone, the uncertainty of ordering of the variables can only achieve limited accuracy on the ordering. On the other hand, the proposed Bayesian learning and sensing framework effectively elicit expert knowledge to reduce the estimation uncertainty, and with 20 expert comparisons, we could achieve fairly accurate estimation of the ordering of the KPIs.

The driver KPIs that appear in the upstream positions in the learned ordering of the KPIs, correspond to solid recruiting criteria such as technical skills and work attitudes, indexes of internal resource and coordination, index of reward systems, and leadership index. It is reasonable for these KPIs to be drivers of the identified downstream KPIs. For instance, selection of employees who have sufficient technical skills, effective internal coordination, and reward system that motivate employees, all could have influence on the downstream indexes such as employee turnover rate. Among the downstream KPIs, what is worth noting is the "multi-functional employee". It is a qualification index of employees and can further lead to stability of business. The rationale behind this KPI to measure stability of business is based on the "personal successors system" mentioned in [75], when employees increase their qualification, they are able to occupy two positions, while at the same time, organization has personal successors on all positions such that it can run in normal mode when some employees suddenly quit the jobs. Knowing that this KPI could be driven by some upstream KPIs, quality improvement strategies targeting this KPI could be defined by improving on the corresponding upstream KPIs to remove the root-causes as a more proactive and preventative strategy. It is also interesting to notice that, the KPIs that correspond to the "employee trait" and "hiring process" act as mediators that deliver the influence from the driver KPIs to downstream KPIs. This result seems also reasonable, as these KPIs relate to human capitals' growing potential, which will have impact on the growth (into multi-functional employees) as well as job turnover rate (due to either being overqualified or unqualified) ([7]).

In addition, the validation process of the utility of the expert data is also conducted. The

result is reported in Figure S8. It clearly shows that the expert data is significantly different from random guess, demonstrating its utility for helping the learning of the influential relationships between the KPIs.

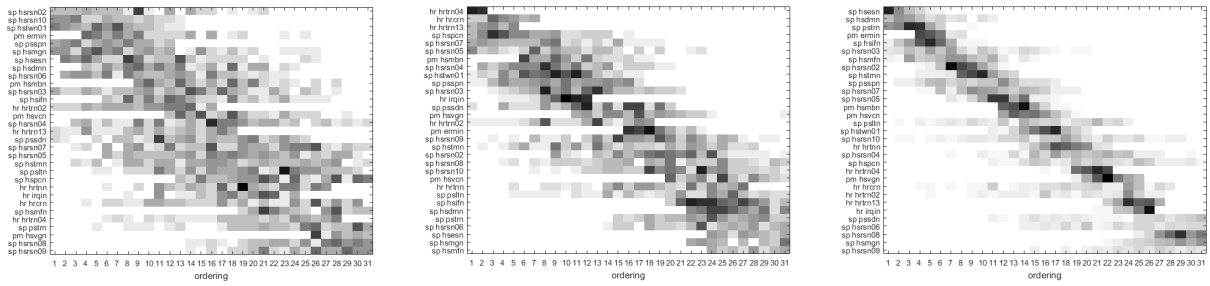


Figure 5.5: Uncertainty of ordering of the KPIs when only observational data is used (top), observation data and 10 expert comparisons are used (middle), observation data and 20 expert comparisons are used (bottom), respectively. Note that, the rows correspond to the KPIs while the numbers in the x-axis represent the ordering of the variables.

5.5.2 Identification of Cascade of Hypermetabolism Reduction Events for Alzheimer's Disease (AD)

In this section, we will implement our method to identify the cascade of hypermetabolism reduction events for Alzheimer's disease. Knowing the cascade of hypermetabolism reduction events will help us understand the progression stages of the disease, which is especially beneficial for early detection of the disease. Although disease progression model of AD has been developed in [77] that gives valuable hypothesis of the cascade of the abnormal clinical events along the AD progression process, such a model is still quite coarse and on the conceptual level, and is undergoing experimental validation. Thus, we propose to conduct quantitative analysis, and particularly focus on the FDG-PET imaging data that can measure the hypermetabolism reduction events, to enrich the disease progression model of AD and provide better resolution of the clinical events that happen along the progression

process.

Our FDG-PET imaging data includes a diverse set of subjects that are on different progression stages. There are 49 AD subjects, 116 mild cognitive impairment (MCI) subjects, and 67 normal aging (NC) subjects. We have identified 42 regions of interest (ROIs) that have been found related to AD. We perform the following data preparation steps: the original PET scan data set could be written as $D \in R^{232\times 42}$, corresponding to FDG-PET measurements of 42 regions of the 232 subjects. In order to derive the hypermetabolism reduction events based on FDG-PET data, we build a $\bar{X}$ chart for each ROI to characterize the normal FDG-PET level for each region. If the FDG-PET measurement of the subject in this ROI is normal, then we label this region of this subject as 0; otherwise, we label it as 1. This is just like the Phase-I analysis of control chart. Through this procedure, we could derive the "event dataset", denoted as $E \in I^{232\times 42}$. We then implement our method on $E \in I^{232\times 42}$ and combine it with the "expert" based on an existing cascade model of hypermetabolism reduction events of AD reported in the literature. e.g. [41]. The final results are shown in Figure S9, which can be interpreted in the same way as Figure **??**. Furthermore, the validation process of the utility of the expert data is also conducted and the result reported in Figure S9 shows that the expert data is significantly different from random guess, demonstrating its utility for helping the learning of ordering of the hypermetabolism reduction events.

Again, from Figure S9 we could observe that the uncertainty of ordering is large, when only observational data is used. With addition of expert knowledge, the uncertainty can be effectively reduced. Figure S10 provides a visualization of the mean ordering of the variables extracted from the posterior distribution. It clearly shows where the neurodegeneration strikes first along the progression of AD. Based on the severity of the degeneration conditions from early to late, we use different colors to highlight these events from yellow (NC stage) through orange (MCI stage) to red (AD stage).

From Figure S10, we can see that the following regions may be involved in early stages of the AD progression. Those regions include the frontal mid orbitalis cortex (node 4), which is an important area involved in the cognitive processing of decision-making; the

hippocampus cortex (node 20), located at the medial temporal lobe, is known to play key roles in the consolidation of information from short-term memory to long-term memory and spatial navigation; the middle temporal gyrus (node 16) serves in contemplating distance, recognition of known faces, and accessing word meaning while reading; the anterior cingulate cortex (ACC) (node 6) is involved in a wide range of cognitive functions such as rational cognitive functions, reward anticipation, decision-making, empathy, impulse control, and emotion; the inferior parietal lobule (node 8) is involved in the perception of emotions in facial stimuli and interpretation of sensory information; the Precuneus (node 9) is involved in episodic memory, visuospatial processing, reflections upon self, and aspects of consciousness. Apparently, many of these regions are related to the memory function, which is one of the early signs of AD. It is also exciting to see that the hippocampus region is involved in the AD progression in early stage, which is consistent with existing knowledge of AD in ([162, 117]) since the hippocampus has been a hallmark of AD. The fact that the Occipital Inferior region shows up at early stage is also interesting. It is known that the occipital inferior region belongs to the sensorimotor network (SMN), which relates to primary visual functions. A number of studies such as [1, 31, 131] have indicated that the functional changes in the visual system might precede the onset of AD, i.e., by impaired functional connectivity in visual systems. Our discovery raises an interesting hypothesis that decline in visual functions may be an early noninvasive biomarker for the diagnosis of AD. Similar findings were previously discovered in [161].

The regions involved in the intermediate progression stages reveal valuable insights regarding the neurological underpinning of the MCI stage. From Figure S10, we can see that these regions include the frontal mid lobe (node 2), which contains most of the dopamine sensitive neurons in the cerebral cortex. Since the dopamine system plays a critical role in memory, planning, and motivation, the reduction in hypermetabolism in this area could result in poorer performance and inefficient functioning during memory tasks. On the other hand, the occipital inferior lobe (node 13) is a critical visual related cortex. Impairment of this region can cause visual hallucinations or blindness. Its relatedness to MCI has been

pointed out in recent studies ([78, 159]). Such finding also provides evidence for the "sensory deprivation hypotheses" in [97], which proposes that sensory (including visual) underload may reduce opportunity for intellectual stimulating exchanges with the environment, and eventually reduce the general level of cognitive ability. In addition, the fusiform gyrus (node 19) is related to processing of color information, face, body, and word recognition. As many of these regions relate to the recognition ability, this is consistent with the scientific evidence from neuropsychological studies of AD that found that word recognition tasks are sensitive tests to identify the individuals who will develop AD soon. We also notice that, the regions such as the prefrontal lobe (node 1, 3,5), the parietal lobe(node 7, 10), the occipital lobe (node 11, 12), are involved in the late stages of AD progression. This is also consistent with the literature that has found these regions are usually less affected by AD in early stages ([56]).

5.6 Conclusion

In this paper, we propose a first of its kind method that can systematically elicit expert knowledge, optimally matched to observational data that has been collected, to identify the influential relationships between variables. This work is motivated by the success of the BN models in characterizing a wide range of complex systems, which use DAG structure to represent how variables influence each other. The BN models have also been found very useful for a number of quality control and monitoring tasks, as evidenced by the existing works in [95, 158, 99, 100, 94]. However, as the common approach for learning the DAG structure of a BN is merely using the observational data, it has been found that the theoretical bottleneck of this approach lies on the fact that only part of the influential relationships can be identified. On the other hand, in many applications, it is possible that knowledgeable expert could provide crucial information regarding the influential relationships among the variables. However, there has been a lack of systematical method that can automate the expert knowledge elicitation process, integrate it with existing learning methods from observational data, and further optimize it. Thus, we develop a Bayesian learning and sensing framework that can

combine both observational data and expert elicitation data in the form as a posterior distribution of orderings of the variables. Such a Bayesian learning framework will further lead to a systematical optimization formulation to automate the expert elicitation process. We conduct extensive numerical experiments on simulated data and real-world data to demonstrate the utility of our method and show its superior performance than baseline approaches.

