A FAULT DIAGNOSIS TECHNIQUE FOR COMPLEX SYSTEMS
USING BAYESIAN DATA ANALYSIS

A Thesis
Presented to
The Academic Faculty

by

Young Ki Lee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Aerospace Engineering

Georgia Institute of Technology
April 2008

Copyright © 2008 by Young Ki Lee
A FAULT DIAGNOSIS TECHNIQUE FOR COMPLEX SYSTEMS
USING BAYESIAN DATA ANALYSIS

Approved by:

Prof. Vitali V. Volovoi, Committee Chair
School of Aerospace Engineering
Georgia Institute of Technology

Prof. Dimitri N. Mavris, Advisor
School of Aerospace Engineering
Georgia Institute of Technology

Mr. Ted Fisher
Gas Turbine Application Engineering
General Electric Energy Services

Prof. Ming Yuan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Hongmei Chen
School of Aerospace Engineering
Georgia Institute of Technology

Date Approved: April 2008
To my parents, brother, and sisters.
ACKNOWLEDGEMENTS

This dissertation could not have been completed if it were not for my teachers, colleagues, friends, and family. I wish to thank all of them from deep in my heart.

I would like to thank Professor Dimitri Mavris for allowing me to join ASDL and supporting years of my study. Especially, His philosophy of research and the scientific method has been very inspiring during this thesis work.

I would also like to thank Professor Vitali Volovoi. He was a research engineer, first, and promoted to an assistant professor later. He ended up being my co-advisor. No matter a position he was in, he always has been a great supporter and advisor in many ways.

I would particularly like to thank Dr. Ming Yuan for helping me break through all the statistical obstacles during this work. The method presented in this dissertation is based on his advise given in the very first meeting with him, which was only a half hour long. He also inspired my interest in statistics.

I have had several great supporters at GE Energy Services ever since my internship there. Mr. Michael Sullivan was my supervisor while I was an intern. This research was initiated under his supervision. Mr. Ted Fisher was so kind enough to be one of the thesis committee members in spite of his busy schedule in his organization. His encouragement was a great motivation for me to go on with this research.

This dissertation is mostly based on my work with the GE project team. I would like to thank the team leader Dr. Hongmei Chen and the team members, especially, Brian Kestner and Stephanie Mma. Hongmei brought me to the Bayesian world and has been supportive not only as the team leader but also as a committee member. Brian was my research partner when I started this project and provided me enormous help regarding to gas turbines and GE gas turbine tool GTP. He deserves a part of the credit for one of the applications presented in this dissertation. Stephanie ran GTP and created regression equations for me while I could not access to the GE propriety tool.
I would also like to thank to the members of the AFOSR status matching project team: Mr. Russell Denney, Dr. Sriram Rallabhandi, Damon Rousis, and Jacob Chang. One of the applications in this dissertation was accomplished through the collaboration with them.

Among my colleagues I would like to thank Drs. Brian German and Jack Zentner, and Tae Choi. Brian was so kind and sincere enough to help me whenever I asked even though he was not obligated to do so. Jack, a fellow statistics enthusiast, provided me a mathematica code that he developed during his doctoral research. The mathematica code inspired the graph-based correlation plots I used in this dissertation. Tae set up LaTex on my computer for me and taught me how to use it. I wish the best luck in his career.

My final thanks are toward many friends in my personal life: Carlos, Eddie, Josh, Kike, Marco, Mickey, and Sharock, to name seven in alphabetical order. I apologize to whom I did not name here. Without them I could have not finished my study with this much joy.

For the rest of my life I will do my best paying everyone back for what he or she has done for me.
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1</td>
<td>The Method of Least Squares</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Sensitivity of a Linear System of Equations</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Regularization Techniques</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Statistical Inversion Theory</td>
</tr>
<tr>
<td>3.3</td>
<td>Bayesian Statistics</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Prior distributions</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Computation of Posterior Distributions</td>
</tr>
<tr>
<td>3.4</td>
<td>Model Uncertainty</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Model Selection</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Bayesian Model Averaging</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Bayesian Model Selection and Averaging in Linear Regression</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Difference Between Regression and Fault Identification</td>
</tr>
<tr>
<td>IV</td>
<td>RESEARCH FORMULATION</td>
</tr>
<tr>
<td>4.1</td>
<td>Observations and Research Questions</td>
</tr>
<tr>
<td>4.2</td>
<td>A Fault Identification Method Inspired By Model Selection and Averaging</td>
</tr>
<tr>
<td>4.3</td>
<td>Boundary of the Thesis</td>
</tr>
<tr>
<td>V</td>
<td>PRELIMINARY STUDY</td>
</tr>
<tr>
<td>5.1</td>
<td>General Theory of Bayesian Networks</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Discrete Networks</td>
</tr>
<tr>
<td>5.1.2</td>
<td>A Linear Gaussian Model</td>
</tr>
<tr>
<td>5.2</td>
<td>Example Application</td>
</tr>
<tr>
<td>5.3</td>
<td>Implementation</td>
</tr>
<tr>
<td>5.3.1</td>
<td>The Basic Network</td>
</tr>
<tr>
<td>5.3.2</td>
<td>The Bias-Augmented Network</td>
</tr>
<tr>
<td>5.4</td>
<td>Chapter Summary</td>
</tr>
<tr>
<td>VI</td>
<td>PROPOSED METHOD</td>
</tr>
<tr>
<td>6.1</td>
<td>Formulation</td>
</tr>
<tr>
<td>6.1.1</td>
<td>A Bayesian Model for the System of Linear Equations</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Multiple Models</td>
</tr>
<tr>
<td>6.1.3</td>
<td>Posteriors $p(M</td>
</tr>
</tbody>
</table>
List of Tables

1  Jeffreys’ Interpretation of Bayes Factor ............................ 44
2  Guidelines Regarding to Bayes Factor by Kass and Raftery ............ 44
3  Laboratory Settings of the Quadruple-Tank Process ..................... 60
4  Maximum A Posterior Estimate of each Variable from the Bias-Augmented Network with Various $\sigma_B$: the Quadruple-Tank Process ............... 67
5  Maximum A Posterior Estimate of Each Variable from the Various Networks: the Quadruple-Tank Process .................................. 68
6  The State Variables and the Observable of the Turbojet Engine ........... 85
7  Correlation Coefficients: the Turbojet Engine with the Degraded Compressor 92
8  Models with a Posterior Probability above 0.02: the Turbojet Engine with the Degraded Compressor .................................................. 96
9  Variable Setting for the Multiple Degraded Components .................. 102
10 Models with the Eight Highest Posterior Probabilities: the Turbojet Engine with the Degraded Major Components ............................... 102
11 The State Variables and the Observable of the GE 7FA+e Gas Turbine ... 103
12 Ranges of the State Variables and the Bias Variables of the GE 7FA+e Gas Turbine ................................................................. 108
13 Variable Setting for Case 1 of the GE 7FA+e Gas Turbine .................. 109
14 Correlation Coefficients: GE 7FA+e Gas Turbine ........................... 113
15 Variable Setting for Case 2 of the GE 7FA+e Gas Turbine .................. 122
16 Variable Setting for Case 3 of the GE 7FA+e Gas Turbine .................. 124
17 Comparison of the Current Method with Others ............................. 137
**List of Figures**

1. Worldwide Commercial Jet Operation Through 2005 ........................................... 1
2. World Annual Traffic ..................................................................................................... 2
3. Chimerical Jet Airplane Accident Rate from 1959 through 2005 ................................. 2
4. NASA Aviation Safety Program Organization .................................................................. 4
5. Fatalities by CAST/ICAO Taxonomy Accident Categories During 1987 Through 2005 .................................................................................................................. 7
6. OSACBM Architecture .................................................................................................. 10
7. A Simple Bayesian Network ......................................................................................... 15
8. Basic and Bias-Augmented Models in the Bayesian Network Context .......................... 15
9. Example of Control Chart ............................................................................................ 17
10. Model-Based Diagnosis ............................................................................................... 21
11. The Smearing Effect ................................................................................................. 23
12. Similarity in Signature Pattern Causing the Isolability Issue ..................................... 24
13. Difference in Signature Magnitude Causing the Sensitivity Issue .............................. 24
15. The Linguistic Values of “Age” .................................................................................. 27
16. Transformation from a Circle to an Ellipse by Mapping Through the Matrix $A$ ........ 32
17. Example of Ill-Conditioned Systems ......................................................................... 33
18. Examples of Conditional Mean and Maximum a Posterior Estimates ..................... 35
19. Rejection Sampling ..................................................................................................... 40
20. Gibbs Sampling .......................................................................................................... 41
21. A Hierarchical Bayesian Model .................................................................................. 47
22. Example of Regression Problems .............................................................................. 48
23. Example of Fault Identification Problems .................................................................. 49
24. Discretization of a Continuous Variable ..................................................................... 57
25. Schematic of Quadruple-Tank Process ....................................................................... 59
26. Basic Network of the Quadruple-Tank Process ...................................................... 61
27. Simulated Measurements with $(X_1, X_2) = (3\sigma_X, -3\sigma_X)$: Case 1 of the Quadruple-Tank Process ................................................................. 63
# LIST OF SYMBOLS OR ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence.</td>
</tr>
<tr>
<td>AIC</td>
<td>An Information Criterion.</td>
</tr>
<tr>
<td>ANNs</td>
<td>Artificial Neural Networks.</td>
</tr>
<tr>
<td>AvSP</td>
<td>Aviation Safety Program.</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion.</td>
</tr>
<tr>
<td>BMA</td>
<td>Bayesian Model Averaging.</td>
</tr>
<tr>
<td>BN</td>
<td>Bayesian Network.</td>
</tr>
<tr>
<td>CAIV</td>
<td>Cost As an Independent Variable.</td>
</tr>
<tr>
<td>CBR</td>
<td>Case-Based Reasoning.</td>
</tr>
<tr>
<td>COMPASS</td>
<td>Condition Monitoring and Performance Analysis Software System.</td>
</tr>
<tr>
<td>CPD</td>
<td>Conditional Probability Distribution.</td>
</tr>
<tr>
<td>DoD</td>
<td>Department of Defense.</td>
</tr>
<tr>
<td>DOE</td>
<td>Design of Experiments.</td>
</tr>
<tr>
<td>DR</td>
<td>Data Reconciliation.</td>
</tr>
<tr>
<td>FAA</td>
<td>Federal Aviation Administration.</td>
</tr>
<tr>
<td>FDA</td>
<td>Fisher Discriminant Analysis.</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic Algorithm.</td>
</tr>
<tr>
<td>GE</td>
<td>General Electric.</td>
</tr>
<tr>
<td>GED</td>
<td>Gross Error Detection.</td>
</tr>
<tr>
<td>GPA</td>
<td>Gas Path Analysis.</td>
</tr>
<tr>
<td>GRC</td>
<td>Glenn Research Center.</td>
</tr>
<tr>
<td>GTP</td>
<td>Gas Turbine Performance.</td>
</tr>
<tr>
<td>HM</td>
<td>Health Management.</td>
</tr>
<tr>
<td>HUMS</td>
<td>Health Usage Monitoring System.</td>
</tr>
<tr>
<td>IVHM</td>
<td>Integrated Vehicle Health Management.</td>
</tr>
<tr>
<td>JPD</td>
<td>Joint Probability Distribution.</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo.</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimate.</td>
</tr>
<tr>
<td>NASA</td>
<td>National Aeronautics and Space Administration.</td>
</tr>
<tr>
<td>NPSS</td>
<td>National Propulsion System Simulation.</td>
</tr>
<tr>
<td>NTSB</td>
<td>National Transportation Safety Board.</td>
</tr>
<tr>
<td>OSACBM</td>
<td>Open System Architecture Condition Based Maintenance.</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis.</td>
</tr>
<tr>
<td>PITEX</td>
<td>Propulsion IVHM Technology Experiment.</td>
</tr>
<tr>
<td>SARS</td>
<td>Severe Acute Respiratory Syndrome.</td>
</tr>
<tr>
<td>SLRT</td>
<td>Sequential Likelihood Ratio Test.</td>
</tr>
<tr>
<td>SPRT</td>
<td>Sequential Probability Ratio Test.</td>
</tr>
<tr>
<td>SSME</td>
<td>Space Shuttle Main Engine.</td>
</tr>
<tr>
<td>SSVS</td>
<td>Stochastic Search Variable Selection.</td>
</tr>
<tr>
<td>USAF</td>
<td>United States Air Force.</td>
</tr>
<tr>
<td>VAATE</td>
<td>Versatile Affordable Advanced Turbine Engine.</td>
</tr>
<tr>
<td>WLS</td>
<td>Weighted Least Squares.</td>
</tr>
</tbody>
</table>
This research develops a fault diagnosis method for complex systems in the presence of uncertainties and possibility of multiple solutions. Fault diagnosis is a challenging problem because data used in diagnosis contain random errors and often systematic errors as well. Furthermore, fault diagnosis is basically an inverse problem so that it inherits unfavorable characteristics of inverse problems: The existence and uniqueness of an inverse solution are not guaranteed and the solution may be unstable. The weighted least squares method and its variations are traditionally used for solving inverse problems. However, the existing algorithms often fail to identify multiple solutions if they are present. In addition, the existing algorithms are not capable of selecting variables systematically so that they generally use the full model, which may contain unnecessary variables as well as necessary variables. Ignoring this model uncertainty often gives rise to, so called, the smearing effect in solutions. The proposed method solves the inverse problem using Bayesian inference. An engineering system can be parameterized using state variables. The probability of each state variable is inferred from observations made on the system. A bias in an observation is treated as a variable, and the probability of the bias variable is inferred as well. To take the uncertainty of model structure into account, multiple Bayesian models are created with various combinations of the state variables and the bias variables. The results from all models are averaged according to how likely each model is. The method is demonstrated for two applications: the status matching of a turbojet engine and the fault diagnosis of an industrial gas turbine. In the status matching application only physical faults in a turbojet engine are considered, whereas in the fault diagnosis application sensor biases are considered as well as physical faults. The proposed method is tested in various faulty conditions using simulated measurements. Results show that the proposed method identifies physical faults and sensor biases simultaneously. Overall, there is a clear improvement in ability to identify correct solutions over the full model that contains all state and bias variables.
Chapter I

INTRODUCTION

Although there was a decrease in world-wide jet operations 2001 and 2002 due to the 9/11 terrorist attacks and the outbreak of SARS (Severe Acute Respiratory Syndrome), the reduced traffic has recovered, and jet operations have returned to an increasing trend [16]. According to a forecast done by the Airbus company [1], revenue passenger kilometers will increase by 5.9% annually over the twenty-year period from 2004 to 2023 as shown in Figure 2. While air traffic has increased, the accident rate of the worldwide commercial jet fleet has decreased because of the advances in technology and the increase in reliability. Despite the low accident rate, the absolute number of accidents is expected to be large due to the large volume of operations. An increased awareness of aviation safety motivates the aerospace research agencies and industry to pursue new technologies falling into the category of Health Management (HM).

In addition to the safety issue, HM has been pursued with a second motivation: the economic benefit. The Department of Defense (DoD) operates approximately 25,000 aircraft powered by 50,000 engines. Overhaul & Maintenance magazine reported in its November issue in 2003 that, according to an estimate by the Logistics Management Institute, a non-profit consulting organization, the U.S. military spent $20 billion dollars on the maintenance

![Figure 1](image)

**Figure 1:** Worldwide Commercial Jet Operations Through 2005 [16]
Figure 2: World Annual Traffic [1]

Figure 3: Commercial Jet Airplane Accident Rate Through 2005 [16]
and parts for its aircraft yearly at all levels from field to depot [20]. The military envisions that HM systems will determine if their assets need maintenance or disposal and predict when the useful life of parts ends in near future. The prediction of useful life eliminates unnecessary maintenance or disposal of usable parts and further enables efficient logistics.

With these two different motivations, HM has been researched by various government agencies and the aerospace industry.

1.1 Increased Attention to Aviation Safety

In February 1997, the White House Commission on Aviation Safety and Security, chaired by the then vice-president Al Gore, recommended to the then president Clinton that government and industry “should establish a national goal to reduce the aviation fatal accident rate by a factor of five in ten years and conduct safety research to support that goal [115].” The commission also urged the Federal Aviation Administration (FAA), the National Aeronautics and Space Administration (NASA), and the industry to combine their efforts to meet this challenge.

In December 1997, another commission, the National Civil Aviation Review Commission, listed a number of specific issues on aviation safety and again urged the FAA and the aviation industry to develop a strategic plan to resolve these issues [90].

In response to this challenge, government agencies and the aviation industry have initiated safety-related programs both separately and collaboratively. In April 1998, the FAA announced the Safer Skies initiative to achieve the accident rate reduction goal by 2007. The main idea was to find the prevailing root causes of accidents and to determine the best actions to prevent the situations that lead to accidents. The project took a systematic data-driven approach to accident and incident data including National Transportation Safety Board (NTSB) reports [43]. The NTSB investigates all civil aviation accidents that involve death, serious injury, or substantial damage.

In the mean time, NASA initiated the Aviation Safety Program (AvSP) [105] in 1999 to study new technologies for improving safety. NASA’s goal is the reduction of the fatal aircraft accident rate by 80 percent in 10 years and by 90 percent in 25 years. The program
Figure 4: NASA Aviation Safety Program Organization (recreated from [109])

has been led by the NASA Langley Research Center, and the six projects has been performed at multiple NASA research centers: (1) aviation system monitoring and modeling, (2) system-wide accident prevention, (3) single aircraft accident prevention, (4) weather accident prevention, (5) accident mitigation, and (6) synthetic vision [105]. Each of these areas focuses on the development of technologies to improve safety in various aspects of aviation systems. The program organization chart is shown in Figure 4.

The single aircraft accident prevention element develops and supports the aircraft-specific technologies that will reduce the accident rate. This element focuses on the following research areas: (1) vehicle health management and flight critical systems design, (2) propulsion health management, and (3) control upset prevention and recovery [105]. The vehicle health
management and flight critical systems design area develops the techniques that detect the
degradation in system components before it results in a catastrophic event. Likewise, the
propulsion health management area develops techniques specifically for propulsion systems.

1.2 Emergence of Affordability

The U.S. military has consistent pressure on its budget every year. The financial pressure
on the military led the Department of Defense (DoD) to propose an initiative in 1995 for
defense systems acquisition: Cost As an Independent Variable (CAIV). Its main purpose
is to reduce the life cycle costs of defense systems. CAIV Working Group describe CAIV
as a strategy that “entails setting aggressive, realistic cost objectives for acquiring defense
systems and managing risks to obtain those objectives [19].”

The new philosophy in defense systems acquisition altered the research and development
direction of defense systems. In 1999, the U.S. Air Force (USAF) initiated the Versatile
Affordable Advanced Turbine Engine (VAATE) program [39] in collaboration with the De-
partment of Energy (DoE) and the NASA. The goal of the VAATE program is a 10-fold
improvement in affordability by 2017. Whereas its preceding program, Integrated High
Performance Turbine Engine Technology Program, uses capability as a metric, the VAATE
program uses Cost Capability Index as a metric, which is defined as capability divided by
cost [38]. The VAATE program consists of three focus research areas: versatile core, dura-
bility, and intelligent engines. As part of the intelligence engines area, an engine health
management plan comprising four elements is proposed: (1) maintenance process control,
(2) diagnostics, (3) prognostics, and (4) active control [39].

1.3 Health Management

1.3.1 Health Management for Aerospace Systems

Although the AvSP and VAATE programs have been initiated with different motivations,
safety and affordability, respectively, Health Management (HM) is the technology envisioned
to fulfill the goals of both programs. Health management in the field of aerospace engineer-
ing has been referred to various names: Vehicle Health Management, Integrated Vehicle
Health Management (IVHM), and Integrated Systems Health Management, etc. Additionally, Structural Health Management [114] is restrictively used in the particular discipline, and so is Propulsion Health Management [109]. As prognosis emerges as a critical element, the new term Prognosis and Health Management [55] is coined.

Nowadays many aerospace systems are equipped with the on-board Health Management (HM) systems, and the aerospace systems operators use the ground-based HM systems to maintain them.

Due to the complexity of vehicles and the high risk of mission, condition monitoring systems have been extensively used in the space explore programs led by NASA. Especially, the NASA Glenn Research Center (GRC) has focused on propulsion systems such as Space Shuttle Main Engine (SSME) [83]. The GRC is currently leading the propulsion system health management task in the AvSP program and conducting the Propulsion IVHM Technology Experiment (PITEX) program collaboratively with the Ames and Kennedy Research Centers [85].

As well as the propulsion systems for spacecrafts, the turbine engines for aircraft are also operated or maintained with aid of the HM systems. The USAF operates a variety of aircraft and engines developed in different period from TF33 developed in 1948 for B-52 Stratofortress to F119 for F-22 Raptor. The wide spread of technology makes their maintenance and logistics tasks especially complicated. In order to facilitate the maintenance, the USAF has been used the computer program CETADS (Comprehensive Engine Trending and Diagnostic System) [58]. CETADS has capabilities to flag trends that indicate an impending failure, and to extrapolate the data for forecasting future events. Their turbine engine health management research is continued in the intelligent engines task of the VAATE program.

The health management system for helicopters is called the Health and Usage Monitoring System (HUMS). Helicopters are equipped with safety-critical rotating machinery. Their propulsion systems produce lift as well as thrust. Military, the biggest operator of helicopters, and the helicopter industry started developing HUMS in reaction to their concern over the airworthiness of helicopters. Since the first flight with a certified HUMS in United Kingdom in 1991, the acceptance of HUMS is continuously growing. In 1999, the
FAA issued an advisory circular to provide guidance for HUMS installation [37].

1.3.2 Role of Health Management Systems

1.3.2.1 Safety Perspective

Among the fatal accidents occurred to worldwide commercial jet fleet from 1987 though 2005, 11 percent of the on-board fatalities was caused by system/component failure or malfunction. The system/component failure or malfunction is the third main cause of the on-board fatalities among nineteen causes as shown in Figure 5. Health Management (HM) is envisioned to reduce this portion of fatalities.

The efforts to detect and overcome system/component failure or malfunction motivated some damage adaptive technologies such as Fault Detection, Isolation, and Reconfiguration [15], Fault Detection, Isolation, and Accommodation [95], and Self-Repairing Flight Control System [121]: the predecessors of HM. With these technologies accompanied by intelligent control systems, on-board HM systems detect incipient faults and accommodate the impact of faults in the control logic to avoid catastrophic events.
1.3.2.2 Economic Perspective

The role of Health Management (HM) systems on ground is to support maintenance. The maintenance strategy of machinery has evolved as the machinery becomes modernized. The evolution of maintenance strategy can be traced through three generations: corrective, preventive, and predictive maintenance generations [88].

The corrective maintenance is to fix a machine when it is not functional. The first generation covers the World War II era. After the war the lack of manpower led the modernization of machine and the automation of manufacturing process. The more industry depended on machines, the deeper its failure affected on the productivity and quality of production. Thus, people felt the need of failure prevention, and industry started perform overhaul with time interval to prevent the loss of assets. This is called preventive maintenance. As machinery takes scheduled maintenance throughout its lifetime and skilled maintenance experts are needed to perform the maintenance, maintenance cost dramatically increased. In order to make profit out of assets, the owner of the assets has to maximize their utility and minimize maintenance cost. The maximum utility can be achieved by reducing unnecessary repair on normally operating machine and unscheduled emergency repair. In other words, maintenance needs to be done “just-in-time.” The new maintenance strategy is called the predictive maintenance. One maintenance strategy does not replace another; the strategy varies a machine to machine and a part to part. The decision is made based on the trade-off that takes cost and effectiveness into consideration.

In order to achieve the “just-in-time” maintenance, the machinery should be continuously monitored, its condition should be diagnosed, and the failure time in the future should be predicted. These needs led the integration of three different elements: condition monitoring, diagnosis, and prognosis. These elements mainly constitute a HM system.

1.3.3 Elements of a Health Management System

In today’s aerospace development and manufacturing environment, many complex mechanical or electrical sub-systems are integrated into an air vehicle. While aircraft manufacturers such as Boeing and Airbus take a role as an integrator, many contractors supply sub-systems,
components, and parts. Some of the sub-systems, components, and parts, especially avionics and electronics, may be embedded with their own elementary diagnostic functions. If each sub-system supplier provides a diagnostic system with proprietary data interface and algorithm, the integrator has to rework extensively on the interfaces and algorithms to make them compatible or to connect with other sub-systems.

This difficulty in integration has led the aerospace industry to an “open systems” approach. A broad definition of open systems is a system design approach that facilitates the integration and interchangeability of components provided from various sources [116]. In 1999, a consortium led by Boeing initiated the development of an open architecture enabling multiple vendors to competitively contribute to a Health Management system. The architecture was entitled Open System Architecture for Condition Based Maintenance (OSACBM). The OSACBM defines required components of a HM system as well as the interfaces and protocols between the components. The architecture consists of seven layers shown in Figure 6. Each layer has a capability to request data from other layers. The sensor module collects sensor measurements, control information, and other types of raw data. The signal processing module transforms the raw data collected from the sensor module into other useful forms. For example, it may perform Fast Fourier Transformation to transform signals into frequency domain or it may take average over time to filter out noise. The condition monitoring module receives the output of the precedent modules, and extracts features that indicate the condition of the system. The extracted features may be compared with expected values or operational limits. The health assessment module performs diagnosis; it determines if the monitored system is deteriorated or not. If the system’s health is deteriorated, the health assessment module reports one or more possible fault conditions. The prognostic module projects the current health of the monitored system into future and forecasts the time of failure. The decision support module recommends maintenance actions. The presentation module interfaces the layers with the user. The first five modules are crucial whereas the last two modules subsidiary.
1.3.4 Challenges in the Development of a Health Management System

The seven layers of OSACBM explained in §1.3.3 are interlinked with each other. The output of one layer becomes the input of another layer. Therefore, an error in one layer triggers a chain reaction; the error propagates to other layers.

Measurements obtained from a sensor are inevitably corrupted by random noise. Sensor measurements can be systematically biased or totally wrong when the sensor is miscalibrated or malfunctioning, although these events are less likely than the random noise. Thus, measurements have to be validated before they are fed to other layers; invalid measurements have to be reconciled or excluded if the reconciliation is impossible.

Most diagnosis algorithms use sensor measurements to estimate the health condition of the object to be diagnosed. The algorithms are not perfect either. Even with valid measurements, they can diagnose a system incorrectly. A diagnosis commits two types of errors: false and missed diagnoses, if the diagnostic decision is binary choice between faulty
and normal. Both the false or missed diagnoses are harmful. A false diagnosis will cause the removal of usable parts, and consequently, increase the maintenance cost, which is the direct opposite of what Health Management (HM) intends for. On the other hand, a missed diagnosis may lead a correctable fault to a failure of system and, eventually, a catastrophic event. Furthermore, the error in diagnosis adversely affect on the next element, prognosis. Prognosis projects the current state of a system to the future. If the current state of the system is wrongfully diagnosed, the subsequent prognosis cannot be right.
Chapter II

DATA VALIDATION AND FAULT DIAGNOSIS

In the previous chapter the importance of valid data and correct diagnosis was emphasized. Poor quality of data corrupts all other elements of a health management system. Incorrect diagnosis give rise to incorrect prognosis.

This chapter discusses about data validation and fault diagnosis in detail and presents numerous techniques currently used in various filed.

2.1 Data Validation

A sensor consists of several parts such as a sensing device, transducer, and communication device. When one of these parts malfunctions, the sensor signal differs from normal signals. The difference appears as drift, bias, short, spike, and precision degradation.

Mathematically speaking, when a sensor measures a property, the measured value $y$ equals to the sum of the true value $x$ and the random error $\varepsilon$:

$$ y = x + \varepsilon $$

(1)

The random error $\varepsilon$ is usually modeled with a normal distribution with zero mean and a constant standard deviation $\sigma$: $N(0, \sigma)$. The standard deviation is a parameter closely related to the precision of the sensor. The inverse of variance, $1/\sigma^2$, is called precision.

When the sensor is improperly calibrated, a measurement from the sensor becomes biased. The bias, which is a systematic error, affects the mean of the measurement. When the bias is taken into consideration, Equation 1 becomes

$$ y = x + b + \varepsilon $$

(2)

where $b$ is the unknown bias.
2.1.1 Data Reconciliation

In the field of chemical engineering, a procedure, called data reconciliation (DR) [89], has been developed to remove random errors in measurements obtained from chemical processes. Physics enforces the properties in a chemical process to follow physical laws, e.g., the mass and energy balances. Therefore, the measurements of the properties have to satisfy the process constraints. Random errors contained in the measurements, however, give rise to a violation of the constraint. DR adjusts the measurements until they satisfy the constraints. A reconciled value is an estimate of the true value.

2.1.1.1 Weighted Least Squares Method

One of the approaches to solve the DR problem is the Weighted Least Squares (WLS) method. The formulation of the WLS method is as follows:

\[
\text{minimize} \quad (y - x)^T W (y - x) \\
\text{subject to} \quad g(x) = 0 \\
\text{by changing} \quad x
\]

where \( y \) is a vector of measurements, \( x \) is a vector of the estimate of the true value, \( W \) is the matrix of weighting factors, and \( g \) is a constraint. Assuming the random error \( \varepsilon = y - x \) follows a normal distribution with zero mean, the objective function becomes equivalent to the maximum likelihood function, \( \max p(y|x) \), where \( p(y|x) \) is the conditional probability of having \( y \) given \( x \) [67]; the DR problem becomes a maximum likelihood estimation problem.

If the constraint \( g(x) \) is linear and is written as \( Ax = 0 \), the analytical solution to the above problem can be obtained as follows [89]:

\[
\hat{x} = y - W^{-1} A^T (AW^{-1}A^T)^{-1} Ay
\]

2.1.1.2 Gross Error

In the DR formulation using the WLS, the error is assumed to be random with zero mean. As well as the random error, the process may experience gross error. A gross error refers a systematic error in a process, for example, a mass leakage. A statistical definition of gross error is “an error whose occurrence as realization of a random variable is highly unlikely
Also, a gross error affects the mean value of measurements, and its magnitude is considered larger than that of random error.

Sensor measurements containing gross errors should be excluded in the DR because the gross error in a measurement affect on others that do not contain any gross error. Consequently, the measurements are reconciled to less accurate values. Gross Error Detection (GED) is a complementary procedure to detect any gross error in measurements before they are reconciled. Various statistical GED methods for a linear process are reviewed in [89].

2.1.1.3 Robust Estimation

GED seeks to detect any gross error in measurements and to exclude it from DR. Robust estimation, however, instead of excluding the gross error, attempts to diminish the influence of it in DR. Thus, DR using the robust estimation does not require an extra GED procedure.

Instead of the normal distribution assumption for random error, robust estimation uses other distributions such as the contaminated Gaussian distribution [120] and the Cauchy distribution. These distributions are reported to be less sensitive to a gross error [91].

2.1.2 Bayesian Networks for Data Validation

An alternative view on the problem of data reconciliation is possible by means of more structured and graphical representation of statistical relationships, such as Bayesian networks. A Bayesian networks (BN) or Bayesian belief network is a graphical probabilistic model originated from the field of Artificial Intelligence (AI) [23]. A BN is a directed, acyclic graph associated with conditional probabilities. The graph consists of nodes and edges. Each node represents a random variable or event. A directed edge connect a parent to a child if one probabilistically depends on the other. The probabilistic relationship is quantified by a conditional probability.

For data validation, a process or system producing the data can be modeled with a BN. The BN models for the two sensor models 1 and 2 are shown in Figure 8. Mehranbod et al. [82][81] applied the bias-augmented one in the sensor validation of a chemical process.

Relationships between measurement of various properties can be obtained from statistics such as a regression equation or from laws of nature such as the mass balance equation.
Bickmore [13] developed a Bayesian network using regression equations obtained from rocket engine data.

2.2 Fault Diagnosis

Thus far, we have focused on sensors and data that the sensors measure rather than the system we are trying to diagnose. From this point, we change our focus to the system on which the sensor is located. Fault diagnosis is a process that determines the condition of a system or components constituting the system based on data obtained from the system. Although the words, failure and fault, are used as synonyms in some literature, strictly speaking they are different. Whereas a failure typically means a complete breakdown, a fault means a change in the state of the system in question or a deterioration in function, which may still be tolerable [45].
Fault diagnosis consist of three tasks: detection, isolation, and identification. A generally accepted definition of each task is as follows [45]:

1. Detection: the indication that something is going wrong in the object to be diagnosed
2. Isolation: the determination of the location of the fault
3. Identification: the determination of the magnitude of the fault

The isolation task implicitly includes the detection task, and the identification task encompasses the other two. Three tasks may require different amount or types of data, and they provide different depths of diagnosis results. For example, whereas the detection and isolation tasks provide qualitative results, and the identification task results in a quantitative answer. Accordingly, different types of algorithms are used in each task.

2.2.1 Detection

Detection of a fault is the most primitive task among the three tasks. Two types of detection techniques has been widely used in various fields: limit checking and hypothesis testing.