There are many future directions we could exploit. For example, one important direction is to extend this framework to other optimal design methods such as $A-$Optimal or $D-$Optimal to fit needs from a broad range of application domains. New optimization models will be constructed accordingly. Also, note that the ordering of a DAG is not unique, resulting ambiguity in the representation of the ordering information. Fortunately, any ordering of the variables could facilitate the learning of the BN structure. This motivates us to pursue another direction which is to develop a better probabilistic model to characterize the distribution of the ordering of the variables. In this study, we propose to use a surrogate vector of the ordering of variables, which is essentially a relaxation of the ordering as a permutation set. As relaxation has been a common and effective tool for gaining numerical performance and computational feasibility, there is always a possibility that more statistical power could be gained if the relaxation could be tightened or not used at all. In addition, we may extend our method to be able to interact with multiple experts that have different accuracy levels of expert knowledge. Last but not least, it is worthy of pointing it out that the Bayesian learning framework demonstrated in the Figure 5.3 is generic. Thus, we could plug in any Bayesian network structure learning algorithm to learn the prior distribution of the orderings from observational data. Also, instead of Bootstrap, we could sample orderings of variables using Markov Chain Monte Carlo (MCMC) such as in [153, 46, 48]. In addition, there is an alternative approach to sample for DAGs rather than orderings of the variables in Algorithm 1. The output will be a collection of DAG structures then, denoted as $\{\hat{G}_0^i\}_{i=1,\cdots,m}$, that provide a good sample-based representation of the posterior distribution of the DAG structure. To integrate this DAG-sampling framework with the proposed Bayesian learning framework, we need a procedure to derive the prior distribution of ϕ. Specifically,

we recognize that there is a resemblance between the learned DAG $\hat{G}_0^i$ with the pairwise comparison data obtained from expert knowledge elicitation, i.e., a learned DAG $\hat{G}_0^i$ could be viewed as a collection of pairwise comparison data by defining that $w_k =$ count of directed edges (i, j) that shows up in all samples $\hat{G}_0^i$, and $y_k = \frac{1}{w_k} \sum_{(i,j)\in \hat{G}_0^1} sgn(i \to j)\cdot$, where k corresponds to the directed edge (i, j). Here $sgn(y)$ is used to indicate the direction of arcs. Following this line, we could derive the prior distribution of $\boldsymbol{\phi}$. Last but not least, it is also of interest to study how to choose hyperparameters in our model. Selecting hyperparameters for prior distribution in Bayesian models has been a practical challenge. In our case, the two parameters of the inverse-Gamma distribution, a_0, b_0, are selected by prior knowledge. One thing we notice is that the results are not sensitive to our choices of these two parameters, particularly when the sample size of the observational data is large and expert data is of high quality. All these directions are worthy of exploiting to further enhance and enrich the proposed methodology, that can combine observational data and expert knowledge for better learning of the influential relationships between variables.

Chapter 6

CONCLUSIONS AND FUTURE RESEARCH

This dissertation contributes to generic methodology development for analyzing some general complex datasets from high dimensional networked system that are ubiquitous in manufacturing and healthcare. It also contributes to domain knowledge discovery for Alzheimers disease research and quality control in manufacturing.

More specifically, table 6.1 lists the detailed problem settings and assumptions and the original contributions for individual work in each chapter.

6.1 Future research

There are still a lot of potential works in order to improve the diagnostic monitoring in the future.

6.1.1 Selective Sensing via Crowdsourcing

Problem Description

As the scope of machine learning applications has increased, the complexity of the prediction tasks (classification, regression) that are commonly tackled has grown dramatically. On one dimension, many classification problems involve hundreds or even thousands or even thousands of possible classes. On another dimension, researchers have spent considerable effort developing new features sets for particular applications. We want to take an active and adaptive approach to combine multiple classifiers/features at testing time, according to the defined value of classifiers and particular constraints.

Table 6.1: a summary of contributions for each chapter

Chapter	Background & Assumption	Major contribution
Chap 2	continuous system measurements; known system structure	a diagnostic monitoring method that conducts fault detection and diagnosis simultaneously; outstanding performance in both fault detection and diagnosis.
Chap 3	mixed type system measurements; unknown system structure; known sensor network structure	development of the safe screening framework that will greatly reduce the computational load of real-time fault diagnosis.
Chap 4	multimodal biomarkers; unknown system hierarchy	application of the mixed type Bayesian network to model the interactions among heterogeneous multimodal biomarkers.
Chap 5	unknown system structure; multivariate Gaussian distribution	systematically elicit expert knowledge to identify the influential relationships among variables

Related Work

To solve a complex prediction problem, many researchers have resorted to ensemble methods, in which mulitiple classifiers are combined to achieve an accurate prediction decision. For example, the Viola-Jones classifier uses a cascade of classifiers, each of which focuses on different spatial and appearance patterns. Boosting comducts a committee of weak classifiers, each of which focuses on different input distributions. Multiclass classification problems are very often reduced to a set of simpler decisions, including one-vs-one, one-vs-all, error-correcting output codes, or tree based approaches. Intuitively, different classifiers provide different expertise in making certain distinctions that can inform the classification task.

In real world, the online gathering of customer feedback is one form of crowdsourcing, as soliciting and displaying reviews helps buyers and sellers alike in their decision-making. The growth of services like Amazon's Mechanical Turk and CrowedFlower indicates that crowdsourcing can be used for increasingly diverse efforts in the large-scale gathering and integration of human judgments.

algorithm

Value and cost of classifier

We adopt the value of classifier in [**?**] as the information metric in equation 6.8. The value of classifier $V(h_i|m_O)$ for a classifier h_i given the observed classifier responses m_O is the combination of the expected reward of the state informed by h_i and the computational cost of h_i. Formally,

$$\begin{aligned} I(\cdot) =& V(h_i|m_O) \\ =& \int P(m_i|\boldsymbol{m}_O) R(P(Y|m_i, \boldsymbol{m}_O)) dm_i - \frac{1}{\tau} C(h_i|\mathcal{O}) \\ =& \underbrace{E_{m_i \sim P(m_i'|\boldsymbol{m}_O)}[R(P(Y|m_i, \boldsymbol{m}_O))]}_{\text{value: } \theta_i'} - \underbrace{\frac{1}{\tau} C(h_i|\mathcal{O})}_{\text{cost: } \sigma_i'} \end{aligned} \tag{6.1}$$

where the first part, θ'_i, is expected reward of $p(Y|m_i, \boldsymbol{m}_\mathcal{O})$, where the expectation is with respect to the posterior of M_i given $\boldsymbol{m}_\mathcal{O}$.

There are two ways to define the reward $R : p \to \mathcal{R}$, residual entropy and classification loss.

Residual Entropy

From the information-theoretical point of view, we want to reduce the uncertainty of the class variable Y by observing classifier responses. Therefore, a natural way to define the reward is to consider the negative residual entropy, that is the lower the entropy the higher the reward. Formally, given some posterior distribution $p(Y|\boldsymbol{m}_\mathcal{O})$, we define

$$R(p(Y|\boldsymbol{m}_\mathcal{O})) = -H(Y|\boldsymbol{m}_\mathcal{O}) = \sum_y p(y|\boldsymbol{m}_\mathcal{O}) \log p(y|\boldsymbol{m}_\mathcal{O}) \tag{6.2}$$

Classification Loss

From the classification loss point of view, we want to minimize the expected loss when choosing classifiers to evaluate. Therefore, given a loss function $\Delta(y, y')$ specifying the penalty of classifying an instance of class y to y', we can define the reward as negative of the minimum expected loss:

$$R(p(Y|\boldsymbol{m}_\mathcal{O})) = -\min_{y'} \sum_y p(y|\boldsymbol{m}_\mathcal{O}) \Delta(y, y') \tag{6.3}$$

Item Response Theory

In Rasch model, let θ'_i be the ability of classifier i, σ'_i be the difficulty of classifier i, $Y_i \in \{0, 1\}$ be a dichotomous variable that indicates the prediction is correct or incorrect, p_{i1} and p_{i0} be the probability that classifier i predicts correctly and incorrectly respectively. And let us define the prediction success rate be P_{i1}/P_{i0}, thus in Rasch model, the odds could be defined as:

$$\text{odds}_i = \frac{p_{i1}}{p_{i0}} = \frac{\theta'_i}{\sigma'_i} \tag{6.4}$$

After taking logs of both side on equation 6.4, we get:

$$\text{logit}_i = \theta_i - \sigma_i \tag{6.5}$$

where $\theta_i = \log(\theta_i')$ and $\sigma_i = \log(\sigma_i')$.

Based on equation 6.4 and equation 6.5, it could be able to deduce the probability of successful prediction of classifier i:

$$p_{i1} = \frac{\exp(\theta_i - \sigma_i)}{1 + \exp(\theta_i - \sigma_i)} \tag{6.6}$$

where this function is known as the one-parameter logistic function (1PL) in item response theory. It has the nice mathematical property that its values remain between 0 and 1 for any argument between $-\infty$ and $+\infty$, this makes it appropriate for predicting probabilities, where are always numbers between 0 and 1.

Thus the item information function of the 1PL model is defined as:

$$I_i = p_{i1} \cdot p_{i0} \tag{6.7}$$

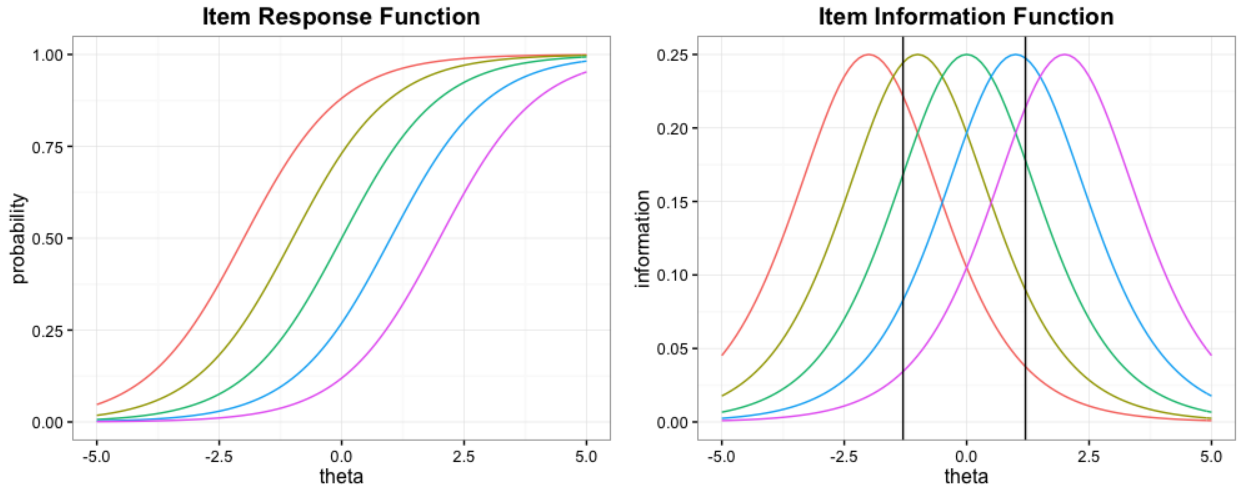


Figure 6.1: information curves with different difficulties

Figure 6.1 (right figure) shows the item characteristic curves of four items whose difficulties (-2,-1,1,2) are more or less evenly spread over the most important part of the ability range. The four curves run parallel to each other and never cross.

Shadow Tests

The following model is for the assembly of the shadow test when the current ability estimate is $\hat{\theta}_k$ and the $(k+1)$st classifier needs to be selected [**?**]:

$$\begin{aligned} \max_{x_i, i=1,\ldots,m} & \sum_{i=1}^{m} I_i(\hat{\theta}_k) x_i \\ \text{subject to } & \sum_{i \in V_c} x_i \geq n_c \\ & \sum_{i=1}^{m} x_i = n \\ & x_i = 1, \text{ for all } i \in S_k \\ & x_i \in \{0, 1\} \end{aligned} \tag{6.8}$$

The objective function in equation 6.8 requires the model to select the shadow test with optimal information at the current ability estimate, $\hat{\theta}_k$. The explanations of constraints are listed below:

1. For each category c, the total number of selected tests should be greater than a specified number n_c

2. For whole test, the total number of selected tests should be equal to a specified number n

3. For those tests that have been already picked up, the decision variables are set to 1

4. Binary optimization problem

Learning $p(m'_i|\boldsymbol{m}_\mathcal{O}, y)$

Given the subset of the training set $\{(x^j, y^j = y)\}_{j=1}^{N_y}$ corresponding to the instances from class y, we denote $m_i^j = h_i(x^j)$, then our goal is to learn $p(m'_i|\boldsymbol{m}_\mathcal{O}, y)$ from $\{(\boldsymbol{m}^j, y^j = y)\}_{j=1}^{N_y}$.