2.2.1.1 Limit Checking

Limit checking technique monitors measurements, and when any of the measurements exceeds a preset limit, a warning or an alarm is raised. The control chart from the field of statistical process control, for example, is a tool for limit checking. It was invented by a physicist, Walter A. Shewhart, in the 1920s [104]. The fundamental idea of control chart is that when a process is normally operating, a measurement will vary in a random pattern within a control limit which typically set at $\pm 3\sigma$ from the mean value. If a sequence of measurements shows a regular pattern or goes beyond the control limit, it suggests that there may be a special cause. Shewhart set the $3\sigma$ limit based on an empirical study that, for most probability distributions, at least 99 percent of observations fall within $\pm 3\sigma$ from the mean. Figure 9 shows an example of control charts.
2.2.1.2 Hypothesis Testing

Hypothesis testing is a statistical procedure to make a rational decision. In a hypothesis test, a null hypothesis, $H_0$, and an alternative hypothesis, $H_1$, are posed. Given data, it is decided which hypothesis describes the data the best.

A likelihood ratio is often used as a criterion to that decision. Given the data $D$, the likelihood ratio is the ratio of conditional probabilities written as following:

$$\Lambda(D) = \frac{P(D|H_1)}{P(D|H_0)}$$  \hspace{1cm} (4)

The likelihood ratio is compared to a threshold in order to make a decision in favor of one hypothesis against the other. This logic can be easily extended to more than two hypotheses.

Thus far, the data is implicitly assumed to be a block of sample points. When a time-critical decision is desired to be made, the test can be performed as a new data point arrives. The number of data points required to make a decision varies with the performance specifications: the false-alarm probability and the detection probability [66]. This procedure is called Sequential Likelihood Ratio Testing (SLRT) or Statistical Probability Ratio Testing (SPRT). The technique was developed by a mathematician Abraham Wald in the 1940s [127]. It was used in the development of statistical signal detection theory. In the early 90s, Argonne National Laboratory developed a sensor fault detection tool for nuclear power

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{control_chart.png}
\caption{Example of Control Chart [104]}
\end{figure}
plants using the SPRT [48].

2.2.2 Isolation

Fault detection is a binary classification problem in which a decision to be made has only two choices: normal or abnormal. On the other hand, fault isolation has to pinpoint the origin of an abnormality if it presents. Fault isolation results must be one out of many possible sources of the abnormality. Thus, it is a multi-category classification problem. Many algorithms from statistics and AI are adopted to fault isolation.

2.2.2.1 Multivariate Statistical Methods

Principal Component Analysis (PCA) is a multivariate statistical method widely applied in process monitoring of chemical plants or power plants. It is well known as a dimensionality reduction technique. Consider multidimensional samples\(^1\) obtained from a process. A few principal components, i.e., eigenvectors, often explain most of the variability in the samples. When the samples are projected onto the principal components, the multidimensional samples can be visualized in two or three dimensional space. Each variable contributes on the total variability. The higher is the contribution of a variable, the more likely is the variable the source of an abnormality.

PCA determines the principal components that explain the variability of data the best. However, the principal components are not necessarily distinguish normal data from abnormal data well because they analyze the normal data and the abnormal data together [6]. Fisher Discriminant Analysis (FDA) is another multivariate method that seeks for the direction optimally classifying data points into different categories. It is reported that FDA has better performance in fault isolation than PCA [24][54]. FDA, however, requires a pre-processing of data. In order for FDA to isolate a fault properly, the samples have to be sorted out and grouped into clusters using, for example, PCA or other cluster analysis methods. Then, each cluster is compared individually to the cluster obtained from a normal operation. The comparison determines what makes an abnormal cluster apart from the normal one.

\(^1\)The word “sample” is often used in a singular form to indicate multiple points as a group. However, the author will use it in a plural form as well.
2.2.2.2 Artificial Intelligence Techniques

Many classification methods have been researched in the field of Artificial Intelligence (AI), and some of them are adapted to fault isolation. To name four, there are expert systems, Case-Base Reasoning (CBR), Artificial Neural Networks (ANNs), and Bayesian networks (BNs).

Expert systems emulate how human experts work in a computer model or program. An expert system consists of a knowledge base and an inference engine. The knowledge base is a storage of factual and heuristic knowledge that the human experts may use in their problem solving process. The knowledge is represented by a set of “if-then” rules. The inference engine manipulates the knowledge to perform reasoning, i.e., to find a cause for a symptom.

As the size of system to be modeled increases, the number of rules required is increases. The large knowledge base increases the chance that some of rules conflict to others. Expert systems have been applied to various complex applications such as space shuttle main engine [4], gas turbine engines [112], automobile engines [42], and electric power systems [86], etc.

Case-Based Reasoning (CBR) is a recent approach to solve a problem and to acquire knowledge. It finds the most similar problem from previously solved ones stored in a case base, and adapts the solution of the most similar one. It also stores the newly solved problem in the case base so that the knowledge grows over time. General Electric (GE) has developed CBR systems for medical equipments [27], locomotives [123] and gas turbine diagnostics [29].

Artificial Neural Networks (ANNs) were inspired by biologic nerve systems. They are used to model complex phenomena or to solve problems such as pattern recognition in the computer science field. A neural network consists of interconnected artificial neurons which cooperatively produce output. In diagnosis, the network is trained to map the nonlinear relationship between health parameters and measurable variables. The training data are usually generated from computer simulations. When an operating condition of the object to be diagnosed affects significantly on the nonlinear relationship, the network has to be retrained as the operating condition changes for better accuracy. Various types of ANNs have been applied in many different systems such as gas turbine engines [136], rocket engines [132], helicopter rotor systems [40], aircraft icing problems [100], and industrial power plants.
The Bayesian network (BN) technique described in §2.1.2 also have been applied to fault diagnosis of complex systems as well. The variables describing a system to be diagnosed and their interdependency are obtained from mathematical equations or expert knowledge. In diagnostic mode, the observables, e.g., sensor measurements, are entered as evidence, then the probability distributions of others are updated. Since the 1990’s, the BN technique has been applied to fault diagnosis for automobiles, rocket engines, and gas turbine engines [101, 12, 96].

2.2.3 Identification

Whereas detection and isolation are qualitative tasks, fault identification is a quantitative task. Once a fault is detected and isolated, it is desired to know how severe the fault is. The severity of fault is quantified with the deviation of estimated health parameters from the baseline values obtained from the normal condition. Because it is impossible to physically measure the health parameters, they must be estimated from measurable quantities. The measurable quantities of a system are functions of the health parameters and the operating conditions. As described in §2.1, when the measurable quantities are measured by sensors, random errors are added to the measurements. Furthermore, if the sensor is faulty, the measurement contains a systematic error, i.e., bias. The relationship can be written in a vector form:

\[ y = f(u, o) + b + \varepsilon \]  

(5)

where \( u \) is a vector of health parameters, \( o \) a vector of operating condition parameters, \( b \) a vector of bias, \( \varepsilon \) a vector of random error. The objective of identification is to estimate unknown \( u \) from known \( y \). Therefore, it poses an inverse problem.

Mathematical or physics-based models often approximate the function \( f \). Parameter estimation techniques are used to find the unobservable parameters \( u \) and \( b \) that drive the difference between the model prediction \( f \) and the measurement \( y \), which is called residual, to zero. Various parameter estimation techniques have been applied to fault diagnosis of linear and nonlinear systems.
2.2.3.1 Linear Systems

Although the function $f$ for some complex systems is nonlinear, it can be linearized with acceptable loss of accuracy. For linear systems, the WLS method described in §2.1.1.1 and the Kalman filter technique have been widely used.

The WLS method is introduced in the field of gas turbine engine diagnosis by Urban in 1972 [122]. Urban developed a program called GPA (Gas Path Analysis) using the WLS to adjust component health parameters (efficiency and flow) to match sensor measurements. In late 1970s, GE Aircraft Engines developed a similar program called TEMPER [31].

The Kalman filters are recursive filters that estimate the state of dynamic systems from a series of noisy measurements. A Kalman filter consists of two phases: predict and update. In the predict phase, the current state is predicted from the state at the previous time step. In the update phase, the measurement at the current time step is used to correct the predicted state.

In the 1980s, Rolls Royce developed COMPASS (COndition Monitoring and Performance Analysis Software System) using the Kalman filtering technique to estimate turbofan engine health parameters and sensor bias from measurements [93]. The basic Kalman filtering technique and its variations have extensively been researched in the field of gas turbine engine diagnosis [109, 108].

2.2.3.2 Nonlinear Systems

There has been an approach in which fault diagnosis is treated as an optimization problem. In this approach, a numerical optimizer finds health parameters and sensor biases that make the difference between estimation from a mathematical model and measurements negligible.
The nomenclature in the figure is equivalent to that of Equation 5.

This approach is known to be computationally more intensive than the classical estimation techniques such as WLS and Kalman filtering. Zedda and Singh [137] investigated the use of Genetic Algorithm (GA) in gas turbine diagnosis. Later, Sampath and Singh [98] developed a hybrid approach of ANN and GA to reduce the computational burden on the GA. In their hybrid approach, several ANNs perform heuristic classification and reduce the size of the problem that the GA has to consider. Then, the GA is applied to the reduced problem, and estimates health parameters and sensor biases.

2.2.4 Complicating Factors for the Existing Algorithms

In the field of fault diagnosis, several methods are borrowed from mathematics and computer science. Each method has its own strength and weakness, but they share the common issues: the smearing effect, the isolability, and the sensitivity.

First, “smearing” refers to the spread in diagnosis over multiple causes, instead of the diagnosis pinpointing one, and subsequent inconclusiveness of results. The inconclusiveness makes the user harder to interpret diagnosis results and make a diagnostic decision. The smearing effect is typical especially in the results from the Weighted Least Squares method [31].

Second, the isolability is the ability of a diagnosis algorithm to distinguish a certain fault among many others [45]. It becomes a hard task when more than one faults produce similar signature such as in Figure 12, which is supposed to be a distinctive characteristic of a fault. A fault diagnosis algorithm has to either tiebreak them in a logical way or report all of the potential solutions and confidence associated with them.

Third, the sensitivity is the degree of susceptibility to fault signature. Some faults may pronounce themselves, e.g., large magnitude signature, but others may not, e.g., small magnitude signature as shown in Figure 13. With the presence of uncertainty, the small magnitude signature may easily be concealed in measurement noises [45][32].

The smearing effect is the deficiency that prevails among the existing algorithms, but the isolability and the sensitivity are merely the properties that strongly affect on the quality
Figure 11: The Smearing Effect
Figure 12: Similarity in Signature Pattern Causing the Isolability Issue

Figure 13: Difference in Signature Magnitude Causing the Sensitivity Issue
of algorithm. Any newly developed algorithm has to be examined on these issues, and its limitations have to be provided to the user.

2.3 Reasoning under Uncertainty

In real world problems, fault diagnosis algorithms have to make decisions based on uncertain data or uncertain models. For this reason, several methods have been proposed in the field of AI for dealing with the uncertainty and reasoning with different kinds of uncertainty. This section reviews some of the reasoning methods.

Among the reasoning methods, it is the probability theory that has the longest history. According to the probability theory, the problem domain of interest can be represented probabilistically by a joint distribution over all variables involved to the domain. Although it is well understood and mathematically sound, the complete specification of the joint probability, however, requires unrealistically large number of probabilities. The number of required probabilities grows exponentially with the number of variables involved.

Instead of joint probabilities, conditional probabilities are the basic expressions of the problem domain of interest in the Bayesian formalization of probability theory. In Bayesian probability theory, a proposition probabilistically depends only on the information relevant to the proposition. The relevance is represented by a conditional probability. We are ultimately interested in posterior distributions that are conditional probabilities “logically later in the particular chain of inference being made [60].” The posterior probability on a hypothesis \( H \) given a set of relevant evidence \( E \) can be calculated with the Bayes rule:

\[
P(H|E) = \frac{P(E|H)P(H)}{P(E)}
\]

The posterior \( P(H|E) \) answers to the query: “Given that I know \( E \), what is my belief in \( H \)?”

A Bayesian network (BN) is the formalization of the Bayesian probability theory based on the probabilistic conditional independence assumption [51]. In the simple BN structures in 14, the nodes A and C are probabilistically independent given B. The independence assumption converts a global joint probability into a function of local conditional probabilities, and consequently, the number of probabilities required for the complete specification
reduces, although its computation is still complex.

The Dempster-Shafer theory adds the concept of “ignorance” in the probability theory [103]. According to the theory, when the evidence $E$ is unreliable with a certain probability $p$, the degree of belief on the claim that $H$ is false has zero degree of belief instead of $(1-p)$ degree of belief, and the zero degree of belief does not mean the hypothesis $H$ is false. It merely means that the unreliable evidence $E$ gives no reason to believe that the hypothesis $H$ is false. Although it may represent actual word more precisely, the Dempster-Shafer theory is mathematically more complex than the Bayesian probability theory.

The certainty factor formalism is an approach that was intended to use with rule-based expert systems [73]. In order to accommodate uncertainty, a certainty factor is associated to each rule. The certainty factor of a rule “if $E$ then $H'$, $CF(H, E)$, is the degree of belief to which the evidence $E$ confirms the hypothesis $H$. When $E$ increases the probability of $H$, i.e., $P(H|E) > P(H)$, then $0 < CF(H, E) \leq 1$; the belief on the proposition increases. On the other hand, when $E$ contributes against $H$, i.e., $P(H|E) < P(H)$, then $-1 \leq CF(H, E) < 0$; the belief on the proposition decreases. The certainty factor of 1 means the complete certainty that the proposition is true, while the certainty factor of -1 the complete certainty that the proposition is false. The certainty factor of zero means no change in belief.

It is computationally more efficient than the probability theory, which requires full specification of probabilities. Despite of its successfully application in MYCIN [106], the certainty
factor formalism has been criticized for its mathematical inconsistency [30].

Fuzzy logic aims at modeling the imprecise reasoning that is crucial in the human ability to make decisions with uncertain environment and imprecise information. It is a multivalued logic in which everything is a matter of degree. For example, “Age” is the linguistic variable whose values are “young,” “old,” and “not young” and so on. In fuzzy logic, the variable “Age” can be in multiple states with different degrees, as shown in Figure 15. The fuzzy inputs are fed to fuzzy inference engines. The fuzzy outputs from the inference engines fed to defuzzifiers that convert the fuzzy outputs to crisp outputs. Although it has been used in control applications, there are few publications regarding to the use of fuzzy logic for reasoning under uncertainty in the real world [35].

Figure 15: Linguistic Values of “Age” [135]
In the previous chapter fault diagnosis and its three sub-tasks are defined. In addition various techniques used for the sub-tasks are surveyed. Among the three sub-tasks fault identification is the task that plays an important role in deciding a further action for either operation or maintenance. For example, a power plant owner may want to operate his gas turbine to make a bigger profit even if a minor fault occurs in it. In this case the power plant owner must know whether the fault is tolerable or an immediate action is required.

Fault identification is a quantitative task in which an estimator determines unknown or unmeasurable quantities of a system from known and measurable data. The direction of information in this task is from the output to the input of the system, which is the reverse of how the system works or how we often model the system. Fault identification is an inverse problem.

This chapter, first, explains mathematical characteristics of inverse problems. Then, it introduces a traditional inverse solver, the method of least squares, and regularization, a technique that is conceived to overcome adverse characteristics of inverse problems, and statistical inversion theory, which convert a estimation problem to a probabilistic inference problem. Finally, it provides a survey of several relevant topics in Bayesian statistics on which statistical inversion theory is based.

3.1 Characteristics of Inverse Problems

Tarantola [113] generalizes a scientific procedure for the study of a physical system with the following three steps:

1. Parameterization of the system
2. Forward modeling
3. Inverse modeling
First, the parameterization of the system is to discover the minimal set of parameters that completely characterizes the system. Such a set is called model parameters. Second, forward modeling is to discover a physical law that allows us to predict the outcome of the system given the model parameters. Third, inverse modeling is to estimate the model parameters when the outcome of the system is observed.

Problems that involve the forward and inverse modeling are referred to as direct and inverse problems, respectively. Keller [71] provides a definition of direct and inverse problems in a historical point of view. Keller defines two problems as direct and inverse problems if the formulation of one involves the solution of the other. Among the two problems the direct problem is the one that has been studied extensively than the other while the inverse problem is the one that is less studied or understood than the other. On the other hand, Bertero’s definition [10] is based on causal relationships. A direct problem is formulated based on a physical law specifying a cause-effect consequence. The corresponding inverse problem is to find the unknown cause of known effect. Hansen’s point of view is more or less similar to that of Bertero. Hansens [52] describes that inverse problems involve finding the internal structure of a system from the observed behavior of the system or determining the unknown input of the system from the known output.

Direct problems are perceived to be much easier than inverse problems, due to the following two properties: locality and causality. Laws of nature are often expressed as a system of algebraic or differential equations. The equations are local in a sense that they express the dependency of the function describing a system and its derivatives on the outcome of the system at a given point, i.e., at a given model parameters. They are causal in a sense that the outcome depends on the model parameters. On the contrary, inverse problems are often not local and/or not causal [68]. Bertero and Boccacci [11] argue that the conceptual difficulty associated with inverse problems due to a loss of information. A forward modeling always involves a loss of information or an increase in entropy. Consequently, an inverse modeling of the same system becomes different from the forward modeling, and the inverse problem requires the recovery of the lost information. The argument of Bertero and Boccacci is an analogy of forward and inverse modeling to an irreversible thermodynamic
The conceptual difficulty of inverse problems imposes adverse characteristics on inverse problems. According to [50], a problem is well-posed if the following conditions are met:

1. a solution exists,
2. the solution is unique, and
3. the solution depends continuously on the data.