1. if $\mathcal{O} = \emptyset$, then $p(m'_i|\boldsymbol{m}_\mathcal{O}, y)$ reduces to the marginal distribution $p(m'_i|y) = \mathcal{N}(\mu_y, \sigma_y^2)$, and based on maximum likelihood estimation, we have:

$$\mu_y = \frac{1}{N_y}\sum_j m_i^j$$
$$\sigma_y^2 = \frac{1}{N_y}\sum_j (m_i^j - \mu_y)^2$$

2. if $\mathcal{O} \neq \emptyset$, we assume that $p(m'_i|\boldsymbol{m}_\mathcal{O}, y)$ is a linear Gaussian, i.e., $\mu_y = \omega_y^T \boldsymbol{m}_\mathcal{O}$:

$$\hat{\omega_y} = (\bar{M}_\mathcal{O}^T W \bar{M}_\mathcal{O} + \lambda I)^{-1} \bar{M}_\mathcal{O}^T W \bar{M}_i$$
$$\hat{\sigma_y} = \frac{1}{\sum_{j=1}^{N_y} w_j} \sum_{j=1}^{N_y} w_j \|m_i^j - \hat{\omega_y}^T \boldsymbol{m}_\mathcal{O}^j\|^2$$

where $w_j = e^{-\frac{\|\boldsymbol{m}_\mathcal{O} - \boldsymbol{m}_\mathcal{O}^j\|^2}{\beta}}$

Learning $p(m'_i|\boldsymbol{m}_\mathcal{O})$

$$p(m'_i|\boldsymbol{m}_\mathcal{O}) = \sum_y p(m'_i|\boldsymbol{m}_\mathcal{O}, y) p(y|\boldsymbol{m}_\mathcal{O})$$

where $p(y|\boldsymbol{m}_\mathcal{O})$ is the posterior over Y given some observation $\boldsymbol{m}_\mathcal{O}$ which is tracked over iterations.

Learning $p(Y|m_i, \boldsymbol{m}_\mathcal{O})$

$$p(Y|m_i, \boldsymbol{m}_\mathcal{O}) \propto p(m_i, \boldsymbol{m}_\mathcal{O}|Y)p(Y) = p(m_i|\boldsymbol{m}_\mathcal{O}, Y)p(\boldsymbol{m}_\mathcal{O}|Y)p(Y)$$

where all terms are available by caching previous computations.

Expected Results

We performed experiments on a collection of the UCI Wine-Quality data set: red wine (1599 samples) and white wine (4898 samples), each with 11 features. All tasks are multiclass classification problems with considering quality as the output variable. 70% of the samples are randomly chosen for training, and the remaining 30% for testing. To reduce statistical variability, results are averaged over 5 repetitions.

From the baselines, we have one-vs-one with max win, one-vs-all, DAGSVM and a tree-based method. These methods vary both in terms of what set of classifiers they use and how those classifiers are evaluated and combined.

We compare different methods in terms of both the classification accuracy and the number of evaluated classifiers. For our algorithm and the random selection baseline, we show the accuracy over iterations as well.

6.1.2 Optimal Sensor Allocation

Problem Description

Recent advances in technology have led to a continuously increasing in the complexity of systems. The failures within these systems can cause disruption to the operational functionality. Fault location has therefore become a first objective in engineering applications. Effective diagnostic approaches, which bring the system back at the lowest cost, can decrease downtime and consequently, enhance the operational functionality. However, designing an effective diagnosis system does not start after the system design, but it has to be done during the system design. Indeed, the performance of a diagnostic system highly depends on the number and location of sensors.

Sensors placement problem is to establish the objective function which can be calculated easily based on some special criteria as well as constraint conditions of sensor resources, and use some algorithms to optimize the objective function to get effective sensors distribution. Of course, the objective function and constraints should be simple in the formulation and easy

in calculating, researches about sensors placement focus on such issues: fault modeling or prediction of cause-effect behavior of the system; Reasonable optimization criterion to meet design specifications; The efficient algorithms used for solving the optimization problem.

Algorithm

Compressive sensing is a novel sensing approach, as opposed to the traditional sampling method that follows the Shannon-Nyquist rate in signal processing. Under the compressive sensing paradigm, the signal can be reconstructed from a smaller number of samples than the Shannon-Nyquist rate.

Many observation selection objectives satisfy submodularity, an intuitive diminishing returns property adding a sensor to a small deployment helps more than adding it to a large deployment. Examples include mutual information for spatial prediction and placing sensors for outbreak detection.

BIBLIOGRAPHY

[1] MW. Albers, GC. Gilmore, and J. Kaye. At the interface of sensory and motor dysfunctions and Alzheimer's Disease. *Alzheimer's and dementia: the journal of the Alzheimer's Association*, 11(1):70–98, 2015.

[2] Omir Correia ALVES JUNIOR and Ricardo J RABELO. A kpi model for logistics partners' search and suggestion to create virtual organisations. *International journal of networking and virtual organisations*, 12(2):149–177, 2013.

[3] K. J. Åström and T. Hägglund. Revisiting the ziegler–nichols step response method for pid control. *Journal of process control*, 14(6):635–650, 2004.

[4] Richard G Baraniuk. Compressive sensing [lecture notes]. *IEEE signal processing magazine*, 24(4):118–121, 2007.

[5] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research*, 9:1823–1840, 2008.

[6] Kaveh Bastani, Zhenyu Kong, Wenzhen Huang, and Yingqing Zhou. Compressive sensing–based optimal sensor placement and fault diagnosis for multi-station assembly processes. *IIE Transactions*, 48(5):462–474, 2016.

[7] BE. Becker, MA. Huselid, and D. Ulrich. *The HR scorecard: linking people, strategy and performance*. Harvard Business Review Press, 2001.

[8] R. L. Berger, L. T. Li, S. C. Hicks, J. A. Davila, L. S. Kao, and M. K. Liang. Development and validation of a risk-stratification score for surgical site occurrence and surgical site infection after open ventral hernia repair. *Journal of the American College of Surgeons*, 217(6):974–982, 2013.

[9] CK Blair, AR Folsom, David S Knopman, MS Bray, TH Mosley, E Boerwinkle, Atherosclerosis Risk in Communities (ARIC) Study Investigators, et al. Apoe genotype and cognitive decline in a middle-aged cohort. *Neurology*, 64(2):268–276, 2005.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[11] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[12] K. Breuing, C. E. Butler, S. Ferzoco, M. Franz, C. S. Hultman, J. F. Kilbridge, M. Rosen, R. P. Silverman, D. Vargo, Ventral Hernia Working Group, et al. Incisional ventral hernias: review of the literature and recommendations regarding the grading and technique of repair. *Surgery*, 148(3):544–558, 2010.

[13] Wray Buntine. Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, UAI '91, pages 52 – 60, 1991.

[14] J. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[15] Jian Cai, Xiangdong Liu, Zhihui Xiao, and Jin Liu. Improving supply chain performance management: A systematic approach to analyzing iterative kpi accomplishment. *Decision support systems*, 46(2):512–521, 2009.

[16] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[17] AR. Cano, A. Masegosa and Moral S. A method for integrating expert knowledge when learning bayesian networks from data. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICSPART B: CYBERNETICS*, 41(5), 2011.

[18] Giovanna Capizzi and Guido Masarotto. A least angle regression control chart for multidimensional data. *Technometrics*, 53(3):285–296, 2011.

[19] W Matthew Carlyle, Douglas C Montgomery, and George C Runger. Optimization problems and methods in quality control and improvement. *Journal of Quality Technology*, 32(1):1, 2000.

[20] Kewei Chen, Napatkamon Ayutyanont, Jessica BS Langbaum, Adam S Fleisher, Cole Reschke, Wendy Lee, Xiaofen Liu, Gene E Alexander, Dan Bandy, Richard J Caselli, et al. Correlations between fdg pet glucose uptake-mri gray matter volume scores and apolipoprotein e $\varepsilon 4$ gene dose in cognitively normal adults: a cross-validation study using voxel-based multi-modal partial least squares. *Neuroimage*, 60(4):2316–2322, 2012.

[21] Jie Cheng, David A Bell, and Weiru Liu. Learning belief networks from data: An information theory based approach. In *Proceedings of the sixth international conference on Information and knowledge management*, pages 325–331. ACM, 1997.

[22] Jie Cheng, G Grainer, J Kelly, DA Bell, and W Lius. Learning bayesian networks from data: An information-theory based approach, 2001. *URL citeseer. ist. psu. edu/628344. html*.

[23] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498, 2002.

[24] David Maxwell Chickering, David Heckerman, and Christopher Meek. A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, UAI'97, pages 80–89, 1997.

[25] Smart Manufacturing Leadership Coalition. Smart manufacturing, manufacturing intelligence and demand-dynamic performance. 2011.

[26] Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[27] Sanjeeb Dash, Dmitry M Malioutov, and Kush R Varshney. Screening for learning classification rules via boolean compressed sensing. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3360–3364. IEEE, 2014.

[28] Jim Davis, T Edgar, Y Dimitratos, J Gipson, I Grossmann, P Hewitt, R Jackson, K Seavey, P Porter, R Reklaitis, et al. Smart process manufacturing: An operations and technology roadmap. *Smart process manufacturing engineering virtual organization steering committee, Los Angeles, CA, Tech. Rep*, 2009.

[29] Cassio P. De Campos, Zhi Zeng, and Qiang Ji. Structure learning of Bayesian networks using constraints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 113–120, 2009.

[30] Luis M. De Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7:2149–2187, December 2006.

[31] DP. Devanand, X. Liu, MH. Tabert, G. Pradhaban, K. Cuasay, and K. Bell. Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biological Psychiatry*, 64(10):871–879, 2008.

[32] MG Dik, C Jonker, HC Comijs, LM Bouter, JWR Twisk, GJ Van Kamp, and DJH Deeg. Memory complaints and apoe-ε4 accelerate cognitive decline in cognitively normal elderly. *Neurology*, 57(12):2217–2222, 2001.

[33] J. T. Dipiro, R. G. Martindale, A. Bakst, P. F. Vacani, P. Watson, and M. T. Miller. Infection in surgical patients: Effects on mortality, hospitalization, and postdischarge care. *Am J Heal Pharm*, 55(8):777–781, 1998.

[34] C Montgomory Douglas. Introduction to statistical quality control. *John Wiley & Sons*, 2005.

[35] A Drzezga, T Grimmer, G Henriksen, M Mühlau, R Perneczky, I Miederer, C Praus, C Sorg, A Wohlschläger, M Riemenschneider, et al. Effect of apoe genotype on amyloid plaque load and gray matter volume in alzheimer disease. *Neurology*, 72(17):1487–1494, 2009.

[36] Ricardo Dunia, S Joe Qin, Thomas F Edgar, and Thomas J McAvoy. Identification of faulty sensors using principal component analysis. *AIChE Journal*, 42(10):2797–2812, 1996.

[37] Frederick Eberhardt. Almost optimal intervention sets for causal discovery. *arXiv preprint arXiv:1206.3250*, 2012.

[38] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

[39] L El Ghaoui, V Viallon, and T Rabbani. Safe feature elimination in sparse supervised learning technical report no. Technical report, UCB/EECS-2010-126, EECS Department, University of California, Berkeley, 2010.

[40] M Julia Flores, Ann E Nicholson, Andrew Brunskill, Kevin B Korb, and Steven Mascaro. Incorporating expert knowledge when learning bayesian network structure: a medical case study. *Artificial intelligence in medicine*, 53(3):181–204, 2011.

[41] HM. Fonteijn, M. Modat, MJ. Clarkson, J. Barnes, M. Lehmann, NZ. Hobbs, RI. Scahill, SJ. Tabrizi, S. Ourselin, NC. Fox, and DC. Alexander. An event-based model for disease progression and its application in familial Alzheimer's disease and huntington's disease. *Neuroimage*, 60(3):1880–9, 2012.

[42] W. T. Freeman and J. B. Tenenbaum. Learning bilinear models for two-factor problems in vision. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 554–560. IEEE, 1997.

[43] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954, 2008.

[44] Nir Friedman and Moises Goldszmidt. Learning Bayesian networks with local structure. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI '96, pages 252–262, 1996.

[45] Nir Friedman, Moises Goldszmidt, et al. Discretizing continuous attributes while learning bayesian networks. In *ICML*, pages 157–165, 1996.

[46] Nir Friedman and Daphne Koller. Being Bayesian about network structure. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI '00, pages 201–210, 2000.

[47] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[48] Nir Friedman, Iftach Nachman, and Dana Peér. Learning bayesian network structure from massive datasets: the sparse candidate algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 206–215. Morgan Kaufmann Publishers Inc., 1999.

[49] R. P. Gaynes, D. H. Culver, T. C. Horan, J. R. Edwards, C. Richards, J. S. Tolson, et al. Surgical site infection (ssi) rates in the united states, 1992–1998: the national nosocomial infections surveillance system basic ssi risk index. *Clinical Infectious Diseases*, 33(Supplement 2):S69–S77, 2001.

[50] Dan Geiger and David Heckerman. Learning gaussian networks. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 235–243. Morgan Kaufmann Publishers Inc., 1994.

[51] A. Gelman and J. Hill. Opening windows to the black box. *Journal of Statistical Software*, 40, 2011.

[52] Andrew Gelman, Christian Robert, Nicolas Chopin, and Judith Rousseau. *Bayesian Data Analysis*. Chapman and Hall, 1995.

[53] A. Gibson, S. Tevis, and G. Kennedy. Readmission after delayed diagnosis of surgical site infection: a focus on prevention using the american college of surgeons national surgical quality improvement program. *Am J Surg. Elsevier Inc*, 207(6):832–839, 2014.

[54] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

[55] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[56] RL. Gould, B. Arroyo, R. Brown, A. Owen, E. Bullmore, and RJ. Howard. Brain mechanisms of successful compensation during learning in alzheimer disease. *Neurology*, 67(1), 2006.

[57] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[58] R. W. Haley, D. H. Culver, J. W. White, W. M. Morgan, T. G. Emori, van P. Munn, and T. M. Hooton. The efficacy of infection surveillance and control programs in preventing nosocomial infections in us hospitals. *Am J Epidemiol*, 121(2):182–205, 1985.

[59] Mohmmad Hanafy and Hoda ElMaraghy. Co-design of products and systems using a bayesian network. *Procedia CIRP*, 17:284–289, 2014.

[60] John Hardy. Has the amyloid cascade hypothesis for alzheimer's disease been proved? *Current Alzheimer Research*, 3(1):71–73, 2006.

[61] M. Haridas and M. A. Malangoni. Predictive factors for surgical site infection in general surgery. *Surgery*, 144(4):496–501, 2008.

[62] Douglas M Hawkins. Multivariate quality control based on regression-adiusted variables. *Technometrics*, 33(1):61–75, 1991.

[63] Douglas M Hawkins. Regression adjustment for variables in multivariate quality control. *Journal of Quality Technology*, 25(3):170–182, 1993.

[64] Douglas M Hawkins and Edgard M Maboudou-Tchao. Multivariate exponentially weighted moving covariance matrix. *Technometrics*, 50(2):155–166, 2008.

[65] Jianming He and Wesley W Chu. A social network-based recommender system (snrs). In *Data Mining for Social Network Data*, pages 47–74. Springer, 2010.

[66] David Heckerman. A tutorial on learning with Bayesian networks. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. 1999.

[67] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[68] V. P. Ho, S. L. Stein, K. Trencheva, P. S. Barie, J. W. Milsom, S. W. Lee, and T. Sonoda. Differing risk factors for incisional and organ/space surgical site infections following abdominal colorectal surgery. *Diseases of the Colon & Rectum*, 54(7):818–825, 2011.

[69] David M Holtzman, John C Morris, and Alison M Goate. Alzheimers disease: the challenge of the second century. *Science translational medicine*, 3(77):77sr1–77sr1, 2011.

[70] Kie Honjo, Sandra E Black, and Nicolaas PLG Verhoeff. Alzheimer's disease, cerebrovascular disease, and the β-amyloid cascade. *Canadian Journal of Neurological Sciences*, 39(6):712–728, 2012.

[71] T. C. Horan, R. P. Gaynes, W. J. Martone, W. R. Jarvis, and T. G. Emori. Cdc definitions of nosocomial surgical site infections, 1992: a modification of cdc definitions of surgical wound infections. *Infect Control Hosp Epidemiol*, 13(10):606–608, 1992.

[72] Qiang Huang and Jianjun Shi. Stream of variation modeling and analysis of serial-parallel multistage manufacturing systems. *Ann Arbor*, 1001:48109, 2004.

[73] Shuai Huang, Jing Li, Jieping Ye, Adam Fleisher, Kewei Chen, Teresa Wu, and Eric Reiman. Brain effective connectivity modeling for alzheimer's disease by sparse gaussian bayesian network. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 931–939. ACM, 2011.

[74] Shuai Huang, Jing Li, Jieping Ye, Adam Fleisher, Kewei Chen, Teresa Wu, Eric Reiman, Alzheimer's Disease Neuroimaging Initiative, et al. A sparse structure learning algorithm for gaussian bayesian network identification from high-dimensional data. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1328–1342, 2013.

[75] G. Iveta. Human resources key performance indicators. *Journal of Competitiveness*, 4(1):117 – 128, 2012.

[76] Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[77] CR. Jack, DS. Knopman, WJ. Jagust, LM. Shaw, PS. Aisen, MW. Weiner, RC. Petersen, and JQ. Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *Lancet Neurol.*, 9(1):119–28, 2010.

[78] V. Jelic, SE. Johansson, O. Almkvist, P. Shigeta, M. Julin, A. Nordberg, B. Winblad, and LO. Wahlund. Quantitative electroencephalography in mild cognitive impairment: longitudinal changes and possible prediction of Alzheimer's disease. *Neurobiology of Aging*, 21(4):533 – 540, 2000.

[79] H. Ji, C. Liu, Z. Shen, and Y. Xu. Robust video denoising using low rank matrix completion. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1791–1798. IEEE, 2010.

[80] Wei Jiang and Kwok-Leung Tsui. A theoretical framework and efficiency study of multivariate statistical process control charts. *IIE Transactions*, 40(7):650–663, 2008.

[81] Wei Jiang, Kaibo Wang, and Fugee Tsung. A variable-selection-based multivariate ewma chart for process monitoring and diagnosis. *Journal of Quality Technology*, 44(3):209, 2012.

[82] Frank C Kaminsky, James C Benneyan, Robert D Davis, and Richard J Burke. Statistical control charts based on a geometric distribution. *Journal of Quality Technology*, 24(2):63–69, 1992.

[83] A. E. Kanters, D. M. Krpata, J. A. Blatnik, Y. M. Novitsky, and M. J. Rosen. Modified hernia grading scale to stratify surgical site occurrence after open ventral hernia repairs. *Journal of the American College of Surgeons*, 215(6):787–793, 2012.

[84] Celeste M Karch and Alison M Goate. Alzheimers disease risk genes and mechanisms of disease pathogenesis. *Biological psychiatry*, 77(1):43–51, 2015.

[85] J. M. Keller, M. R. Gray, and J. A. Givens. A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4):580–585, 1985.

[86] K. Kim, B. Kim, and G. Yi. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, 5(1):1, 2004.

[87] B. R. Krieger, D. M. Davis, J. E. Sanchez, J. J.L. Mateka, V. N. Nfonsam, J. C. Frattini, and J. E. Marcet. The use of silver nylon in preventing surgical site infections following colon and rectal surgery. *Dis Colon Rectum*, 54(8):1014–1019, 2011.

[88] Dana C Krueger, Douglas C Montgomery, and Christina M Mastrangelo. Application of generalized linear models to predict semiconductor yield using defect metrology data. *IEEE Transactions on Semiconductor Manufacturing*, 24(1):44–58, 2011.

[89] Rodrigo O Kuljiš. Integrative understanding of emergent brain properties, quantum brain hypotheses, and connectome alterations in dementia are key challenges to conquer alzheimer's disease. *Frontiers in neurology*, 1, 2010.

[90] M. H. Kutner, C. J. Nachtsheim, and J. Neter. Applied linear regression models. *McGraw-Hill/Irwin, Boston, 4th Edition*, 2004.

[91] Helge Langseth and Thomas D Nielsen. Fusion of domain knowledge with data for structural learning in object oriented domains. *Journal of Machine Learning Research*, 4(Jul):339–368, 2003.

[92] Steffen L Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.

[93] E. H. Lawson, B. L. Hall, and C. Y. Ko. Risk factors for superficial vs deep/organ-space surgical site infections: implications for quality improvement initiatives. *JAMA surgery*, 148(9):849–858, 2013.

[94] Jing Li, Jionghua Jin, and Jianjun Shi. Causation-based t^ sup 2^ decomposition for multivariate process monitoring and diagnosis. *Journal of Quality Technology*, 40(1):46, 2008.

[95] Jing Li and Jianjun Shi. Knowledge discovery from observational data for process control using causal bayesian networks. *IIE transactions*, 39(6):681–690, 2007.

[96] Yi Li, Juha O Rinne, Lisa Mosconi, Elizabeth Pirraglia, Henry Rusinek, Susan DeSanti, Nina Kemppainen, Kjell Någren, Byeong-Chae Kim, Wai Tsui, et al. Regional analysis of fdg and pib-pet images in normal aging, mild cognitive impairment, and alzheimers disease. *European journal of nuclear medicine and molecular imaging*, 35(12):2169–2181, 2008.

[97] U. Lindenberger and PB. Baltes. Sensory functioning and intelligence in old age: a strong connection. *Psychology and Aging*, 9(3):339–55, 1994.

[98] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

[99] Kaibo Liu and Jianjun Shi. Objective-oriented optimal sensor allocation strategy for process monitoring and diagnosis by multivariate analysis in a bayesian network. *IIE Transactions*, 45(6):630–643, 2013.