### 3.2 Solving Inverse Problems

#### 3.2.1 The Method of Least Squares

Consider a system in a steady-state, which follows a functional relationship $f$ between its input $\mathbf{x}$ and output $\mathbf{y}$ as follows:

$$
\mathbf{y} = f(\mathbf{x}) + \mathbf{\varepsilon}
$$

(7)

$\mathbf{x}$ is a set of unknown parameters that describe the system. Let us call it the state vector hereafter. $\mathbf{y}$ is a set of observable properties, often measured with sensors. Let us call it the observables. $\mathbf{\varepsilon}$ is the sensor random noise. The random noise is generally assumed to follow a Gaussian distribution with zero mean and a constant variance. The relationship can be approximated to a linear form and written in the matrix form as

$$
\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{\varepsilon}
$$

(8)

where, unlike in Equation 7, $\mathbf{\varepsilon}$ now includes not only the sensor random noise but also the linearization error. The simplest, but not necessarily the most effective, way to solve Equation 8 is to invert the matrix $\mathbf{A}$:

$$
\hat{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{y}
$$

(9)

However, Equation 9 is valid only if $\mathbf{A}$ is invertible, i.e., $\mathbf{A}$ is a nonsingular square matrix. Instead of inverting the matrix $\mathbf{A}$, the method of least squares [110] is commonly used to solve the system of linear equations. The method of least squares finds a solution $\mathbf{x}$ that minimizes

$$
J = (\mathbf{y} - \mathbf{A}\mathbf{x})^T(\mathbf{y} - \mathbf{A}\mathbf{x})
$$

(10)
The objective function is called the misfit of data. In the method of least squares each measurement, each element in the vector \( y \), has same impact on the objective function no matter how accurate each sensor is. In reality each sensor has different precision, and it may be reasonable to weigh precise measurements more than imprecise ones. The method of weighted least squares finds \( x \) that minimizes

\[
J = (y - Ax)^T R^{-1} (y - Ax)
\]

where \( R \) is the covariance matrix of measurement random error, \( \varepsilon \). The inverse of \( R \) is called the precision matrix, and it weighs each measurement differently on the objective function \( J \). The method of weighed least squares is equivalent to maximum likelihood estimation (MLE) if the random error \( \varepsilon \) follows a Gaussian distribution and the measurements at different instants of time are independent to each other [47]. The method of least squares and the method of weighted least squares have their own limitations. First, they are incapable of handling multiple solutions. Second, they are sensitive to errors in the observables \( y \) because the residual \( r = y - Ax \) is squared in the objective function.

### 3.2.2 Sensitivity of a Linear System of Equations

For linear problems a common measure of ill-conditioning is the condition number [126]. Consider the linear system in Equation 8 again. The condition number of the matrix \( A \) is the ratio of the largest singular value to the smallest one [46]. As the condition number increases, the sensitivity increases. When the condition number is too large, the matrix \( A \) is called ill-conditioned.

The condition number can be interpreted geometrically. Let the matrix \( A \) be a \( 2 \times 2 \) matrix, and consider a circle in the two dimensional space. The condition number of the matrix \( A \) is the axis ratio when the circle is transformed by the matrix \( A \). The axis ratio tells that how flat the circle becomes by the transformation [79]. If the condition number is large, the sphere will become an elongated ellipse as depicted in Figure 16.

Another geometric interpretation of condition numbers is given below. Consider a system
of two linear equations $Ax = b$ as defined by

\begin{align*}
a_{11}x_1 + a_{12}x_2 &= b_1 \\
a_{21}x_1 + a_{22}x_2 &= b_2
\end{align*}

(12)

The two equations are lines in the $(x_1, x_2)$ plane. The solution of the system is the intersection point of the two lines. If the matrix $A$ is ill-conditioned, that is, the condition number of the matrix $A$ is large, the slopes of the two lines are nearly same as shown in Figure 17. Consequently, a small perturbation in either $b_1$ or $b_2$ causes a large shift of the solution in the $(x_1, x_2)$ plane. Solutions of an ill-conditioned system of linear equations change drastically with any error contained in data, modeling, and computation.

### 3.2.3 Regularization Techniques

A classical approach for solving ill-conditioned inverse problems is regularization. Regularization techniques seek an approximate solution that is stable by adding a supplementary constraint on $x$. Among regularization techniques Tikhonov regularization is the most widely known. Tikhonov regularization [119] finds the solution $x$ that minimizes

$$J = \|y - Ax\|^2 + \lambda G(x)$$

(13)

where $G$ is a nonnegative function of $x$. A simple form of the objective function $J$ is

$$J = (y - Ax)^T(y - Ax) + \lambda x^T x$$

(14)
where $\lambda$ is a positive constant, called the regularization parameter. The objective function $J$ is the sum of the misfit of data and the constraint that is the square of the 2-norm of $x$. The regularization parameter controls the weight of the constraint in the objective function. Ridge regression [57] is conceptually equivalent to Tikhonov regularization. Lasso regression [117] uses the 1-norm of $x$ as an ad hoc constraint:

$$J = (y - Ax)^T(y - Ax) + \lambda \|x\|_1$$

(15)

Singular value decomposition (SVD) is a popular method to solve underdetermined systems of linear equations. Like the simple Tikhonov regularization SVD uses the misfit of data and the 2-norm of a solution. Underdetermined systems have infinite number of solutions that minimizes the misfit of data. Among these solutions the minimum 2-norm solution is always unique [14]. SVD finds the unique solution. However, the minimum norm solution is not necessarily meaningful in practical problems.

The methods surveyed thus far transform an inverse problem to an optimization problem. As a consequence, a solution from these methods is a point estimate. These methods provide neither a confidence measure in the point estimate nor any information on the solution space.

Figure 17: Example of Ill-Conditioned Systems (recreated from [130])
that might contain alternate solutions. Although some of these methods also have their equivalent counterparts in statistics, a rigorous statistical approach is desirable to solve inverse problems under many sources of uncertainty.

3.2.4 Statistical Inversion Theory

Statistical inversion theory recasts an estimation problem to a statistical inference problem. Each variable is treated as a random variable in this approach. The forward mapping from the unknowns $x$ to the observable $y$ is probabilistically modeled from Equation 16. The mapping is represented by the conditional probability $p(y|x)$. When some observation $y$ is made, the probability of the unknown $x$ can be updated. The updated probability $p(x|y)$ is called a posterior probability, which is of primary interest in statistical inversion theory. Bayes rule links the conditional probability and the posterior probability as the following:

$$p(x|y) \propto p(y|x)p(x)$$ (16)

The prior probability of $x$, $p(x)$, represents a prior knowledge on $x$ before the arrival of new information $y$. Once the posterior distribution is known, several point estimates can be obtained from the posterior.

Most popular statistical estimates are conditional mean and maximum a posterior (MAP) [68]:

$$x_{CM} = \int_x xp(x|y)dx$$ (17)

$$x_{MAP} = \arg \max_{x \in X} p(x|y)$$ (18)

Nonetheless, these point estimates reveal only a part of the information that the posterior distribution contains so that they can be misleading. Two examples of misleading point estimates are given in Figure 18. The posterior probability of $x$ in these examples is multimodal. In Figure 18(a) the MAP estimate is the higher point between the two modes; however, the conditional mean does not indicate either mode. The conditional mean is between the two modes, where the probability density is almost zero. It is highly unlikely that $x$ is any value near the conditional mean. On the other hand, the conditional mean in Figure 18(b) indicates one of the mode, while the MAP estimate is the higher mode between the
two. However, the probability that \( x \) is near the MAP estimate is negligible. The area under that portion of the curve is almost zero; therefore, it is a bad estimate of \( x \). Bayesian statistics is primarily interested in probability distributions instead of point estimates.

Equations 17 and 18 can be simplified with additional assumptions. When \( x \) and \( \varepsilon \) are assumed to follow normal distributions with covariance matrices \( P \) and \( R \), respectively, the conditional mean equals to the MAP estimate, which can be obtained by minimizing

\[
J = (x - \mu_x)^T P^{-1} (x - \mu_x) + (y - Ax)^T R^{-1} (y - Ax)
\]  

(19)

where \( \mu_x \) is the mean of the prior distribution of \( x \) [110]. The solution should not only match the data but also be close to the prior mean. The solution is called a minimum variance Bayes estimate. When only vague knowledge of \( x \) is available in prior, the vague knowledge is represented by large variance and small precision. In the extreme case in which \( P^{-1} \) is
negligible, the objective function is equivalent to that of the method of least squares. When
$P$ and $R$ are identical matrices and $\mu_x$ is zero, the objective function equals to that of the
simple form of Tikhonov regularization.

### 3.3 Bayesian Statistics

Statistical inversion theory is based on Bayes rule: a posterior probability is proportional to
a likelihood multiplied by a prior probability. Thus, Bayesian statistics involves assigning
prior probabilities to variables, constructing likelihood functions, and computing posterior
probabilities.

Because Bayesian statistics use prior probabilities, it results contains a certain degree of
subjectivity. Historically, this subjectivity has drawn the criticism about Bayesian statistics
from the frequentist [21]. Furthermore, computing posterior probabilities is so burdensome
that the practicality of Bayesian statistics for large scale problems is debatable even nowa-
days. However, Bayesian statistics has been gaining a great deal of attention since several
efficient algorithms are conceived to compute posteriors analytically and numerous sampling
techniques are developed to compute posteriors approximately in the late 1980s [92, 75].

#### 3.3.1 Prior distributions

Bayesian statistics uses prior information explicitly, and that distinguishes Bayesian statis-
tics from other statistical methods. In Bayesian statistics uncertain belief is represented by
a probability distribution. A prior distribution is a probability on the information before
any data is collected. In parametric approaches at least a type of the prior distribution
should be specified beforehand, and sometimes the parameters that define the distribution
should be specified as well, for example, the mean and variance of a normal distribution.
The prior probability can be determined from historical data or the subjective knowledge of
experts.

There have been efforts to find formal rules of determining prior distributions. The
fundamental idea is originated from Jeffreys [70]. Jeffreys believes that there exists “the
initial stage of knowledge”. When a hypothesis is considered, Jeffreys describes, a person in
this stage has no opinion about whether the hypothesis is true or not. If there is no sufficient
reason to choose one hypothesis over another, the probabilities of the two hypotheses should be equal [63]. This is an application of the principle of insufficient reason [99]. Jeffreys applied the principle of insufficient reason to estimation problems as well. Suppose that a parameter is to be estimated and that there is insufficient reason to prefer a particular value over others. In this case it is objective that all possible values of the parameter have the equal probability. Based on the above argument, in 1946, Jeffreys proposed the celebrated Jeffreys’ prior [62]. Suppose that \( p_\theta(y) \) is a probability density function of \( Y \), which depends on the parameter \( \theta \). Jeffreys’ prior for \( \theta \) is

\[
p(\theta) \propto |I_\theta|^{1/2}
\]  

(20)

where \( I_\theta \) is the Fisher information. The Fisher information is defined by

\[
I_\theta = E_\theta(s_\theta^2)
\]  

(21)

where \( s_\theta = \partial \log p_\theta(y) / \partial \theta \) and \( E_\theta \) refers to expectation with respect to \( p_\theta \) [129]. Jeffreys’ prior was followed by several variants over the decades, and they are reviewed in [70]. Priors determined by these rules are referred to as reference priors.

Reference priors are often improper, which means it is not integrated to one, although Jeffreys did not consider this to raise a fundamental difficulty. An alternative of a reference prior is a diffuse proper prior. Diffuse proper priors are locally uniform and integrated to one [17]. For example, a normal distribution with a large variance is a diffuse proper prior. Although diffuse proper priors work fine in many problems, Kass and Wasserman [69] point out that they do not fundamentally resolve difficulties arising from the use of improper reference priors.

Box and Tiao [17] discuss that when a large amount of data is available, a posterior is dominated by the likelihood function so that adverse effects caused by a reference prior will be diminished. However, it is often difficult to know whether a posterior is data dominated or not [70].
3.3.2 Computation of Posterior Distributions

In Bayesian statistics posterior distributions are the ultimate solution. Calculating posterior distributions involves integrations. For example, the denominator of Bayes rule is a marginal distribution, which requires an integration. These integrations are often analytically intractable and multi-dimensional. In practice asymptotic approximation methods or numerical integration using sampling are often used.

3.3.2.1 Asymptotic Approximate Integration

When the amount of data is large, the likelihood overpowers the prior. As a consequence, the prior has little effect on the posterior. When the likelihood is peaked and the prior is reasonably flat, the posterior of a random variable $y$ can be approximated with a normal distribution $N(\mu, \Sigma)$ where the mean $\mu$ is the mode of the posterior distribution $p(\theta|y)$ and the covariance matrix $\Sigma$ is the inverse of the Fisher information matrix [21]. The Fisher information matrix [28] is written as

$$I_{ij}(y) = -\left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \{p(y|\theta)p(\theta)\} \right]_{\theta=\hat{\theta}}^\prime$$

(22)

When the prior $p(\theta)$ is reasonably flat, it can be ignored.

The normal approximate of a posterior distribution is inaccurate when the posterior distribution differs significantly from a normal distribution. An alternative asymptotic technique is Laplace’s method [74]. Suppose that we wish to approximate an integration

$$I = \int_\Theta L(\theta)p(\theta)d\theta$$

(23)

where $L(\theta)$ is a likelihood function. The integration can be approximate as

$$I \approx L(\hat{\theta})p(\hat{\theta})(2\pi)^{d/2}|H|^{1/2}$$

(24)

where $\hat{\theta}$ is the mode of the posterior $p(\theta|y)$ and $H$ is the matrix of second derivatives of the log-posterior evaluated at $\hat{\theta}$ [118].
3.3.2.2 Monte Carlo Methods

Let us consider a random variable $x$ following a probability distribution $p(x)$. The expectation of a function $\phi(x)$ is written as

$$E[\phi(x)] = \int \phi(x)p(x)dx \tag{25}$$

When $\{x_r\}_{r=1}^R$ is a set of sample drawn from $p(x)$, the integration can be approximated using the sample as follows:

$$\int \phi(x)p(x)dx \approx \frac{1}{R} \sum_{r=1}^{R} \phi(x_r) \tag{26}$$

As $R$ approaches to $\infty$, the approximate becomes close to the exact value.

When $x$ is high-dimensional, it is hard to draw samples directly from $p(x)$ [76]. There are several techniques that draw approximate samples of $p(x)$ from a simpler distribution. In these techniques $p(x)$ is referred to as the target density from which we ultimately wish to draw samples. the simpler density is referred to as the proposal density. These techniques are extremely useful in Bayesian statistics. For example, let us suppose that $p(x)$ is a posterior distribution. We can draw samples from $p(x)$ without calculating it.

Rejection sampling [125] is a general and widely used method among the sampling techniques. For simplicity it is explained with a univariate case. Suppose we have a proposal density $q(x)$ that satisfies the following condition:

$$cq(x) > p(x) \text{ for all } x \tag{27}$$

where $c$ is a constant. The procedure of rejection sampling is as follows:

1. Sample $x$ from $q(x)$ and $u$ from $U(0,1)$.

2. If $u$ is greater than $\frac{p(x)}{cq(x)}$, accept $x$. Otherwise, reject $x$. Repeat the procedure.

The accepted samples are independent samples from the target density $p(x)$.

Figure 19 shows how rejection sampling works graphically. A point $(x, u)$ is generated randomly from the proposal density and the uniform distribution, respectively. If the point is between the two curves $cq(x)$ and $p(x)$, it is rejected. Or if the point is below the curve $p(x)$, it is accepted.
Rejection sampling works well if the proposal density is a good approximate of the target density. If the proposal density is quite different from the target density, the constant \( c \) should be large. Consequently, the frequency of rejection becomes large as well. It may be difficult to find a proposal density that is acceptable approximate of the target density but still easy to sample from.

In contrast, Markov Chain Monte Carlo (MCMC) methods do not require a proposal density looking similar to the target density. The MCMC methods involve a Markov process in which a sequence of state \( x \) is generated iteratively. Therefore, unlike rejection sampling, they have the concept of convergence.

The Metropolis-Hastings algorithm \([84, 53]\) is a well known MCMC method. It uses the proposal density \( q(x'; x^t) \) where \( x' \) is a tentative new state, and \( x^t \) the current state. After a tentative new state \( x' \) is generated, it is decided whether to accept the tentative new state or not according to acceptance rate \( a \):

\[
a = \frac{p(x') \cdot q(x^t; x')} {p(x^t) \cdot q(x'; x^t)} \tag{28}
\]

If \( a \) is greater or equal to one, the tentative new state is accepted. Otherwise, the tentative new state is accepted with probability \( a \). After sufficient many iterations the state can be
thought of a sample from the target density $p(x)$.