[100] Kaibo Liu, Xi Zhang, and Jianjun Shi. Adaptive sensor allocation strategy for process monitoring and diagnosis in a bayesian network. *IEEE Transactions on Automation Science and Engineering*, 11(2):452–462, 2014.

[101] Val J Lowe, Bradley J Kemp, Clifford R Jack, Matthew Senjem, Stephen Weigand, Maria Shiung, Glenn Smith, David Knopman, Bradley Boeve, Brian Mullan, et al. Comparison of 18f-fdg and pib pet in cognitive impairment. *Journal of Nuclear Medicine*, 50(6):878–886, 2009.

[102] David. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, NY, USA, 1969.

[103] A M Tucker and Yaakov Stern. Cognitive reserve in aging. *Current Alzheimer Research*, 8(4):354–360, 2011.

[104] Rami Mahdi and Jason Mezey. Sub-local constraint-based learning of Bayesian networks using a joint dependence criterion. *Journal of Machine Learning Research*, 14(1):1563–1603, January 2013.

[105] N. N. Mahmoud, R. S. Turpin, G. Yang, and W. B. Saunders. The impact of surgical-site infections in the 1990s: attributable mortality, excess length of hospitalization, and extra costs. *Infect Control Hosp Epidemiol*, 20(11):725–730, 1999.

[106] N. N. Mahmoud, R. S. Turpin, G. Yang, and W. B. Saunders. Impact of surgical site infections on length of stay and costs in selected colorectal procedures. *Surg Infect (Larchmt)*, 10(6):539–544, 2009.

[107] Dmitry Malioutov and Kush Varshney. Exact rule learning via boolean compressed sensing. In *International Conference on Machine Learning*, pages 765–773, 2013.

[108] D. L. Malone, T. Genuit, J. K. Tracy, C. Gannon, and L. M. Napolitano. Surgical site infections: reanalysis of risk factors. *J Surg Res*, 103(1):89–95, 2002.

[109] Bruce G Marcot, J Douglas Steventon, Glenn D Sutherland, and Robert K McCann. Guidelines for developing and updating bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research*, 36(12):3063–3074, 2006.

[110] Robert L Mason, Nola D Tracy, and John C Young. Decomposition of t2 for multivariate control chart interpretation. *Journal of quality technology*, 27(2):99–1108, 1995.

[111] Robert L Mason and John C Young. Improving the sensitivity of the t2 statistic in multivariate process control. *Journal of Quality Technology*, 31(2):155, 1999.

[112] Alejandro Maté, Juan Trujillo, and John Mylopoulos. Conceptualizing and specifying key performance indicators in business strategy models. In *Proceedings of the 2012 conference of the center for advanced studies on collaborative research*, pages 102–115. IBM Corp., 2012.

[113] Stefano Monti and Gregory F Cooper. Learning hybrid bayesian networks from data. In *Learning in graphical models*, pages 521–540. Springer, 1998.

[114] Stefano Monti and Gregory F Cooper. A multivariate discretization method for learning bayesian networks from mixed data. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 404–413. Morgan Kaufmann Publishers Inc., 1998.

[115] John C Morris, Catherine M Roe, Chengjie Xiong, Anne M Fagan, Alison M Goate, David M Holtzman, and Mark A Mintun. Apoe predicts amyloid-beta but not tau alzheimer pathology in cognitively normal aging. *Annals of neurology*, 67(1):122–131, 2010.

[116] Lisa Mosconi. Glucose metabolism in normal aging and alzheimers disease: methodological and physiological considerations for pet studies. *Clinical and translational imaging*, 1(4):217–233, 2013.

[117] Y. Mu and FH. Gage. Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Molecular Neurodegeneration*, 6(85), 2011.

[118] Kevin P Murphy. Active learning of causal bayes net structure. 2001.

[119] Benedetta Nacmias, Valentina Berti, Irene Piaceri, and Sandro Sorbi. Fdg pet and the genetics of dementia. *Clinical and Translational Imaging*, 1(4):235–246, 2013.

[120] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[121] Y. Nesterov. Introductory lectures on convex optimization. *Springer Science & Business Media*, 87, 2004.

[122] Daniel Nikovski. Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):509–516, 2000.

[123] Xianghui Ning and Fugee Tsung. Monitoring a process with mixed-type and high-dimensional data. In *Industrial Engineering and Engineering Management (IEEM), 2010 IEEE International Conference on*, pages 1430–1432. IEEE, 2010.

[124] B. A. Olshausena, C. Cadieub, J. Culpepperc, and D. K. Warlandd. Bilinear models of natural images. 2007.

[125] HI Patel. Quality control methods for multivariate binomial and poisson distributions. *Technometrics*, 15(1):103–112, 1973.

[126] Judea Pearl. Causality: models, reasoning and inference. *Econometric Theory*, 19(675-685):46, 2003.

[127] N. Peek and A. Abu-Hanna. Clinical prognostic methods: trends and developments. *Journal of Biomedical informatics*, 48:1, 2014.

[128] E. N. Perencevich, K. E. Sands, S. E. Cosgrove, E. Guadagnoli, E. Meara, and R. Platt. Health and economic impact of surgical site infections diagnosed after hospital discharge. *Emerg Infect Dis.*, 9(2):196–203, 2003.

[129] T. D. Pinkney, M. Calvert, D. C. Bartlett, A. Gheorghe, V. Redman, G. Dowswell, W. Hawkins, T. Mak, H. Youssef, C. Richardson, et al. Impact of wound edge protection devices on surgical site infection after laparotomy: multicentre randomised controlled trial (rossini trial). *BMJ*, 347(7):f4305, 2013.

[130] Carmel A Pollino, Owen Woodberry, Ann Nicholson, Kevin Korb, and Barry T Hart. Parameterisation and evaluation of a bayesian network for use in an ecological risk assessment. *Environmental Modelling & Software*, 22(8):1140–1152, 2007.

[131] KL. Possin. Visual spatial cognition in neurodegenerative disease. *Neurocase*, 16(6):466–487, 2010.

[132] Francesco Rinaldi. Mathematical programming methods for minimizing the zero-norm over polyhedral sets. *Sapienza, University of Rome. url: http://www. math. unipd. it/˜ rinaldi/papers/thesis0. pdf*, 2009.

[133] Teemu. Roos, Tomi. Silander, Petri. Kontkanen, and Petri. Myllymaki. Bayesian network structure learning using factorized NML universal models. In *Information Theory and Applications Workshop*. IEEE, 2008.

[134] P. Sanger, A. Hartzier, R. Lordon, C. Armstrong, W. Lober, H. Evans, and W. Pratt. A patient-centered system in a provider-centered world: challenges of incorporating post-discharge wound data into practice. *Journal of the American Medical Informatics*, 0:1–13, 2016.

[135] P. C. Sanger, G. H. van Ramshorst, E. Mercan, S. Huang, A. Hartzler, C. A. L. Armstrong, R. J. Lordon, W. B. Lober, and H. L. Evans. A prognostic model of surgical site infection using daily clinical wound assessment. *Journal of the American College of Surgeons*, 2016.

[136] Richard Scheines, Clark Glymour, and Frederick Eberhardt. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. 2005.

[137] Richard Scheines, Peter Spirtes, and Clark Glymour. A qualitative approach to causal modeling. In *Qualitative simulation modeling and analysis*, pages 72–97. Springer, 1991.

[138] Mark Schmidt, Alexandru Niculescu-Mizil, and Kevin Murphy. Learning graphical model structure using l1-regularization paths. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, pages 1278–1283, 2007.

[139] Frank Schraml, Kewei Chen, Napatkamon Ayutyanont, Roontiva Auttawut, Jessica BS Langbaum, Wendy Lee, Xiaofen Liu, Dan Bandy, Stephanie Q Reeder, Gene E Alexander, et al. Association between an alzheimers disease-related index and apoe $\varepsilon 4$ gene dose. *PloS one*, 8(6):e67163, 2013.

[140] Elisabeth MC Schrijvers, Peter J Koudstaal, Albert Hofman, and Monique MB Breteler. Plasma clusterin and the risk of alzheimer disease. *Jama*, 305(13):1322–1326, 2011.

[141] Thomas Schulz, Łukasz Radliński, Thomas Gorges, and Wolfgang Rosenstiel. Defect cost flow model: a bayesian network for predicting defect correction effort. In *Proceedings of the 6th International Conference on Predictive Models in Software Engineering*, page 16. ACM, 2010.

[142] Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

[143] Frederick T Sheldon, Robert K Abercrombie, and Ali Mili. Evaluating security controls based on key performance indicators and stakeholder mission. In *Proceedings of the 4th annual workshop on Cyber security and information intelligence research: developing*

strategies to meet the cyber security and information intelligence challenges ahead, page 41. ACM, 2008.

[144] M. Signoretto, R. Van de Plas, B. De Moor, and J. AK Suykens. Tensor versus matrix completion: a comparison with application to spectral data. *Signal Processing Letters, IEEE*, 18(7):403–406, 2011.

[145] A. Smola and B. Schlkopf. A tutorial on support vector regression. *NeuroCOLT Technical Report NC-TR-98-030, Royal Holloway College, University of London, UK*, 1998.

[146] Yaakov Stern. What is cognitive reserve? theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*, 8(3):448–460, 2002.

[147] Yaakov Stern. Cognitive reserve and alzheimer disease. *Alzheimer Disease & Associated Disorders*, 20(2):112–117, 2006.

[148] Yaakov Stern. Cognitive reserve in ageing and alzheimer's disease. *The Lancet Neurology*, 11(11):1006–1012, 2012.

[149] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

[150] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.

[151] Marc Teyssier. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. In *Proceedings of the Fifty-seventh Annual Conference on Uncertainty in Artificial Intelligence*, UAI '05, pages 584–590, 2005.

[152] T. Toma, R. Bosman, A. Siebes, N. Peek, and A. Abu-Hanna. Learning predictive models that use pattern discovery - a bootstrap evaluative approach applied in organ functioning sequences. *Journal of biomedical informatics*, 43(4):578–586, 2010.

[153] Simon Tong and Daphne Koller. Active learning for structure in bayesian networks. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pages 863–869, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[154] G. H. van Ramshorst. Wound failure in laparotomy: New insights. *Erasmus University Rotterdam*, 2014.

[155] G. H. van Ramshorst, M. C. Vos, D. den Hartog, W. C. J. Hop, J. Jeekel, S. E. R. Hovius, and J. F. Lange. A comparative assessment of surgeons tracking methods for surgical site infections. *Surg Infect (Larchmt)*, 14(2):181–187, 2013.

[156] C. van Walraven and R. Musselman. The surgical site infection risk score (ssirs): a model to predict the risk of surgical site infections. *PLoS one*, 8(6):e67167, 2013.

[157] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, 1991.

[158] Sylvain Verron, Jing Li, and Teodor Tiplica. Fault detection and isolation of faults in a multivariate process with bayesian network. *Journal of Process Control*, 20(8):902–911, 2010.

[159] K. Vlcek and J. Laczo. Neural correlates of spatial navigation changes in mild cognitive impairment and alzheimers disease. *Frontiers in Behavioral Neuroscience.*, 8(89), 2014.

[160] Kaibo Wang and Wei Jiang. High-dimensional process monitoring and fault isolation via variable selection. *Journal of Quality Technology*, 41(3):247, 2009.