Gibbs sampling is a special case of the Metropolis-Hastings algorithm, which every proposal is accepted [76]. For a system with two variables Gibbs sampling updates the state of the two variables as the following sequence:

$$
x_{1}^{t+1} \sim p(x_{1}^{t}|x_{2}^{t})
$$

$$
x_{2}^{t+1} \sim p(x_{2}^{t}|x_{1}^{t+1})
$$

The state is updated one variable at a time using the corresponding conditional density. The state updating sequence is graphically depicted in Figure 20. After enough number of updating the state is thought of a sample from the posterior distributions. As many sequences should be created as the desired number of samples.
3.4 Model Uncertainty

3.4.1 Model Selection

“In my mind the biggest challenge is to avoid unnecessary complexity and gratuitous engineering. My personal mission is to make things as simple as possible.”

— Andreas Bechtolsheim, a Sun Microsystems co-founder

Finding a form of relationship between responses and predictors has been one of fundamental problems in statistics. The true functional form is always unknown in real world problems. A viable alternative to the true function is an approximating function, called model. Numerous models can possibly approximate the true function equally well. The objective of the use of statistics in this matter is to search for appropriate models among many competing models.

What makes a model more appropriate or better than others? Suppose we are about to build a model based on some data. A criterion often used is how well the model fits the data. However, the model that fits the data the most is not always the best one. One can build a model that passes all the data points merely by adding a large number of terms, or basis functions. A large number of basis functions allow us to fit the data better but not necessarily to predict well. The model may horribly fit the region between the data points. This problem is called overfitting in statistics. Simply adding more basis functions does not lead us closer to the true function.

Occam’s razor, sometimes spelled Ockham’s razor, refers a philosophical notion that “an explanation of the fact should be no more complicated than necessary.” [61] The notion is often referred to as the principle of parsimony [18]. In the context of model comparison, the principle of parsimony means that a simpler theory or hypothesis, or a subsequent model from such a theory or hypothesis should be favored over a more complex one when all other things are same. It is a notion that the complexity of a model should be taken into consideration as well as how well the model fits data. This philosophical notion inspired several model selection criteria.

The likelihood function can be used to compare competing models [34]. Consider two models with an unknown parameter vector $\theta$. The maximum likelihood estimator finds
that maximizes \( p(D|\theta) \), given data \( D \). The larger is the maximum likelihood function \( p(D|\hat{\theta}) \), the better is the model according to the maximum likelihood criterion. However, the maximum likelihood criterion measures only how well a model fit data so that it favors a complex model over simpler models.

Akaike [2] found that maximum log likelihood is “biased upward” as an estimator of the target model. He also found that under certain conditions the bias is approximately equal to the number of parameters in the model. Akaike introduced a correction term to the maximum log likelihood and proposed an criterion called An Information Criterion (AIC):

\[
AIC = -2 \ln(L(\hat{\theta}|D)) + 2p
\]  

(30)

where \( L \) is the likelihood function, \( \hat{\theta} \) the maximum likelihood estimate, and \( p \) is the number of parameters. A model yielding the minimum AIC is favored over others.

In addition to the number of parameters, Schwarz [102] suggests a criterion that reflects the effect of sample size. The Schwarz criterion is to choose the model that maximizes

\[
S = \ln(L(\hat{\theta}|D)) - \frac{1}{2}p \ln n
\]  

(31)

where \( n \) is the sample size.

Bayesian information criterion (BIC) is Schwarz criterion multiplied by minus 2 [69]:

\[
BIC = -2 \ln(L(\hat{\theta}|D)) + p \ln(n)
\]  

(32)

When eight or more data points are available, BIC favors a simple model over more complex ones. The discrepancy between AIC and BIC becomes remarkable as the number of data points increases.

Bayesian model comparison uses posterior probabilities as a criterion. Let us consider two models \( M_0 \) and \( M_1 \), based on a null hypothesis \( H_0 \) and an alternative \( H_1 \), respectively, with unknown parameter vector \( \theta \). The posterior probability of \( M_i \) is

\[
P(M_i|D) = \frac{p(D|M_i)p(M_i)}{p(D)} = \frac{\int_\Theta p(D|\theta,M_i)p(\theta|M_i)d\theta P(M_i)}{p(D)}
\]  

(33)

Between the two models the model with a higher posterior probability is favored.
The ratio of the posterior distributions of the two models is written as

$$\frac{P(M_1|D)}{P(M_0|D)} = \frac{p(D|M_1)}{p(D|M_0)} \frac{P(M_1)}{P(M_0)} \tag{34}$$

The likelihood ratio $p(D|M_1)/p(D|M_0)$ is defined as Bayes factor [69]:

$$B_{10} = \frac{p(D|M_1)}{p(D|M_0)} \tag{35}$$

When the two models are equally probably before any data is collected, in other words, $P(M_1) = p(M_0)$, the Bayes factor $B_{10}$ is equal to the posterior ratio in favor of $M_1$. If $B_{10}$ is greater than one, it means that $M_1$ is relatively more plausible in light of the data $D$. On the other hand, if $B_{10}$ is less than one, it means that $M_0$ is relatively more plausible.

Thus far we were concerned if Bayes factor is greater than one or not. How much greater or lesser than one was not our concern. A few researchers attempted to translate the magnitude of Bayes factor to a qualitative scaling system. Jeffreys [63] suggests the following interpretation of Bayes factor:

Instead of the use of 10 base logarithm, Kass and Raftery [69] suggest the use of twice the natural logarithm of Bayes factor and provide their own interpretation similar to that of Jeffreys.

Jaynes claims that base 10 logarithms give immediate and intuitive meaning to human being [60]. Indeed, decibel is easy for engineers in certain fields to understand.
3.4.2 Bayesian Model Averaging

“... ask a hundred people to answer a question or solve a problem, and the average answer will often be at least as good as the answer of the smartest member.”

— James Surowiecki in “The Wisdom of Crowds”

The previous section several criteria for comparing models are surveyed. The analysis may select a model based on one of the criteria among a pool of competing models. Would one of the models in the pool be absolutely correct? In regarding to this question there are two perspectives: $M$-closed and $M$-open perspectives [9]. Suppose that there exists a number of competing models for some data. People who have the $M$-closed perspective suppose that at least one of the competing models is absolutely correct; the model describe how the data is generated. Let us call the model as the true model hereafter. In contrast, people with the $M$-open perspective suppose that none of the models is absolutely correct. They just hope that some of them are, at least, close to the truth.

Let us take the $M$-closed perspective just for now. The ultimate goal of modeling is not only to find a functional form of the model but also to estimate parameters associated with the model and to use the model in the prediction of data. What happens if a wrong model is selected instead of the true model? The risk associated with model selection is due to the uncertainty whether the true model is selected. To take the model uncertainty into consideration, model averaging finds all competing models or some promising ones and averages their result instead of selecting the “best” model. A Bayesian approach to model averaging is to average the results from the competing model according to how likely each model is.

Suppose that $\theta$ is a vector of unknown parameters, which is of main interest. Given data $D$ the posterior probability of $\theta$ is

$$p(\theta|D) = \sum_{m \in M} p(\theta|D, m)P(m|D)$$

(36)

where $p(\theta|D, m)$ is the posterior of $\theta$ given a model $m$ and, $p(m|D)$ the posterior of the model $m$. Equation 36 is the marginalization of the joint distribution $p(\theta, M)$ with respect
to the model variable $M$ given the data $D$. According to the second axiom of Kolmogorov [72], the sum of the second factor in the right hand side of Equation 36 over all models $m \in M$ is one as written in
\[ \sum_{m \in M} P(m|D) = 1 \] (37)

Therefore, Equation 37 is merely a weighted average of $p(\theta|D, m)$, the posterior of $\theta$ given a model $m$, with $P(m|D)$, the posterior of model $m$, as the weighting factor.

Now change our perspective to the $M$-open perspective. What if all the models are wrong? Even though all the models are wrong, it may be worth comparing tentative theories or hypotheses, and the subsequent models.

### 3.4.3 Bayesian Model Selection and Averaging in Linear Regression

Variable selection is a crucial issue in regression analysis. Using a subset of variables is sometimes better than using all of them. Let us call the model with all variables as the full model hereafter. Classical approaches to variable selection are based on sequences of hypothesis tests, for example, stepwise regression [56, 33]. In recent years several Bayesian approaches have been developed as well regarding to this issue.

Most of the Bayesian approaches use hierarchical modeling. Conceptually speaking, a hierarchical model for linear regression consists of three variables: the model variable $M$, the vector of regression coefficients $\beta$, and the data $D$ as shown in Figure 21. The model variable $M$ is a categorical variable with as many categories as the number of models. The model variable $M$ is connected to the vector of regression coefficients $\beta$ so that a regression coefficient becomes either zero or a value in a prescribed range as the state of the model variable changes. If the regression coefficient is zero, the corresponding predictor is deleted from the regression model. $\beta$ and $M$ are connected to the data $D$. The hierarchical model is completed with the probabilities in direction of the solid arrows in Figure 21: $p(\beta|M)$ and $p(D|\beta)$. When $D$ is instantiated, the posterior distributions of $\beta$ and $M$, $p(\beta|D)$ and $p(M|D)$, respectively, can be updated. This inference is shown as the dotted arrows in Figure 21.

Mitchell and Beauchamp [87] assigns a mixture of two uniform distributions for the
prior distribution of each regression coefficient: a point mass at zero and a diffuse uniform distribution over a range. This kind of prior distribution is referred to as the “spike and slab” distribution. If the posterior of a regression coefficient is concentrated zero, the corresponding predictor is deleted from the regression equation.

If the number of models is large, averaging over all the models using Equation 36 is computationally burdensome. To alleviate the computational burden the summation in Equation 36 is often approximated using a Markov Chain Monte Carlo (MCMC) method.

George and McCulloch [44] uses a normal mixture prior for each regression coefficient. Model averaging is approximated using Gibbs sampling. The procedure is named as SSVS (stochastic search variable selection).

Raftery, Madigan, and Hoeting [94] adapted a MCMC approach originally developed by Madigan and York [77] to investigate dependency between variables in discrete graphical models. Their algorithm is named as the MC³ (Markov chain monte carlo model composition) algorithm.

While in many studies each predictor enters the model independently, some authors take the correlations among predictors into consideration. Chipman [25] adapted SSVS with the modification of priors to incorporate the relations between predictors. Yuan and Lin [133] uses a mixture of a point mass at zero and a double exponential distribution for the
prior distribution of each regression coefficient. The use of double exponential distribution accommodates heavy tail probability. Like George and McCulloch [44], Gibbs sampling is used for estimating posterior probabilities efficiently [134].

3.4.4 Difference Between Regression and Fault Identification

Bayesian model selection and averaging have been researched in statistics mostly in terms of linear regression. Although fault identification is a regression type of problem, the author feels that it would be better to clarify differences between regression and fault identification.

The main difference between these two problems is what we have in hand what we want to know using what we have. Regression is to determine regression coefficients given data consisting of responses and factors. For example, Figure 22 shows an one dimensional regression problem. The goal of this regression problem is to determine the regression coefficients $\beta_0$ and $\beta_1$ given pairs of $x$ and $y$.

In contrast, fault identification is to determine factors given responses and regression coefficients. For example, Figure 23 shows an example of fault identification problems. The goal of this problem is to determine the factor $x$ given the observable $y$ and the coefficients $a_0$ and $a_1$. The coefficients $a_0$ and $a_1$ are characteristics of the system. They can be obtained from expert opinion, or regression of historical data, experiments, or computer simulations.
Figure 23: Example of Fault Identification Problems
Chapter IV

RESEARCH FORMULATION

Chapter 2 provides an extensive, but not exhaustive, survey on numerous algorithms for fault diagnosis. The direction of the literature study is narrowed down to fault identification in Chapter 3. Chapter 3 contains an introduction of Bayesian statistics and several Bayesian approaches in linear regression. While these Bayesian approaches in linear regression is concerned about the model uncertainty, most of the literature in fault identification appear not to take the model uncertainty into consideration.

In this chapter we will formulate a research problem that is inspired by the Bayesian approaches in linear regression. The research problem is intended to answer several research questions raised throughout the literature survey. A hypothesis is posed to address these research questions.

4.1 Observations and Research Questions

Throughout the literature study in Chapters 2 and 3, several observations are made.

The first observation is regarding to the data quality issue. A system can be diagnosed using observable behaviors of the system, for example, measurements obtained from the system. Measured data always contains random sensor noise and sometimes other types of error as well. Poor quality in the data propagates down to fault diagnosis and, subsequently, results poor quality diagnosis. Therefore, data should be validated before it is used for fault diagnosis.

How can data be validated? Data is compared to what it is expected to be: based on intuition, comparison with redundant data or a computer simulation. This comparison involves an implicit assumption: The physical condition of the system is known. Otherwise, the deviation of the data from what it is expected is indeterminable whether it is because of a bad sensor or an abnormality in the system. The first observation is:
Even though fault diagnosis and data validation are interdependent, they are often performed separately.

The next three observations are regarding to fault identification and the existing algorithms. A fault identification problem is often transformed to an optimization problem. A solution of the optimization problem is the one minimizes or maximizes an objective function and, at the same time, satisfies some constraints. The solution is a point in the solution space. The uniqueness of the solution is often not guaranteed. The point solution could be one of multiple solutions, and the rest of the potential solutions are discarded unless the optimizer is designed to explore the solution space and to report all potential solutions. The second observation is:

Even though there is the possibility of multiple solutions in a fault identification problem, formulating the problem as optimization and finding a point solution often fails to report multiple solutions.

The third observation is also regarding to the selection of unknown variables to be identified. If some unnecessary variables are chosen, the existing algorithms appear to overestimate the unnecessary variables and underestimate necessary ones. This phenomenon is called the smearing effect. If some necessary variables are missed, the resulting solution is wrong. Whether a variable is necessary or not depends on how the data is generated in the system. Of course, it is absolutely impossible to know which variables are necessary or unnecessary beforehand in real world problems. There is an ambiguity regarding to the selection of variables. The third observation is:

While formulating a fault identification problem, there is ambiguity regarding to the selection of variables. Incorrectly selected variables can give rise to the smearing effect or incorrect results.

The above observations raise the following questions, which are worth researching.

1. How can we identify faults with data containing error\(^1\)?

\(^1\)By “error” the author means the difference between a measurement and the true quantity.
2. How can we identify multiple solutions of fault identification if they are present?

3. How can we take account for the ambiguity in selecting variables to be identified?

Answering the above research questions is the ultimate goal of this work.

4.2 A Fault Identification Method Inspired By Model Selection and Averaging

The research questions in the previous section are attempted to be answered using the scientific method: posing hypotheses and collecting supporting evidence of the hypotheses. Before we pose hypotheses, let us walk through each research question in turn.

The first research question is:

How can we identify faults with data containing error?

Some authors already attempted to resolve this issue by treating error in data as a variable, which is to be estimated along with state variables. A major drawback of their attempt is that too many variables are to be estimated. A large number of variables not only causes a computational burden, but also increases the dimension of solution space. As a result, the solution is likely to have a stronger smearing effect. The issue cannot be fully resolved just by treating errors as a variable.

Although the first question is not fully answered, let us move on to the next research question and revisit the first one later. The second research question is regarding to the possibility of multiple solutions:

How can we identify multiple solutions of fault identification if they are present?

By multiple solutions we mean an existence of more than one state that matches the data in hand. A problem of the existing optimization-based approaches is that they result in a point solution. The solution is a point in the solution space. It may be possible to find multiple solutions if the optimization repeats with different initial points. One way to avoid the repetition is to deal with probability distributions, instead of point solutions. An estimation problem can be transformed to a statistical inference problem using Bayesian rule. However,
Bayesian statistics is subjective. It will choose a solution closer to prior belief if multiple solutions exist. Although it is acceptable, or even desirable, to use prior belief or knowledge in order to choose a solution over the others, if there is no sufficient reason to favor a particular solution, it make sense to follow objective Bayesianism [8]. According to Jeffreys’ argument on the principle of insufficient reasons in estimation problems, if we use a diffuse proper prior for each variable, all values in its range is equally probable and, consequently, there would be no preference on one solution over another. On the other hand, diffuse priors can give rise to less informative posteriors unless data is abundant.

Like the first research question, the second question is also left partially answered. Let us move on to the next research question again and revisit the first two questions later. The third research question is regarding to the ambiguity in the selection of variables:

How can we take account for the ambiguity in selecting variables to be identified?

It was observed that a wrong set of variable can cause the smearing effect or give rise to an incorrect solution. The cause of the smearing effect is a larger dimension of solution space than it should be. For example, suppose that the true solution is a point on a square-shaped solution space. If we search for the true solution inside a cube of which the square is a facet, the existing algorithms tend to find not the true solution but a point near it. When we restrict our search on the facet, we can reduce the inaccuracy of our estimate caused by the extra dimension. Of course, we never know where the solution is located beforehand. It can be either on a facet, on another facet or inside the cube. Therefore, we need to examine all facets and the space enclosed by the facets. All possible sets of variables should be examined. In the statistical point of view, the previous sentence can be rephrased as: we need to consider multiple statistical models with varying complexity. By doing so the model uncertainty is taken into consideration.

Now let us revisit the first two research questions. The difficulty remaining in the first two questions also can be resolved by taking account for the model uncertainty. First of all, the unsolved difficulty in the first question is the smearing effect strengthened by adding extra error variables. The strengthened smearing effect can be reduced by taking the model
uncertainty into consideration as argued in the last paragraph. Second, the remaining
difficulty in the second question is the inconclusiveness of results caused by the use of
diffuse priors. The increased uncertainty by the use of diffuse priors can be compensated
by suppressing the model complexity. A simple model with diffuse priors gives rise to
more conclusive results than that of complexer ones. Therefore, even though some complex
models still result in inconclusive results, results from the incorporation of multiple models
with different complexity will be better than that of a single complex model.

Based on the above argument, we propose a hypothesis that answers all the three research
questions. Since all three questions are related to each other in one way or another as
described previously, it is hard to address them separately. Thus, a single hypothesis is
posed and addresses all the three questions as follows:

In a state estimation problem, if error is treated as a variable, it can be esti-
mated along with state variables. If the estimation problem is transformed to a
statistical inference problem and a diffuse proper prior is used for each variable,
multiple solutions can be identified at once if they are present. The smearing
effect augmented by the increased number of variables and the diffuse posteriors
caused by the diffuse priors can be mitigated by incorporating multiple statistical
models with varying complexity.

The rest of this thesis will be devoted to describing the implementation of the hypothesis
and presenting substantial supporting evidence of the hypothesis.

4.3 Boundary of the Thesis

To collect supporting evidence of the hypothesis, the abstract hypothesis is transformed to a
mathematical method. The method is applied to a couple of applications, and the results are
examined to see if the hypothesis is true. Any unnecessary complexity, which is not relevant
to the hypothesis, hinders clearer and simpler explanation of results. To maintain this work
manageable and simple, the problem we attempt to solve is under certain conditions.

The first condition is regarding to the time dependency of the problem. If the data
in the problem is time dependent, a nonstationary state estimation technique is required.
Nonstationary problems are often harder to solve than stationary ones. This difficulty can be avoided under the following condition:

It is a steady-state problem.

The second condition is in the same line with the first one. Among the four sensor fault modes, bias and no-signal are steady, while spike and drift are unsteady. No-signal can be seen as a special case of bias; the magnitude of bias in a measurement is exactly same as the true value. To maintain the problem being steady-state, the following condition has to be met:

Only bias is taken account for in the problem among the four sensor fault modes.

The third condition is regarding to models. There are an infinite number of models one can think of. Of course, we cannot consider all of them. To limit the scale of the problem within a manageable size, the following condition has to be satisfied:

There is a pool of predefined models before data is analyzed, and only the models in this pool will be taken into consideration.
A statistical inference problem can be solved in various ways. This chapter explains how differently a statistical inference problem can be solved and which way this dissertation takes.

5.1 General Theory of Bayesian Networks

A Bayesian network (BN) is a graphical framework to model a probabilistic problem of interests. A BN is defined by a pair of a graph and probabilities. The graph consists of nodes and arcs. A node represents a random variable or event; the words node and variable are used synonymously hereafter. An arc indicates a probabilistic dependency from one node (a parent) to the other (a child). The child node is associated with a conditional probability distribution (CPD) dependent on its parents’ state. A node without any parent node is called a root node. A probability distribution associated with a root node is called a prior distribution.

Consider a BN consisting of \( n \) variables, \( \mathbf{X} \). Then the joint probability distribution (JPD) of the variables is a product of CPDs as follows [23]:

\[
p(\mathbf{X}) = \prod_{i=1}^{n} p_i(X_i|X_1, \ldots, X_{i-1})
\]  

(38)

Among the conditioning nodes of \( p(X_i|X_1, \ldots, X_{i-1}) \), all and only the parents of \( X_i \) need be in the conditioning portion. Therefore, the JPD can be simplified as follows [23]:

\[
p(\mathbf{X}) = \prod_{i=1}^{n} p(X_i|\pi_i)
\]  

(39)

where \( \pi_i \) is the set of parents of node \( X_i \).
5.1.1 Discrete Networks

Variables in a Bayesian network can be either discrete or continuous. A discrete variable has a probability table, whereas a continuous variable follows a continuous probability distribution. A continuous variable can be discretized with several discrete states and treated as a discrete variable. However, this discretization can cause several problems.

First, to increase resolution of solutions the continuous variable has to be discretized with a sufficient number of states. The increase in the number of states means an increase of the computational burden. With a limited computational power the resolution of solutions has to be compromised.

Second, characteristics of the continuous variable is fundamentally different from those of a discrete variable. For example, Figure 24 shows a continuous variable that is discretized with two states. Although $x_2$ is much closer to $x_3$, it falls into the same state as $x_1$. Would the discretized variable be a good representation of the original continuous variable?

Figure 24: Discretization of a Continuous Variable
Using continuous variables saves substantial amount of computation because a continuous variable can be parameterized with a few numbers, for example, the mean and variance of a normal distribution. In addition, it does not cause the resolution issue of discrete variables. A difficulty in using continuous variables is to find the right distribution for a continuous variable.

5.1.2 A Linear Gaussian Model

Let us consider a continuous variable $Y$ and its continuous parents $X = \{X_1, \ldots, X_j\}$. It is said that $Y$ has a linear Gaussian model if the CPD of $Y$ is a Gaussian distribution whose mean is a linear function of $X$, and whose variance does not depend on $X$. The CPD is written as follows [28]:

$$p(Y|X_1, \ldots, X_j) = N\left(\mu_Y + \sum_{i=1}^j \beta_i (X_i - \mu_i), \sigma^2\right)$$

(40)

where $\beta_i$ is the regression coefficient of $X_i$ in the regression of $Y$, and $\sigma^2$ is the conditional variance. $\beta$ and $\sigma^2$ are written as

$$\beta = \Sigma_{YX} \Sigma_{XX}^{-1}$$

$$\sigma^2 = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$$

(41)

where $\Sigma_{YY}$ is the unconditional variance of $Y$, $\Sigma_{YX}$ the covariance matrix between $Y$ and $X$, $\Sigma_{XY}$ the transpose of $\Sigma_{YX}$, and $\Sigma_{XX}$ the covariance matrix of $X$. A Bayesian network is said to be a Gaussian Bayesian network if all of its variables are continuous, and all CPDs are linear Gaussians. Gaussian Bayesian networks have several advantages over other types of networks. First, posterior distributions in a Gaussian Bayesian network are analytically derivable. Second, only two parameters are required to define a Gaussian variable.

Among the variables constituting a Bayesian network, some may be observable, and others may not. The observed variables are called evidence, and the unobservable nodes hidden nodes. The probability of hidden variables can be updated when some evidence becomes available; this process is called inference.

Let’s partition variables into two groups the unobserved, $X$, and the observed, $Y$. The mean and covariance matrix of the unobserved conditioned by the observed can be written
as follows [22]:

\[ \mu_{X|Y} = \mu_X + \Sigma_{XY} \Sigma_{YY}^{-1} (Y - \mu_Y) \]  
\[ \Sigma_{X|Y} = \Sigma_{XX} - \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX} \]  

\[ (42) \]
\[ (43) \]

5.2 Example Application

As shown in the previous section, the use of Gaussian variables enables simple and analytic inference. To test whether a linear Gaussian model could be applied to complex engineering systems, it is applied to a simple —but not trivial— problem. The quadruple-tank process [64] is selected as an example application. The quadruple-tank process is a multivariate laboratory process designed to illustrate multivariate control systems. It was developed and built to be used in the control curriculum at Lund Institute of Technology, Sweden, in 1996.

The quadruple-tank process consists of four water tanks, two pumps, and a water reservoir. Each pump draws water off the reservoir at the bottom, and fills up two tanks. Each
Table 3: Laboratory Settings of the Quadruple-Tank Process [64]

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Unit</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1, A_3$</td>
<td>the cross-section of tanks</td>
<td>cm(^2)</td>
<td>28</td>
</tr>
<tr>
<td>$A_2, A_4$</td>
<td>the cross-section of tanks</td>
<td>cm(^2)</td>
<td>32</td>
</tr>
<tr>
<td>$a_1, a_3$</td>
<td>the cross-section of holes</td>
<td>cm(^2)</td>
<td>0.071</td>
</tr>
<tr>
<td>$a_2, a_4$</td>
<td>the cross-section of holes</td>
<td>cm(^2)</td>
<td>0.057</td>
</tr>
<tr>
<td>$k_1, k_2$</td>
<td>the flow coefficient</td>
<td>cm(^3)/Vs</td>
<td>3.33, 3.35</td>
</tr>
<tr>
<td>$\gamma_1, \gamma_2$</td>
<td>the valve settings</td>
<td>-</td>
<td>0.7, 0.6</td>
</tr>
<tr>
<td>$g$</td>
<td>the acceleration of gravity</td>
<td>cm/s(^2)</td>
<td>981</td>
</tr>
</tbody>
</table>

Tanks has a hole at the bottom surface so that the water in the upper tanks pours down to the lower tanks, and the water in the lower tanks flows down back to the reservoir. A nonlinear model of this process can be derived using the mass balance and the Bernoulli’s equation:

\[
\begin{align*}
\frac{dh_1}{dt} &= -\frac{a_1}{A_1} \sqrt{2gh_1} + \frac{a_3}{A_1} \sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1} v_1 \\
\frac{dh_2}{dt} &= -\frac{a_2}{A_2} \sqrt{2gh_2} + \frac{a_4}{A_2} \sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2} v_2 \\
\frac{dh_3}{dt} &= -\frac{a_3}{A_3} \sqrt{2gh_3} + \frac{(1-\gamma_2) k_2}{A_3} v_2 \\
\frac{dh_4}{dt} &= -\frac{a_4}{A_4} \sqrt{2gh_4} + \frac{(1-\gamma_1) k_1}{A_4} v_1
\end{align*}
\]

(44)

where $A_i$ is the cross-section of the tank $i$, $a_i$ the cross-section of the outlet hole, and $h_i$ the water level. The description of parameters are laboratory settings are given in Table 3.

The goal of the laboratory experiment is to control the water level of the two lower tanks. The amount of water drawn from the reservoir is controlled by the voltage applied on each pump. Thus, the inputs to the process is the voltages, $v_1$ and $v_2$, and the outputs of the process is the water levels, $h_1$ and $h_2$. Let us for convinience convert the variables $h$ and $v$ into the form of deviation from a nominal operating condition $h_0$ and $v_0$, and introduce the new variables $y = h - h_0$ and $x = v - v_0$. Then, the linearized equation can be written as

\[
\frac{dy}{dt} = \begin{bmatrix}
-\frac{1}{T_1} & 0 & -\frac{A_3}{A_1 A_3} & 0 \\
0 & -\frac{1}{T_2} & 0 & -\frac{A_4}{A_2 A_4} \\
0 & 0 & -\frac{1}{T_3} & 0 \\
0 & 0 & 0 & -\frac{1}{T_4}
\end{bmatrix} y + \begin{bmatrix}
\frac{\gamma_1 k_1}{A_1} \\
0 \\
0 \\
(1-\gamma_1) k_1
\end{bmatrix} x
\]

(45)

60
The governing equation in the steady state can be obtained by substituting $\frac{dy}{dt} = 0$.

The validation of the proposed method requires the measurement data of the water levels at various operating conditions. Because the actual experimental data are not available, the required data are generated by solving the governing equation. In order to represent the uncertainty in real world, the pump voltage is assumed to fluctuate around the mean value with the standard deviation of 0.01, and the measurement of water level is assumed to be corrupted by random noise with standard deviation of 0.01.

5.3 Implementation

Based on the governing equation of the quadruple-tank process and the sensor models introduced in §2.1, two Gaussian Bayesian networks were developed to model the process: the basic and bias-augmented networks. They were tested with some simulation data to investigate their effectiveness and limitations.

5.3.1 The Basic Network

The basic sensor model was combined to the quadruple-tank process; the resulting network, which will be referred to as the basic network hereafter is shown in Figure 26. The nodes in the upper two rows model the quadruple-tank process, and the ones in the bottom row are measurements of water level of each tank.

In order to complete the network, the probability densities have to be assigned for each

$$T_i = \frac{A_i}{a_i} \sqrt{\frac{2h_i^0}{g}} \quad \text{for all } i = 1, \ldots, 4$$

(46)
node. The prior density for the variable \( X_i \) is a normal distribution with zero mean and the standard deviation of \( \sigma_U \); 
\[ X_i \sim N(0, \sigma_X) \]
The conditional probability density (CPD) of \( Y_i \) is assumed to be a multivariate normal distribution \( N(\mu, \Sigma_Y) \). According to the governing equation, the mean vector of \( Y \) is a function of \( X_1 \) and \( X_2 \):

\[
\begin{align*}
\mu_1 &= 5.2201X_1 + 3.0008X_2 \\
\mu_2 &= 2.8202X_2 + 5.6742X_2 \\
\mu_3 &= 1.1433X_2 \\
\mu_4 &= 0.9363X_1 
\end{align*}
\]

The random error is modeled with a normal distribution with zero mean and a constant standard deviation \( \sigma_\varepsilon \): \( N(0, \sigma_\varepsilon) \). The covariance matrix of \( Y, \Sigma_Y \), can be obtained from Equation 41.

The development of network is completed as the structure is constructed and the CPD is assigned for each node. To validate the completed network a couple of test cases were run. The first case was the change in the voltages provided to Pump 1 and 2; it will be referred to as Case 1 hereafter. The purpose of this test case was to investigate how well the network estimates the hidden nodes \( X \)'s with noisy data \( Y \)'s. It was assumed that the input voltages on Pump 1 and Pump 2 increases and decreases, respectively, by \( 3\sigma_X \) from the operating condition: \( (X_1, X_2) = (3\sigma_X, -3\sigma_X) \). Twenty sample points were generated by adding random errors to the output calculated from the governing equation, and they are shown in Figure 27.

With the twenty sample points, the posterior distributions of \( X_1 \) and \( X_2 \) were updated and shown in Figure 28. In prior the voltage, \( X_i \), was assumed to fluctuate around zero with the standard deviation of 0.01. After the twenty sample points are acquired, the network estimated \( X \)'s accurately with high confidence. Thus, it could be concluded that the basic sensor model implemented in the network is capable of handling data corrupted with random errors, although the accuracy may depend on the amount of data.

The second test case was designed to see how the network behaves with sensor biases as well as the change in the hidden nodes, \( X_1 \) and \( X_2 \); it will be referred to as Case 2 hereafter.
Figure 27: Simulated Measurements with \((X_1, X_2) = (3\sigma_X, -3\sigma_X)\): Case 1 of the Quadruple-Tank Process

The pump voltages are same as in Case 1, but the measurement of water level of Tank 2 is constantly biased by \(3\sigma_Y\): \(Y_2 = X_2 + 3\sigma_Y + \varepsilon\). Twenty sample points were generated in the same way as in Case 1, and they are shown in Figure 29. Like Case 1 the probability distributions moved toward the true value from the prior distributions. The estimated \(X\)'s were, however, not as accurate as the one from Case 1 due to the presence of the bias in \(Y_2\).

To sum up the results from the two test cases, the basic sensor model can handle random errors, but not sensor biases, which persistently exist on measurements and affect on the mean of measurements. Consequently, when a bias is present in a measurement, the solution becomes less accurate than when no biases are present.

5.3.2 The Bias-Augmented Network

Due to the lack of mean to detect sensor biases, the basic network provided less accurate results when a bias is present in a measurement than when no biases are present. To model a bias a bias node can be added to the basic network. A bias node is connected to each measurement. The resultant network will be referred to as the bias-augmented network. Figure 31 shows the bias-augmented network in which four bias nodes are added in addition
Figure 28: Posterior Distribution of $X_1$ and $X_2$ in Case 1 of the Quadruple-Tank Process
Figure 29: Simulated Measurements with \((X_1, X_2) = (3\sigma_X, -3\sigma_X)\) and the Bias in \(Y_2\):
Case 2 of the Quadruple-Tank Process

to the basic network.

Unlike the random noise of a sensor, which can be revealed by experiments or given from the sensor manufacturer, the statistical knowledge on a bias rarely exist. It is hard to obtain an accurate uncertainty associated with them; what is the chance of a mechanic improperly calibrating a sensor by a particular value? Thus, the standard deviation of a bias, \(\sigma_B\), is set to the same value of the random noise. treated as an adjustable parameter to tune the behavior of the network.