[161] Pan. Wang, Bo. Zhou, Hongxiang. Yao, Yafeng. Zhan, Zengqiang. Zhang, Yue. Cui, Kaibin. Xu, Jianhua. Ma, Luning. Wang, Ningyu. An, Xi. Zhang, Yong. Liu, and Tianzi. Jiang. Aberrant intra- and inter-network connectivity architectures in alzheimers disease and mild cognitive impairment. *Scientific Reports*, 5(14824), 2015.

[162] M.J West, P.D Coleman, D.G Flood, and J.C Troncoso. Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *The Lancet*, 344(8925):769 – 772, 1994.

[163] Jieping Ye, Kewei Chen, Teresa Wu, Jing Li, Zheng Zhao, Rinkal Patel, Min Bae, Ravi Janardan, Huan Liu, Gene Alexander, et al. Heterogeneous data fusion for alzheimer's disease study. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1025–1033. ACM, 2008.

[164] Shiming Ye, Yadong Huang, Karin Müllendorff, Liming Dong, Gretchen Giedt, Elaine C Meng, Fred E Cohen, Irwin D Kuntz, Karl H Weisgraber, and Robert W Mahley. Apolipoprotein (apo) e4 enhances amyloid β peptide production in cultured neuronal cells: Apoe structure as a potential therapeutic target. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18700–18705, 2005.

[165] Jie Yu and Mudassir M Rashid. A novel dynamic bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis. *AIChE Journal*, 59(7):2348–2365, 2013.

[166] Daoqiang Zhang, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage*, 59(2):895–907, 2012.

[167] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.

[168] Shiyu Zhou, Qiang Huang, and Jianjun Shi. State space modeling of dimensional variation propagation in multistage machining process using differential motion vectors. *IEEE Transactions on robotics and automation*, 19(2):296–309, 2003.

[169] E. Zimlichman, D. Henderson, O. Tamir, C. Franz, P. Song, C. K. Yamin, C. Keohane, C. R. Denham, and D. W. Bates. Health care-associated infections: A meta-analysis of costs and financial impact on the us health care system. *AMA Intern Med*, 173(22):2039–2046, 2013.

[170] Changliang Zou, Wei Jiang, and Fugee Tsung. A lasso-based diagnostic framework for multivariate statistical process control. *Technometrics*, 53(3):297–309, 2011.

[171] Changliang Zou and Peihua Qiu. Multivariate statistical process control using lasso. *Journal of the American Statistical Association*, 104(488):1586–1596, 2009.

[172] Min Zou and Suzanne D Conzen. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79, 2004.

Appendix A

APPENDIX FOR CHAPTER 3

A.1 Boolean Compressive Sensing

It has been known that the best relaxation for approximating the l_0 norm is to relax it using l_1 norm. It results in the following optimization problem:

$$\begin{aligned} \min \quad & \|\mathbf{x}\|_1 \\ \text{subject to} \quad & \mathbf{y} = (A \vee \mathbf{x}) \otimes \boldsymbol{\xi} \end{aligned}$$

To relax the boolean constraints, we notice that if a vector $\mathbf{x}$ satisfies the constraint that $\mathbf{y} = (A \vee \mathbf{x}) \otimes \boldsymbol{\xi}$, then it also satisfies the linear inequalities: $A_{\mathbb{T}}\mathbf{x} + \boldsymbol{\xi}_{\mathbb{T}} \geq \mathbf{1}$ and $A_{\mathbb{F}}\mathbf{x} = \boldsymbol{\xi}_{\mathbb{F}}$, where $\mathbb{T} = \{i | y_i = 1\}$ is the set of out of control sensor signals, $\mathbb{F} = \{i | y_i = 0\}$ is the set of in control sensor signals, and $A_{\mathbb{T}}$ and $A_{\mathbb{F}}$ are the corresponding subsets of rows of A. Putting the relaxations together we have the following formulation:

$$\begin{aligned} \min \quad & \sum_{i=1}^{p} x_i \\ \text{subject to} \quad & A_{\mathbb{T}}\mathbf{x} + \boldsymbol{\xi}_{\mathbb{T}} \geq \mathbf{1} \\ & A_{\mathbb{F}}\mathbf{x} = \boldsymbol{\xi}_{\mathbb{F}} \\ & 0 \leq x_i \leq 1, i = 1, \ldots, p \\ & 0 \leq \boldsymbol{\xi}_i \leq 1, i \in \mathbb{T} \\ & 0 \leq \boldsymbol{\xi}_i, i \in \mathbb{F} \end{aligned}$$

On top of this formulation, as suggested in boolean compressive sensing [**?**], we can further introduce a penalty term on $\boldsymbol{\xi}$ to encourage sparsity of $\mathbf{x}$, leading to the LP primal problem 3.2.

A.2 Duality of Formulation 3.2

The second term in the primal formulation 3.2, $\sum_{j=1}^{q} \boldsymbol{\xi}_j$, could be decomposed into test positive part and test negative part, while the test negative part could be rewritten after substitution of the constraints $A_{\mathbb{F}}\mathbf{x} = \boldsymbol{\xi}_{\mathbb{F}}$. This leads to

$$\begin{aligned}
\sum_{i=1}^{p} x_i + \lambda \sum_{j=1}^{q} \boldsymbol{\xi}_j &= \sum_{i=1}^{p} x_i + \lambda \sum_{j\in\mathbb{T}} \boldsymbol{\xi}_j + \lambda \|\boldsymbol{\xi}_{\mathbb{F}}\|_1 \\
&= \sum_{i=1}^{p} x_i + \lambda \sum_{j\in\mathbb{T}} \boldsymbol{\xi}_j + \lambda \|A_{\mathbb{F}}\mathbf{x}\|_1 \\
&= \sum_{i=1}^{p} x_i + \lambda \sum_{j\in\mathbb{T}} \boldsymbol{\xi}_j \\
&\quad + \lambda \|[\mathbf{a}_{\mathbb{F}}^1, \ldots, \mathbf{a}_{\mathbb{F}}^p] \cdot [x_1, \ldots, x_p]^T\|_1 \\
&= \sum_{i=1}^{p} x_i + \lambda \sum_{j\in\mathbb{T}} \boldsymbol{\xi}_j + \lambda \sum_{i=1}^{p} \|\mathbf{a}_{\mathbb{F}}^i\|_1 x_i \\
&= \sum_{i=1}^{p} \left(1 + \lambda \|\mathbf{a}_{\mathbb{F}}^i\|_1\right) x_i + \lambda \sum_{j\in\mathbb{T}} \boldsymbol{\xi}_j
\end{aligned}$$

Furthermore, since the upper bounds of x_i and $\boldsymbol{\xi}_i$ are redundant, we could rewrite the primal formulation as

$$\begin{aligned}
\mathcal{P}(\mathbf{x}, \boldsymbol{\xi}) = \min \quad & \sum_{i=1}^{p} \left(1 + \lambda \|\mathbf{a}_{\mathbb{F}}^i\|_1\right) x_i + \lambda \sum_{j\in\mathbb{T}} \boldsymbol{\xi}_j \\
\text{such that} \quad & A_{\mathbb{T}}\mathbf{x} + \boldsymbol{\xi}_{\mathbb{T}} \geq \mathbf{1} \\
& 0 \leq x_i, i = 1, \ldots, p \\
& 0 \leq \boldsymbol{\xi}_i, i \in \mathbb{T}
\end{aligned}$$

By introducing dual variables $\mu_i, i = 1, \ldots, t$, we could derive the dual problem as

$$\begin{aligned} \mathcal{D}(\boldsymbol{\mu}) = \max \quad & \sum_{i=1}^{t} \mu_i \\ \text{such that} \quad & \boldsymbol{\mu} A_{\mathbb{T}} \leq \mathbf{1}_m + \lambda \mathbf{1}^T A_{\mathbb{F}} \\ & 0 \leq \mu_i \leq \lambda, i = 1, \ldots, t \end{aligned}$$

Appendix B

APPENDIX FOR CHAPTER 5

B.1 Proof of Lemma 1

To show the convexity of (5.5), one needs to show that the constraint defines a convex set and the objective is a concave (convex) function if it is a maximization (minimization) problem. The linear constraint in (5.5) defines a polyhedron which is convex apparently. Therefore, we only need to prove the objective is a concave function, i.e., to prove that $\boldsymbol{\lambda}_1(\boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0 + \boldsymbol{\Lambda}_0 + \sum_{l=1}^{|\boldsymbol{B}^*|} v_l \boldsymbol{a}_l^T \boldsymbol{a}_l)$ is a concave function. Denote $\boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0 + \boldsymbol{\Lambda}_0 + \sum_{l=1}^{|\boldsymbol{B}^*|} v_l \boldsymbol{a}_l^T \boldsymbol{a}_l$ as $\boldsymbol{L}(\boldsymbol{v})$. If we could prove that $-\boldsymbol{\lambda}_1(L)$ is a convex function of L, then it is known that $-\boldsymbol{\lambda}_1(L(\boldsymbol{v}))$ is convex, since $L(\boldsymbol{v})$ is a linear transformation in terms of $\boldsymbol{v}$ which does not affect the convexity (or concavity) of the objective function. Thus, to show that $-\boldsymbol{\lambda}_1(L(\boldsymbol{v}))$ is convex, first, note that we could write $-\lambda_1(\boldsymbol{L}(\boldsymbol{v}))$ using the definition for the minimal eigenvalue of a positive semi-definite matrix as

$$\begin{aligned} -\boldsymbol{\lambda}_1(L(\boldsymbol{v})) &= -\inf_{\|z\|=1} \boldsymbol{z}^T L(\boldsymbol{v}) \boldsymbol{z} \\ &= -\inf_{\|z\|=1} \boldsymbol{z}^T L(\boldsymbol{v}) \boldsymbol{z} \\ &= \sup_{\|z\|=1} \langle -\boldsymbol{z}\boldsymbol{z}^T,\ L(\boldsymbol{v}) \rangle. \end{aligned}$$

To prove a function is convex, it is equivalent to show that its epigraph is a convex set. The epigraph of $-\lambda_1(\boldsymbol{L}(\boldsymbol{v}))$ is

$$\begin{aligned} \text{epigraph}(-\boldsymbol{\lambda}_1(L(\boldsymbol{v}))) &= \text{epigraph}(\sup_{\|z\|=1} \langle -\boldsymbol{z}\boldsymbol{z}^T,\ L(\boldsymbol{v}) \rangle) \\ &= \cap_{\|z\|=1}\ \text{epigraph}(\langle -\boldsymbol{z}\boldsymbol{z}^T, L(\boldsymbol{v}) \rangle). \end{aligned}$$

Since $\langle -\boldsymbol{z}\boldsymbol{z}^T, L(\boldsymbol{v})\rangle$ is a linear function, its epigraph is a half space above the hyperplane defined by the linear function. So it is a convex set. As we know, the intersection of convex sets is still a convex set. Therefore, we have that epigraph$(-\boldsymbol{\lambda}_1(L(\boldsymbol{v})))$ is a convex set, which proves the convexity of the function $-\boldsymbol{\lambda}_1(L(\boldsymbol{v}))$. It completes the proof.

B.2 Proof of Lemma 2

To show the function $\omega(\xi)$ is convex. First, we could see that the domain $\Omega = \{\xi | \exists \boldsymbol{v}, \mathbf{1}^T\boldsymbol{v} \leq \xi, \mathbf{0} \leq \boldsymbol{v} \leq \mathbf{1}\}$ is a convex set. Then let $\xi_1, \xi_2 \in \Omega$, then we could find two vectors $\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}$ such that $\mathbf{1}^T\boldsymbol{v}^{(1)} = \xi_1, \mathbf{1}^T\boldsymbol{v}^{(2)} = \xi_2$ and $\mathbf{0} \leq \boldsymbol{v}^{(1)} \leq \mathbf{1}, \mathbf{0} \leq \boldsymbol{v}^{(2)} \leq \mathbf{1}$. By the definition of convexity of a function, for any $\alpha \in (0,1)$, $\alpha\xi_1 + (1-\alpha)\xi_2 \in \Omega$, we need to show that

$$\omega(\alpha\xi_1 + (1-\alpha)\xi_2) \leq \alpha\omega(\xi_1) + (1-\alpha)\omega(\xi_2)$$

This can be shown by the following induction:

$$\begin{aligned}
&\omega(\alpha\xi_1 + (1-\alpha)\xi_2) \\
&\quad = \inf\{f(\boldsymbol{v}) | \mathbf{1}^T\boldsymbol{v} = \alpha\xi_1 + (1-\alpha)\xi_2, \mathbf{0} \leq \boldsymbol{v} \leq \mathbf{1}\} \\
&\quad \leq \inf\{f(\boldsymbol{v} = \alpha\boldsymbol{v}^{(1)} + (1-\alpha)\boldsymbol{v}^{(2)}) | \mathbf{1}^T\boldsymbol{v} = \alpha\xi_1 + (1-\alpha)\xi_2\} \\
&\quad \leq \alpha \inf\{f(\boldsymbol{v}^{(1)}) | \mathbf{1}^T\boldsymbol{v}^{(1)} = \xi_1, \mathbf{0} \leq \boldsymbol{v}^{(1)} \leq \mathbf{1}\} \\
&\qquad + (1-\alpha) \inf\{f(\boldsymbol{v}^{(2)}) | \mathbf{1}^T\boldsymbol{v}^{(2)} = \xi_2, \mathbf{0} \leq \boldsymbol{v}^{(2)} \leq \mathbf{1}\} \\
&\quad = \alpha\omega(\xi_1) + (1-\alpha)\omega(\xi_2)
\end{aligned}$$

Therefore we proved the convexity of $\omega(\xi)$, and we used the convexity of $f(\boldsymbol{v})$ in the last inequality.

B.3 Proof of Lemma 3

It can be seen that $\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T\boldsymbol{a}_k + \boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0$ is a singular positive semi-definite symmetric matrix. To see that, for any vector $\boldsymbol{z}$, we could show that $\boldsymbol{z}^T(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k\boldsymbol{a}_k^T\boldsymbol{a}_k + \boldsymbol{B}_0^T\boldsymbol{W}_0\boldsymbol{B}_0)\boldsymbol{z} = \sum_{k=1}^{|\boldsymbol{B}^*|}(v_k+w_k)(\boldsymbol{a}_k\boldsymbol{z})^2 \geq 0$, since $\forall k, v_k \geq 0, w_k \geq 0$. Since $\boldsymbol{B}_0\mathbf{1} = \mathbf{0}$, the vector $\mathbf{1}$ is in the null

space of $\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0$ and the corresponding eigenvalue is 0. On the other hand, it is known that for the matrix $\sigma_0^{-2}\boldsymbol{I}$, it has σ_0^{-2} as an eigenvalue and $\mathbf{1}$ as an eigenvector. Then we could draw the bounds of the smallest eigenvalue of $\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \sigma_0^{-2}\boldsymbol{I}$ as $\lambda_1(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0) + \lambda_1(\sigma_0^{-2}\boldsymbol{I}) \leq \lambda_1(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \sigma_0^{-2}\boldsymbol{I}) \leq \lambda_1(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0) + \lambda_p(\sigma_0^{-2}\boldsymbol{I})$, which is based on the Theorem 8.1.5 in [55]. This essentially imply that $\sigma_0^{-2} \leq \lambda_1(\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \sigma_0^{-2}\boldsymbol{I}) \leq \sigma_0^{-2}$. Therefore, the smallest eigenvalue of the matrix $\sum_{k=1}^{|\boldsymbol{B}^*|} v_k \boldsymbol{a}_k^T \boldsymbol{a}_k + \boldsymbol{B}_0^T \boldsymbol{W}_0 \boldsymbol{B}_0 + \sigma_0^{-2}\boldsymbol{I}$ is σ_0^{-2} with $\mathbf{1}$ as an eigenvector.

B.4 Bootstrap the observational data

Algorithm 1 Bootstrap the observational data

procedure Bootstrap(m), where m is the number of bootstrapping times

$i = 1$

While $i \leq m$ **do**

Re-sample the n instances from data set D_0^{obs} with replacement.

Apply an appropriate BN structural learning algorithm (i.e., the DAGlearn algorithm can be used for continuous BNs while the K2 algorithm can be used for discrete BNs) on the re-sampled dataset to learn $\hat{\phi}_0^i$

$i = i + 1$

end while

return $\{\hat{\phi}_0^i, i = 1, 2, \ldots, m\}$

end procedure

B.5 Facts of benchmark networks

Facts of Benchmark Networks from Bayesian Network Repository (BNR)

Data	**# Nodes**	**# Arc**	**Avg Degree**
Asia	8	8	2.00
Child	20	25	1.25
Insurance	27	52	3.85
Mildew	35	46	2.63
Alarm	37	46	2.49
Barley	48	84	3.50

B.6 Summary of experimental results when $\sigma^2 = 2$

Dataset	Scenario under $\sigma^2 = 2$	**Variance Reduction Ratio** $\pm s.d.$		
		@3th Iteration	**@6th Iteration**	**@9th Iteration**
Alarm	$nQuery = 1$	0.15 ± 0.10	0.15 ± 0.11	0.14 ± 0.12
	$nQuery = 2$	0.09 ± 0.11	0.08 ± 0.10	0.03 ± 0.15
	$nQuery = 4$	0.07 ± 0.06	0.04 ± 0.08	0.21 ± 0.24
Asia	$nQuery = 1$	0.18 ± 0.21	0.29 ± 0.26	1.26 ± 2.99
	$nQuery = 2$	0.20 ± 0.14	0.25 ± 0.25	0.06 ± 0.04
	$nQuery = 4$	0.16 ± 0.17	0.10 ± 0.27	-0.08 ± 0.12
Child	$nQuery = 1$	0.22 ± 0.30	0.25 ± 0.26	0.17 ± 0.23
	$nQuery = 2$	0.14 ± 0.12	0.15 ± 0.14	0.38 ± 0.35
	$nQuery = 4$	0.13 ± 0.13	0.22 ± 0.12	0.53 ± 0.68
Insurance	$nQuery = 1$	0.14 ± 0.12	0.19 ± 0.22	0.27 ± 0.31
	$nQuery = 2$	0.14 ± 0.13	0.16 ±0.18	0.27 ± 0.31
	$nQuery = 4$	0.08 ±0.08	0.13 ±0.12	0.41 ± 0.40
Mildew	$nQuery = 1$	0.12 ± 0.09	0.14 ± 0.08	0.22 ± 0.20
	$nQuery = 2$	0.08± 0.07	0.12 ±0.11	0.19 ± 0.21
	$nQuery = 4$	0.06 ± 0.05	0.07 ± 0.09	0.25 ± 0.22
Barley	$nQuery = 1$	0.12 ± 0.09	0.16 ±0.10	0.18 ± 0.12
	$nQuery = 2$	0.06 ± 0.12	0.05 ± 0.09	-0.03 ± 0.15
	$nQuery = 4$	0.02 ± 0.07	0.01 ± 0.14	0.02 ± 0.18

B.7 Summary of experimental results when $\sigma^2 = 4$

Dataset	Scenario under $\sigma^2 = 4$	**Variance Reduction Ratio** $\pm s.d.$		
		@3th Iteration	**@6th Iteration**	**@9th Iteration**
Alarm	$nQuery = 1$	0.12 ± 0.10	0.21 ± 0.19	0.21 ± 0.12
	$nQuery = 2$	0.12 ± 0.14	0.05 ± 0.09	0.03 ± 0.21
	$nQuery = 4$	0.05 ± 0.08	0.10 ± 0.09	0.20 ± 0.27
Asia	$nQuery = 1$	0.16 ± 0.11	0.47 ± 0.54	1.14 ± 1.56
	$nQuery = 2$	0.19 ± 0.14	0.16 ± 0.24	0.40 ± 1.20
	$nQuery = 4$	0.15 ± 0.09	0.05 ± 0.16	-0.05 ±0.10
Child	$nQuery = 1$	0.14 ± 0.11	0.20 ± 0.15	0.23 ± 0.26
	$nQuery = 2$	0.12 ± 0.07	0.24 ± 0.20	0.74 ± 1.10
	$nQuery = 4$	0.10 ± 0.11	0.20 ± 0.18	0.28 ± 0.45
Insurance	$nQuery = 1$	0.16 ± 0.14	0.21 ± 0.23	0.21 ± 0.18
	$nQuery = 2$	0.17 ± 0.16	0.12 ± 0.09	0.22 ± 0.19
	$nQuery = 4$	0.07 ± 0.05	0.12 ± 0.14	0.52 ± 0.39
Mildew	$nQuery = 1$	0.13 ± 0.08	0.13 ± 0.10	0.15 ± 0.15
	$nQuery = 2$	0.10 ± 0.10	0.11 ± 0.13	0.11 ± 0.21
	$nQuery = 4$	0.10 ± 0.06	0.06 ± 0.09	0.21 ± 0.25
Barley	$nQuery = 1$	0.11 ± 0.09	0.14 ± 0.11	0.16 ± 0.11
	$nQuery = 2$	0.09 ± 0.06	0.07 ± 0.11	0.05 ± 0.16
	$nQuery = 4$	0.02 ± 0.07	0.01 ± 0.07	0.01 ± 0.19