Table 4 lists the maximum a posterior estimate of each variable, in other words, the mean of posterior distribution of each variable obtained with various \(\sigma_B\). As well as the actually biased measurement \(Y_2\), all other measurements were reported to be more or less biased, too. Without knowing the true values in advance, it is hard to determine whether these non-zero mean values of bias nodes are actually biases or some erroneous results. It is an example of the smearing effect, which most existing fault identification algorithms suffer.

Although the bias-augmented network in Figure 31 is the most general form, it does not precisely model how the data are generated in this test case. A network that models the data generation the closest is shown in Figure 32. This network contains a bias node only on
Figure 30: Posterior Distributions of $X_1$ and $X_2$ of the Quadruple Tank Application
Figure 31: Bias-Augmented Network for the Quadruple-Tank Process

Table 4: Maximum A Posterior Estimate of Each Variable from the Bias-Augmented Network with Various $\sigma_B$: the Quadruple-Tank Process

<table>
<thead>
<tr>
<th>$\sigma_B$</th>
<th>$\mu_{X_1}$</th>
<th>$\mu_{X_2}$</th>
<th>$\mu_{B_1}$</th>
<th>$\mu_{B_2}$</th>
<th>$\mu_{B_3}$</th>
<th>$\mu_{B_4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Value</td>
<td>0.03</td>
<td>-0.03</td>
<td>0</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma_B = 0.001$</td>
<td>0.0263</td>
<td>-0.0234</td>
<td>-0.0002</td>
<td>0.0003</td>
<td>-0.0012</td>
<td>0.0006</td>
</tr>
<tr>
<td>$\sigma_B = 0.01$</td>
<td>0.0233</td>
<td>-0.0207</td>
<td>0.0060</td>
<td>-0.0048</td>
<td>-0.0097</td>
<td>0.0064</td>
</tr>
<tr>
<td>$\sigma_B = 0.03$</td>
<td>0.0121</td>
<td>-0.0106</td>
<td>0.0342</td>
<td>-0.0305</td>
<td>-0.0216</td>
<td>0.0171</td>
</tr>
</tbody>
</table>
Figure 32: Bayesian Network Based on the Data Generation Mechanism in Case 2 of the Quadruple-Tank Process

Table 5: Maximum A Posterior Estimate of Each Variable from the Various Networks: the Quadruple-Tank Process

<table>
<thead>
<tr>
<th></th>
<th>( \mu_{X_1} )</th>
<th>( \mu_{X_2} )</th>
<th>( \mu_{B_1} )</th>
<th>( \mu_{B_2} )</th>
<th>( \mu_{B_3} )</th>
<th>( \mu_{B_4} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Value</td>
<td>0.03</td>
<td>-0.03</td>
<td>0</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True Model</td>
<td>0.0280</td>
<td>-0.0267</td>
<td>-</td>
<td>0.0148</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Full Model</td>
<td>0.0233</td>
<td>-0.0207</td>
<td>0.0060</td>
<td>-0.0048</td>
<td>-0.0097</td>
<td>0.0064</td>
</tr>
</tbody>
</table>

\( Y_2 \), which is the actually biased measurement. Let us refer to the former as the full model and the latter as the true model. In both models \( \sigma_B \) is set to 0.01. Table 5 shows maximum a posterior estimate of each variable from the full and true models compared with the true value. The true model is more accurate than the full model because it does not have the semaring effect in its solution.

5.4 Chapter Summary

In this chapter the general theory of Bayesian networks is discussed. It is illustrated how a discretization of a continuous variable can cause issues such as the resolution of solutions, an increased computational burden, and the ambiguity in assigning probabilities on discrete states.

In addition to the discussion of a discretization of continuous variables, a linear Gaussian model is explained. A linear Gaussian model uses only continuous variables, more specifically, Gaussian variables. The linear Gaussian model is applied to the example application,
the quadruple-tank process. It is demonstrated that if biases are not included in a Bayesian network, the solution becomes inaccurate when a bias is present in measurements. This result stresses that Research Question 1 has to be answered properly. This issue is not limited to the Gaussian Bayesian network. Any type of Bayesian network faces this issue.

Even though biases are included in the network, if unnecessary bias variables are added and connected to an unbiased measurement, the solution becomes smeared. When bias variables are added and associated with the measurements that are actually biased, the Bayesian network gives rise to less smeared solutions. This result demonstrates why the model uncertainty has to be taken into consideration and Research Question 3 has to be answered.

A linear Gaussian model gives rise to a normal posterior distribution, which is unimodal. Thus, in the example application, multiple solutions are not identified.

The Matlab code for the Gaussian Bayesian network of the quadruple-tank process is listed in Appendix A.
Chapter VI

PROPOSED METHOD

“If you learn to look at the data in the right way, you can explain riddles that otherwise might have seemed impossible.”

— Steven D. Levitt and Stephen J. Dubner in “Freakonomics”

The abstract hypothesis given in Chapter 4 is transformed to a tangible form in this chapter. This chapter presents a fault identification method, which can identify physical faults and sensor faults. The formulation of the method is based on the preliminary study in the previous chapter.

This chapter explains how to build a Bayesian model for fault identification and how to incorporate multiple models in one framework. It is also suggested how to interpret statistical results obtained using method.

6.1 Formulation

6.1.1 A Bayesian Model for the System of Linear Equations

Let $X$ be a vector of state variables and $Y$ a vector of observations. The state vector and the observables have a functional relationship $f$:

$$Y = f(X) + \varepsilon$$  \hspace{1cm} (48)

when $\varepsilon$ is sensor random noise. Given an operating condition, the relationship $f$ can be linearized as

$$Y = AX + \varepsilon$$  \hspace{1cm} (49)

where $A$ is the coefficient matrix. Now $\varepsilon$ includes not only the random noise but the linearization error as well. When the observable $Y$ is subject to sensor bias $B$, Equation 8 can
be written as follows:

\[ Y = AX + B + \varepsilon \]  

(50)

Let us assume that the conditional probability of \( Y \) given \( X \) and \( B \), \( p(Y|X,B) \), follows a multivariate normal distribution \( N(\mu,\tau) \) where \( \mu \) is the mean vector and \( \tau \) is the precision matrix.

Given \( X \) and \( B \), the mean vector can be written as

\[ \mu|X,B = b_0 + b_1 X_1 + \ldots + b_l X_l + B \]  

(51)

where \( b_i \) is the elements of the coefficient matrix \( A \) and \( l \) is the number of state variables.

With limited data or knowledge the precision matrix \( \tau \) is often hard to define. Thus, \( \tau \) is considered as a variable that is to be inferred from data. A conjugate prior of the precision matrix of a multivariate normal distribution is a Wishart distribution [26]. We use a non-informative and disperse prior for \( \tau \)

\[ \tau \sim W(\Lambda, \nu) \]  

(52)

where \( \Lambda \) is the scale matrix, and \( \nu \) the degree of freedom. The elements of \( \Lambda \) are set to the values that make the Wishart distribution disperse. The probability density function of a Wishart distribution is defined as

\[ p(x) = |\Lambda|^\nu/2 |x|^{(\nu-p-1)/2} \exp \left[ -\frac{1}{2} Tr(\Lambda x) \right] \]  

(53)

where \( Tr(\bullet) \) is the trace of a matrix [111].

Whether the normal distribution assumption is valid or not depends on the applications on which the method is about to applied. Historically, normal distributions are often used not because data actually follow them but because they are mathematically easy to handle. If we were attempting to solve a particular problem, it should be examined what distributions \( Y \) follows. However, it is not the case. We follow the tradition without further justification.

According to the principle of insufficient reasons, a state variable \( X_i \) can be any value in a range, and no particular value is more likely than others in the range. This notion is expressed with a uniform distribution \( U(a,b) \) where \( a \) and \( b \) are the lower and upper
boundaries of the uniform distribution. With the same reason the bias $B$ is assumed to follow a uniform distribution as well.

### 6.1.2 Multiple Models

Let $\theta$ be a combined set of $X$ and $B$, $\theta = \{X_1, X_2, \ldots, X_l, B_1, B_2, \ldots, B_m\}$. One can build a model with either all the state variables and biases, or a subset of them. The total number of possible subsets is $2^n$ where $n$ is the number of elements in $\theta$. As the model variable changes from one model to another, some variables are added and others are removed from the model. When a variable is included in the model, it is to be estimate. In contrast, when a variable is excluded from the model, it is fixed at a prescribed value. The challenge is how to implement this inclusion and exclusion of variables in the model numerically. To do it efficiently, an auxiliary variable $\gamma$ is introduced to the model and connected to each $\theta$. $\gamma$ is a binary variable with two states: zero and one. $\gamma$ controls the mixture of two uniform distributions and assigns it to $\theta$ as follows:

$$\theta | \gamma \sim (1 - \gamma)U(\theta_0 - \delta, \theta_0 + \delta) + \gamma U(a, b)$$  \hspace{1cm} (54)

where $\theta_0$ is the nominal value and $\delta$ is a small number. When $\gamma$ is zero, the uniform prior of the corresponding $\theta$ is concentrated at a particular value. On the other hand, when $\gamma$ is one, the uniform prior covers the range of interest. This virtually emulates the inclusion and exclusion of the corresponding $\theta$ connected to $\gamma$. The mixture of the two uniform distributions, called the **spike and slab prior** [87], is depicted in Figure 33.

The $2^n$ models can be shown in one generic graph as shown in Figure 34. Because $\gamma$, $X$, $Y$, and $B$ are vectors, the nodes representing them can be expanded using each element of the vectors. Unlike other arrows, the arrow between $M$ and $\gamma$ is deterministic. If a particular model $m$ includes $X_i$, then the corresponding $\gamma_i$ is set to one. Otherwise, $\gamma_i$ is set to zero.

### 6.1.3 Posteriors $p(M|Y)$ and $p(\theta|Y)$

In the previous two sections the graphical model is constructed using the state variables and biases, and the observables. In addition, the conditional probability of a child given its parents’ state is specified. Once the observation $Y$ is made, the probability distributions of
**Figure 33:** The Spike and Slab Prior

**Figure 34:** Generic Graphical Model of the Current Formulation
other variables can be updated. The updated probability distribution is analytically derived in this section.

According to the Bayes rule [60], the model posterior \( p(M|Y) \) is written as

\[
p(M|Y) = \frac{p(Y|M)p(M)}{p(Y)} \tag{55}
\]

The marginal probability \( p(Y) \) is independent of the model variable \( M \). Thus, Equation 55 becomes

\[
p(M|Y) \propto p(Y|M)p(M) \tag{56}
\]

\( p(Y|M) \) can be obtained by integrating the likelihood \( p(Y|\theta, M) \) over \( \theta \):

\[
p(Y|M) = \int_{\Theta} p(Y|\theta, M)p(\theta|M)d\theta \tag{57}
\]

After plugging Equation 57 into Equation 56, the model posterior can be expressed as

\[
p(M|Y) \propto p(M) \int_{\Theta} p(Y|\theta, M)p(\theta|M)d\theta \tag{58}
\]

\( p(M), p(Y|\theta, M), \) and \( p(\theta|M) \) are defined in the graphical model shown in Figure 34.

The parameter posterior \( p(\theta|Y) \) is a marginalization of \( p(\theta|Y, M) \) with respect to \( M \):

\[
p(\theta|Y) = \int_{M} p(\theta|m, Y)p(m|Y)dm \tag{59}
\]

Because the model variable \( M \) is discrete, the integration is converted to the summation over all models as follows:

\[
p(\theta|Y) = \sum_{m \in M} p(\theta|m, Y)p(m|Y) \tag{60}
\]

Equation 60 is nothing but a weighted average of the posterior distributions of \( \theta \) from each model according to the posterior of \( M \).

The two factors in the right hand side, \( p(\theta|m, Y) \) and \( p(m|Y) \), are unknown. They should be transformed in terms of the conditional probability distributions specified along with the graphical model. The model posterior \( p(m|Y) \) can be substituted with Equation 56. \( p(\theta|m, Y) \) can be substitute with a function of known quantities using Bayes rule as follows:

\[
p(\theta|m, Y) = \frac{p(Y|\theta, m)p(\theta|m)}{p(Y|m)} \tag{61}
\]
Finally, the parameter posterior is written as

\[
p(\theta|Y) \propto \sum_{m \in M} p(Y|\theta, m)p(\theta|m)p(m)
\]  

(62)

### 6.1.4 Gibbs Sampling

In the previous chapter, the parameter and model posteriors, \(p(\theta|Y)\) and \(p(M|Y)\), are derived. The parameter posterior is a multi-dimensional integration if \(\theta\) is a vector of multiple elements. The model posterior is the summation of over all models. As the number of models increases, the terms to be summed increases as well.

Instead of calculating them analytically, the posteriors are estimated using Gibbs sampling. Gibbs sampling can draw samples approximately from the posteriors even when the posteriors are not known. In Gibbs sampling the states of \(M\) and \(\theta\) are updated throughout the following sequence given the observation \(Y\):

\[
\begin{align*}
\theta_1^{(j+1)} &\sim p(\theta_1^{(j+1)} | \theta_2^{(j)}, \theta_3^{(j)}, \ldots, \theta_n^{(j)}, M^{(j)}, Y) \\
\theta_2^{(j+1)} &\sim p(\theta_2^{(j+1)} | \theta_1^{(j+1)}, \theta_3^{(j)}, \ldots, \theta_n^{(j)}, M^{(j)}, Y) \\
\theta_3^{(j+1)} &\sim p(\theta_3^{(j+1)} | \theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_n^{(j)}, M^{(j)}, Y) \\
&\vdots \\
\theta_n^{(j+1)} &\sim p(\theta_n^{(j+1)} | \theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_{n-1}^{(j+1)}, M^{(j)}, Y) \\
M^{(j+1)} &\sim p(M | \theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_n^{(j+1)}, Y)
\end{align*}
\]

(63)

With a large \(j\) the sequence converges, and the final state of each variable can be thought of a sample from the corresponding posterior distribution. Multiple sequences are created in order to obtain multiple independent samples from the posteriors of \(M\) and \(\theta\). To see the effect of the initial point, multiple sampling chains can be created and start from different points.

### 6.2 Interpretation of Outputs

"The idea is that the computer seems more like an aid rather than a final device. What you are looking for is some guidance, not a model answer."

— Peter Norvig, Director of Security Quality of Google
The only output from the proposed method is Gibbs samples. The Gibbs samples contain all the information we need to know. The challenge is how we interpret them.

First of all, given the observables, the Gibbs samples are drawn from the approximate posterior distribution of each variable. The posterior distribution can be estimated from the Gibbs samples using a density estimate scheme, for example, \textit{ksdensity} [80] in the Matlab statistics toolbox. The posterior distribution of a state variable shows what value the state variable is likely to be and how likely that value is. Posterior distributions can be either unimodal or multimodal, as shown in Figure 35. When the posterior is multimodal, the posterior should be further investigated along with the posteriors of other variables.

When the Gibbs samples are plotted in a coordinate system constituted by two variables, the relationship between the two variables can be visualized. If the bivariate scatter plot has no pattern, it can be thought that the two variables are statistically uncorrelated. If the cloud of sample points forms a straight line, they are linearly co-dependent. Examples of uncorrelated and correlated bivariate scatter plots are shown in Figure 36. The degree of correlation can be measured with correlation coefficients. The Pearson correlation coefficient of two variables $X$ and $Y$ is written:

$$r^2 = \frac{(\Sigma xy - n\bar{x}\bar{y})^2}{(\Sigma x^2 - n\bar{x}^2)(\Sigma y^2 - n\bar{y}^2)}$$

(64)

where $\bar{x}$ and $\bar{y}$ are the sample means of $X$ and $Y$, $s_x$ and $s_y$ are the sample standard deviations of $X$ and $Y$, and $n$ is the number of samples [131]. It should be noted that a correlation coefficient measure how linearly two variables are correlated.

All correlation coefficients between each pair of variables can be visualized with an undirected graph. Figure 37 is an example of a undirected graph. A node represents a variable, and a line between two nodes is the correlation between them. The opacity of the line is proportional to the degree of correlation. The fainter the line is, the weaker the correlation is. The Mathematica code for plotting correlation graphs is listed in Appendix B.

The Gibbs samples of the model variable $M$ are used to determine the posterior distribution of $M$. Since $M$ is a categorical variable, the probability is calculated by counting the...
Figure 35: Examples of Posterior Probability Distribution Plots

(a) Unimodal Posterior Distribution

(b) Multimodal Posterior Distribution
Figure 36: Examples of Bivariate Scatter Plots of Gibbs Samples

(a) Samples from Uncorrelated Variables

(b) Samples from Correlated Variables
Figure 37: Example of Correlation Graphs
number of samples in each category and dividing it by the total number of the samples. The probability can be visualized in a stem plot as shown in Figure 38. The posterior probability tells which models are supported by data compared to others. The higher the probability of a model is, the more the model is supported.