B.8 Evaluation of the proposed Bayesian learning and sensing framework

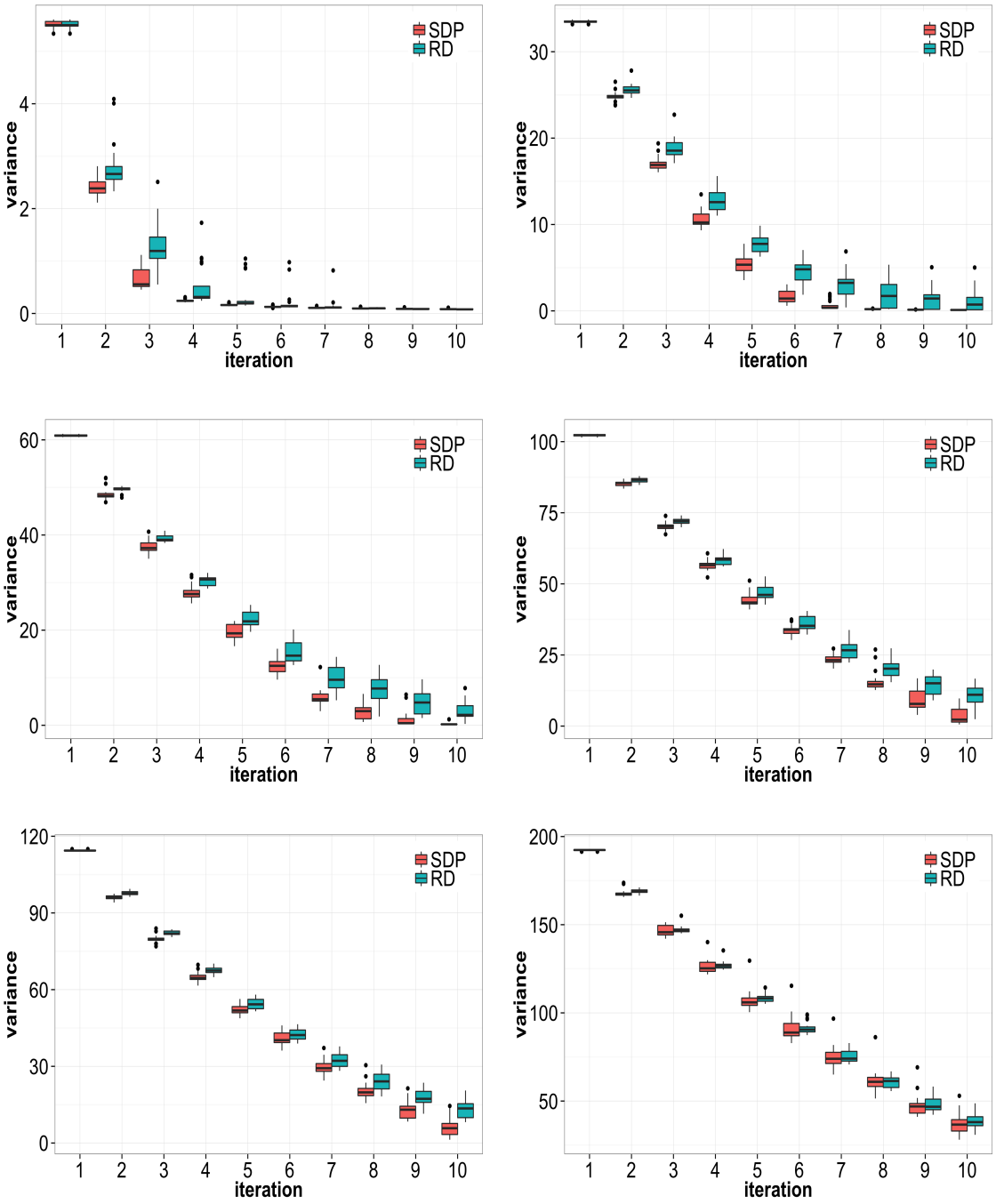
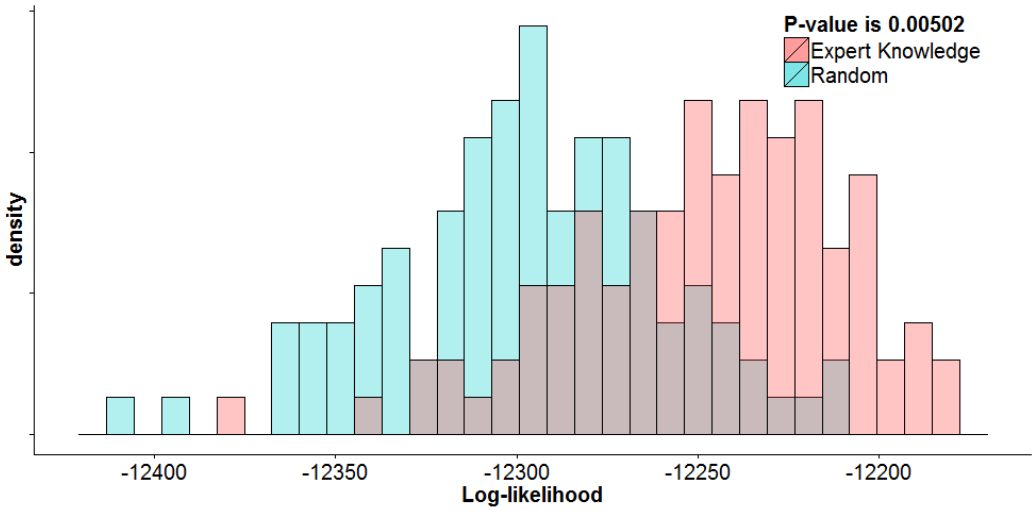


Figure B.1: Evaluation of the proposed Bayesian learning and sensing framework, with either the SDP (red) method or the random selection method (blue), for reducing the variance of estimation of the ordering. Each figure corresponds to a network, which are asia, child, insurance, mildew, alarm, barley, from left to right and top to down.

B.9 Ordering of human resource management key performance indicators

Ordering of Human Resource Management Key Performance Indicators (KPIs)		
	Index	**Meaning**
Driver KPI	hr-hrcrn	Compensation, Reward, and Incentive Systems
	pm-hsmbn	Management Breath of Experience
	sp-hspcn	Reward-Manufacturing Coordination
	sp-hsrsn04	Hiring Criteria: work values and attitudes
	sp-hsrsn10	Hiring Criteria: technical skills
	pm-hsvcn	Cooperation
Mediator KPI	sp-hsrsn02	Employee Trait: teamwork
	sp-hsrsn03	Employee Trait: problem solving
	sp-hsrsn07	Hiring Process: large candidate pool
	sp-hsrsn08	Hiring Process: effective interview
Downstream KPI	sp-hsmfn	Multi-functional Employees
	hr-hrtrn02	Turnover rate of Hourly Employees
	sp-psltn	Supply Lead Time

B.10 Validation in the KPI case study



Validation of the utility of the expert comparison data in the KPI case study. It clearly shows that the expert comparison data is significantly different from random guess.

B.11 Uncertainty of ordering of the hypermetabolism reduction

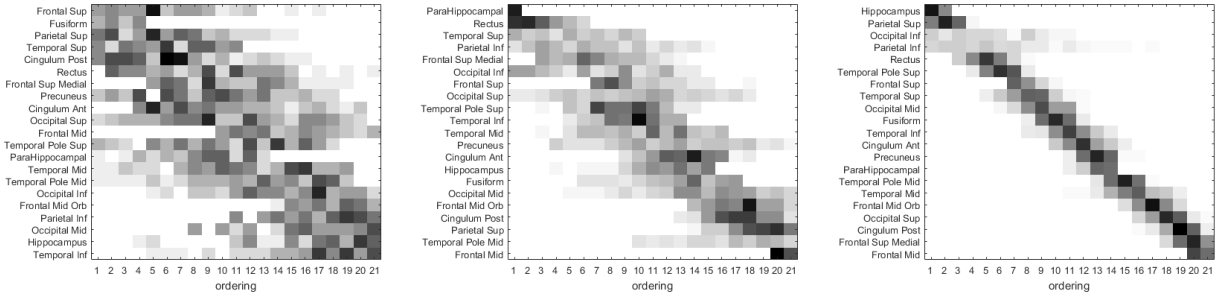


Figure B.2: Uncertainty of ordering of the hypermetabolism reduction events when only observational data is used (top), observational data and 10 expert comparisons are used (middle), and observational data and 20 expert comparisons are used (bottom), respectively. Note that, the rows correspond to the hypermetabolism reduction events while the numbers in the x-axis represent the ordering of the hypermetabolism reduction events.

B.12 Validation in the AD case study

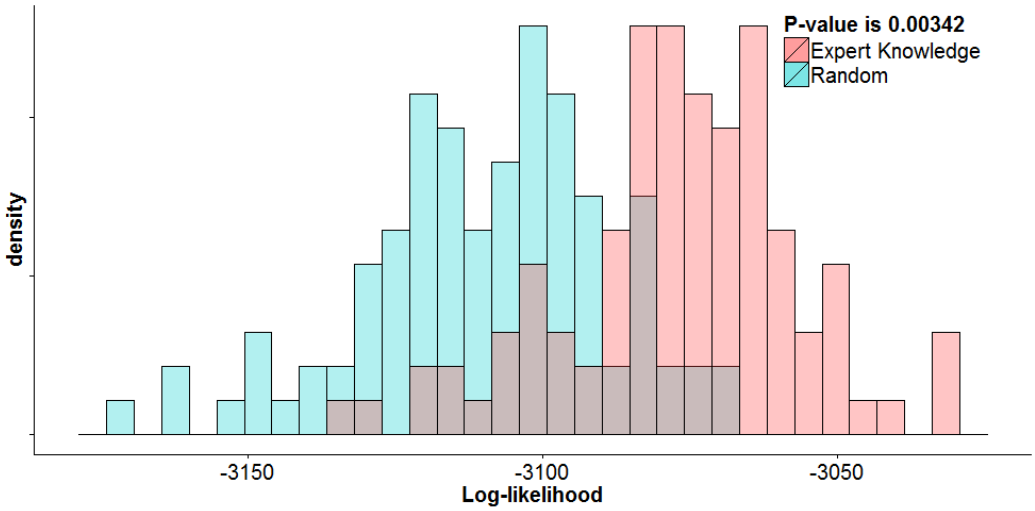


Figure B.3: Validation of the utility of the expert comparison data in the AD case study. It clearly shows that the expert comparison data is significantly different from random guess.

Appendix C

APPENDIX FOR CHAPTER 6

Table C.1: Nomenclature

Notation	**Definition**
$X \in \mathbb{R}^{n \times p}$	instance
$Y \in \mathbb{R}^{n \times 1}$	label
$\mathcal{H}$	a set of trained classifier
$h_i \in \mathcal{H}$	any real-valued classifier
$\mathcal{O}$	the set of classifiers that have already been evaluated (empty at the beginning)
$m'_i \in \mathbb{R}$	the response/margin of the i-th classifier, $h_i(X)$
$\boldsymbol{m}'_{\mathcal{O}} \in \mathbb{R}^{\lvert\mathcal{O}\rvert \times 1}$	the random vector of observed classifiers as $[m'_{o_1}, \ldots, m'_{o_{\lvert\mathcal{O}\rvert}}]^T$
m_i	the actual observed value of classifier i
$\boldsymbol{m}_{\mathcal{O}} \in \mathbb{R}^{\lvert\mathcal{O}\rvert \times 1}$	the vector of actual observed value
$P(Y \lvert \boldsymbol{m}_{\mathcal{O}})$	the posterior over Y
$C(h_i \lvert \mathcal{O})$	the computational cost of evaluating classifier h_i conditioned on the set of evaluated classifiers $\mathcal{O}$
θ'_i	the ability of classifier i
δ'_i	the difficulty of classifier i
θ_i	$\log(\theta'_i)$
σ_i	$\log(\sigma'_i)$
$Y_i \in \{0, 1\}$	a dichotomous variable that indicates the prediction is correct or incorrect
p_{i1}, p_{i0}	the probability that classifier i predicts correctly and incorrectly respectively
m	total amount of classifiers
n	total amount of classifiers that is going to achieve
c	the category of classifiers
n_c	total amount of classifiers that is going to achieve in category c
$I_i(\theta)$	the information of classifier i at the ability θ