The posterior distribution of $\gamma$ indicates how likely the corresponding $\theta$ is included in the model. Initially, all the models under consideration have the same probability since the uniform categorical distribution is assigned to the model variable. Each variable appears in half of the models while it does not in the other half. Therefore, the two states of $\gamma$, zero and one, are equally likely to be true initially. Once some data are given, some models are much more supported by the data than others are. As a consequence, $\gamma$ becomes more likely to be one if the corresponding variable commonly appears in the models with higher posterior probability. The posterior of $\gamma$ can be calculated in the same way as the posterior of $M$ and visualized with a bar plot as shown in Figure 39.
6.3 Process for Building a Hierarchical Bayesian Model for an Engineering System

To apply the proposed method to an engineering system, we can follow the three steps shown in Figure 40:

1. Create a system of linear equations governing the system

2. Determine the coefficient matrix $A$

3. Build a hierarchical Bayesian model

The first step is regrading to selecting state variables the observable. The system is parameterized so that the state of the system can be represented with several state variables $X$. Among available measurements some relavent ones are chosen as the observable $Y$. The relationship between the state variables and the observable can be linearized so that the system is approximated with a system of linear equations $Y = AX + \varepsilon$ where $\varepsilon$ is the random error. Each observable can be biased. A vector of bias variables, $B$, is introduced in the system of linear equations. The outcome of the first step is the system of linear equations,
\[ Y = AX + B + \varepsilon. \]

In the second step the coefficient matrix of the system of linear equations, \( A \), is determined. It can be determined from a regression of experiments or computer simulations. Any kind of reasonable design of experiments is usable. Each point in the design of experiments is either actually experimented or numerically calculated. The results are regressed using the first order linear equations.

The third step is to build a hierarchical Bayesian model using the variables and the system of linear equations determined in the first step, and the coefficient matrix \( A \) determined in the second step. The hierarchical Bayesian model can be shown as a graph. In the graph the state variables, the observable, and the bias variables are connected to one another. At this stage all variables are connected to each other. The resultant graph is referred to as the full model. To emulate a connection or disconnection between two variables, an auxiliary binary variable \( \gamma \) is added. In the graph \( \gamma \) is actually a vector. The model variable \( M \) are added to the graph and connected to \( \gamma \). Once the structure is completed, a probability distribution is assigned to each variable as explained in Section 6.1.
Step 1: Create a system of linear equations governing the system

- Parameterize the state of the system
- Select available measurements
- Set the range of biases

Step 2: Determine the coefficient matrix $A$

- Create a DOE
- 1st order regression

Step 3: Build a hierarchical model

- Build the full Bayesian model
- Expand the full model to a hierarchical model

**Figure 40:** Process for Building a Hierarchical Bayesian Model for Fault Identification
Chapter VII

APPLICATIONS OF THE PROPOSED METHOD

As a proof-of-concept two applications of the proposed method are presented in this chapter. First application is the status matching of a turbojet engine. In this application the unknown health condition of main components of the engine is estimated from the behavior of the engine. The second application is the fault diagnosis of an industrial gas turbine. In this application in addition to the health condition of components, the bias in sensor measurements are taken into consideration as well.

In each application various situations are tested. The data needed to perform the tests are generated using computer simulations. The results from the proposed method, which uses Bayesian model averaging (BMA), are compared with solutions from the true model and a single complex model. The proposed method is also tested with various types of noise, and its sensitivity to the error assumption are examined.

7.1 Status Matching of a Turbojet Engine

It is a common practice for a jet engine operator to determine the status of a gas turbine from test data\[122, 31, 93\]. An engine at a particular status yields the corresponding pattern of physical properties such as temperature and pressure at various locations of the engine. The status of the engine is represented by several parameters that scale engine performance relative to a baseline, for example, a new engine. The parameters are called health parameters. Given test data, the parameters are adjusted until the engine performance matches the test data. This process is usually referred to as status matching [97] or cycle matching [128].

7.1.1 The Problem

The Numerical Propulsion System Simulation (NPSS) [36], developed at NASA, is used for representing a turbojet engine. In this problem seven state variables are to be estimated
from five measurements as listed in Table 6. The state variables are scalars that adjust the performance of each engine component in NPSS. A scalar of one means the performance of the corresponding component is same as the performance at the design condition. The turbojet engine is modeled with a system of linear equations, $Y = AX + \varepsilon$. The unknown outnumber the known. Thus, the system of linear equations is underdetermined. For simplicity no biases are considered in this application. Since there are seven unknown variables, the number of models is $2^7 = 128$.

The nodes $X$ and $Y$ in Figure 34 are expanded using each state variable and the observable as shown in Figure 41. Technically, a state variables are connected to all the observables, even though some of the connections may be not as strong as others. The strength of a connection is the corresponding coefficient in the matrix $A$.

To determine the coefficient matrix $A$, a design of experiments (DOE) is built using NPSS simulations with all the state variables in the range of $[0.96, 1.02]$ except for the nozzle gross thrust coefficient, which is in the range of $[0.95, 0.99]$, while the operating condition is fixed to the sea level. Using the DOE, a first order linear regression equation is created for each observable as a function of the state variables. The created system of linear equations for the turbojet engine is listed in Appendix C.

Two cases are tested using the proposed method. The measurements in the two test case are simulated using NPSS. To simulate a degraded engine, NPSS is run with decreased scalars, and its output is added to Gaussian random numbers, which represent the sensor random noise. The variance of the sensor random noise is set to 0.01% of the NPSS output at the design condition. The process of generating simulated data is depicted in Figure 42.

<table>
<thead>
<tr>
<th>State Variables (X)</th>
<th>The Observable (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressor efficiency scalar (CE)</td>
<td>Fuel flow (WF)</td>
</tr>
<tr>
<td>Compressor flow scalar (CF)</td>
<td>Engine pressure ratio (EPR)</td>
</tr>
<tr>
<td>Burner pressure drop scalar (BPD)</td>
<td>Exhaust nozzle throat area (A8)</td>
</tr>
<tr>
<td>Turbine efficiency scalar (TE)</td>
<td>Exhaust gas temperature (EGT)</td>
</tr>
<tr>
<td>Turbine flow scalar (TF)</td>
<td>Thrust (FN)</td>
</tr>
<tr>
<td>Duct pressure drop scalar (DPD)</td>
<td></td>
</tr>
<tr>
<td>Nozzle gross thrust coefficient (CFG)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 41: Network of the State Variables and the Observable for the Turbojet Engine

Figure 42: Data Generation Process for the Turbojet Engine Application
7.1.2 A Degraded Compressor Case

The first test case is a turbojet engine with the degraded compressor. Both the compressor efficiency and flow scalars are set to 0.98, which is 2% below the design condition. The simulated data for this test case is shown in Figure 43 along with the baseline value at the design condition and ±3σ lines. The EPR, A8, and FN measurements are mostly within the ±3σ range from the baseline whereas the WF and EGT measurements are beyond the ±3σ range.

The proposed method is capable of analyzing multiple data points together. The multiple data points are obtained at multiple instants of time. A solution from the proposed method varies with the number of data points. To see how the number of data points affects on results, data in various numbers are tested with the proposed method. A thousand Gibbs samples are used for burn-in and three thousand Gibbs samples are generated from three chains with different initial states. Figure 44 shows the trace of the three thousand Gibbs samples from each chain in different colors. The three chains appear to be mixed well, and it can be said that the three chains are converged. Figure 45 shows the posteriors of the compressor flow scalar calculated with various numbers of data points. As the number of data points increases, the posterior probability distribution becomes concentrated around the true value 0.98, which is used for generating the data. The results presented hereafter are from the 50 data points.

Figure 46 shows the posterior distributions of all state variables. All the variables have a unimodal posterior concentrated near their true value except for the compressor and turbine efficiency scalars. The posterior of the compressor efficiency scalar has two modes near the true value, 0.98, and the baseline value, one. The mode at the baseline value is given rise to the models that do not include the compressor efficiency scalar. When the compressor efficiency scalar is not included in the model, it is fixed to the baseline value, one. The posterior of the turbine efficiency scalar has two modes as well near the true value, one, and 0.985. The multimodal distributions suggest the possibility of the multiple solutions for $X_{CE}$ and $X_{TE}$.

Now we have two posterior distributions with two modes. There are four ways of pairing
Figure 43: Simulated Data from a Turbojet Engine with the Degraded Compressor; Dashed Line: the Baseline, Dotted Line: ±3σ, Symbol: data
Figure 44: Convergence History: the Turbojet Engine with the Degraded Compressor
Figure 45: Effect of the Amount of Data on the Posterior of the Compressor Flow Scalar: the Turbojet Engine with the Degraded Compressor; Dashed Line: the True Value

of \((X_{CE}, X_{TE})\): (1, 1), (0.98, 1), (1, 0.985), and (0.98, 0.985). A further investigation of which pair is more likely can be made with correlations between variables. A correlation coefficient between each pair of variables is calculated from the Gibbs samples and listed in Table 7. Figure 47 is a graphical representation of Table 7. The opacity of a line between two nodes is proportional to the degree of correlation. The fainter is the color of the line, the weaker is the correlation. The compressor efficiency scalar is strongly correlated to the turbine efficiency scalar. So is the compressor flow scalar to the turbine flow scalar. The duct pressure drop does not appear to be correlated to any other variables.

The strong correlation is assured by the bivariate scatter plot of the Gibbs samples in the \((X_{CE}, X_{TE})\) coordinate system in Figure 48. According to the bivariate scatter plot, the pair of \(X_{CE}\) and \(X_{TE}\) is rarely likely to be (1, 1) or (0.98, 0.985); no samples are near the two points. In contrast, the rest two pairs, (1, 0.985) and (0.98, 1), are on the cloud of the samples. According to the NPSS simulations on the two pairs, they results in nearly same NPSS outputs, which are different less than 0.7%, when everything else is in the design condition. This insignificant difference can not be distinguished due to the uncertainty in the
Figure 46: Posteriors Distributions of the State Variables: the Turbojet Engine with the Degraded Compressor; Dashed Line: the True Value
Table 7: Correlation Coefficients: the Turbojet Engine with the Degraded Compressor

<table>
<thead>
<tr>
<th></th>
<th>CE</th>
<th>CF</th>
<th>BPD</th>
<th>TE</th>
<th>TF</th>
<th>DPD</th>
<th>CFG</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>1.0000</td>
<td>0.2943</td>
<td>0.1068</td>
<td>-0.9589</td>
<td>-0.1981</td>
<td>-0.0010</td>
<td>-0.1240</td>
</tr>
<tr>
<td>CF</td>
<td>0.2943</td>
<td>1.0000</td>
<td>-0.2446</td>
<td>-0.1353</td>
<td>0.6464</td>
<td>-0.0222</td>
<td>-0.2183</td>
</tr>
<tr>
<td>BPD</td>
<td>0.1068</td>
<td>-0.2446</td>
<td>1.0000</td>
<td>-0.0737</td>
<td>0.1871</td>
<td>0.0155</td>
<td>0.0711</td>
</tr>
<tr>
<td>TE</td>
<td>-0.9589</td>
<td>-0.1353</td>
<td>-0.0737</td>
<td>1.0000</td>
<td>0.2963</td>
<td>0.0074</td>
<td>0.0827</td>
</tr>
<tr>
<td>TF</td>
<td>-0.1981</td>
<td>0.6464</td>
<td>0.1871</td>
<td>0.2963</td>
<td>1.0000</td>
<td>-0.0072</td>
<td>-0.1740</td>
</tr>
<tr>
<td>DPD</td>
<td>-0.0010</td>
<td>-0.0222</td>
<td>0.0155</td>
<td>0.0074</td>
<td>-0.0072</td>
<td>1.0000</td>
<td>0.0200</td>
</tr>
<tr>
<td>CFG</td>
<td>-0.1240</td>
<td>-0.2183</td>
<td>0.0711</td>
<td>0.0827</td>
<td>-0.1740</td>
<td>0.0200</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Figure 47: Correlations Between the State Variables: the Turbojet Engine with the Degraded Compressor
measurements and models. With the five measurements in Table 6, it is hard to distinguish the deterioration in the compressor efficiency from the turbine efficiency.

A decision on which solution is more likely can be made by investigating the posterior of $\gamma$. Figure 49 shows the posteriors of $\gamma$ associated with $X_{CE}$ and $X_{TE}$. While the posterior of $\gamma_{TE}$ is more or less inconclusive, $\gamma_{CE}$ is likely to be one fourth as much as it is likely to be zero. It means that $X_{CE}$ often appears in the models that are supported by the data.

Which models are supported by the data, and which models are not? This question can be answered with the model posterior shown in Figure 50. Nearly half of the models result in zero probability, which means they are not worth further consideration. They are not supported by the data at all. Among the other half only twelve models have a posterior probability above 0.02. Table 8 lists the posterior probabilities of the twelve models and the state variables that each model contains. The twelve models count as much as 75% of the total model posterior probability.
The true model of this case is the one with the compressor efficiency and flow scalars. It is ranked the second after the top ranker with the difference as small as 0.002. The top ranker includes the duct pressure drop scalar in addition to the compressor efficiency and flow scalars. The insignificance of the duct pressure drop scalar prevails in the rest of the models.

The top four models commonly include the compressor efficiency and flow scalars. In contrast, the next eight models commonly include the compressor flow scalar and the turbine efficiency scalar. The two groups of models reassure that multiple solutions may exist and that the multiple solutions are related to the compressor efficiency and flow scalars and the turbine efficiency scalar.

To sum up, the current method found each variable to be the most likely at its true value except the compressor efficiency scalar. The compressor and turbine efficiency scalars were found to be highly correlated. The cross-examination of the bivariate scatter plot, the model posterior, and the posterior of $\gamma$ confirmed that either the compressor efficiency or the turbine efficiency is degraded in addition to the compressor flow and that the compressor
**Figure 50:** Posterior Distribution of the Model Variable $M$: the Turbojet Engine with the Degraded Compressor

**Table 8:** Models with a Posterior Probability above 0.02: the Turbojet Engine with the Degraded Compressor

<table>
<thead>
<tr>
<th>Model No.</th>
<th>Probability</th>
<th>Cumulative Prob.</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>99</td>
<td>0.1317</td>
<td>0.1317</td>
<td>CE CF DPD</td>
</tr>
<tr>
<td>97</td>
<td>0.1299</td>
<td>0.2616</td>
<td>CE CF</td>
</tr>
<tr>
<td>113</td>
<td>0.0938</td>
<td>0.3553</td>
<td>CE CF BPD</td>
</tr>
<tr>
<td>115</td>
<td>0.0887</td>
<td>0.4440</td>
<td>CE CF BPD DPD</td>
</tr>
<tr>
<td>107</td>
<td>0.0621</td>
<td>0.5061</td>
<td>CE CF TE DPD</td>
</tr>
<tr>
<td>123</td>
<td>0.0512</td>
<td>0.5573</td>
<td>CE CF BPD TE DPD</td>
</tr>
<tr>
<td>105</td>
<td>0.0457</td>
<td>0.6030</td>
<td>CE CF TE</td>
</tr>
<tr>
<td>121</td>
<td>0.0364</td>
<td>0.6394</td>
<td>CE CF BPD TE</td>
</tr>
<tr>
<td>57</td>
<td>0.0351</td>
<td>0.6746</td>
<td>CF BPD TE</td>
</tr>
<tr>
<td>41</td>
<td>0.0299</td>
<td>0.7044</td>
<td>CF TE</td>
</tr>
<tr>
<td>59</td>
<td>0.0268</td>
<td>0.7312</td>
<td>CF BPD TE DPD</td>
</tr>
<tr>
<td>43</td>
<td>0.0230</td>
<td>0.7542</td>
<td>CF TE DPD</td>
</tr>
</tbody>
</table>
efficiency is more likely to be degraded than the turbine efficiency is.

So far the capability of the current method has been discussed, but the accuracy has not. It is hard to define a simple quantitative measure of the accuracy for the current method, such as a percentage error, because it deals with probability distributions instead of point estimates. Since the true model exactly describes how data is produced, the accuracy of the current method can be measured from how close its results are to those of the true model. However, the closeness is still a qualitative measure.

Figure 51 shows the posteriors of the compressor efficiency and flow scalars from the current method accompanied by the results from the true and full models. The true model results in the posteriors each of which peaks near the true value. They are the most accurate and the least uncertain. In contrast, the full model results in more disperse posteriors for both scalars than others. The last, but not the least, the current method approximates the true model successfully for both scalars, although it gives rise to a multimodal distribution for the compressor efficiency scalar.

While the simulated data was generated, it was assumed that the sensor noise follows a Gaussian distribution. To see how sensitive the current method is to the error assumption, the current method is tested with the uniform noise and the beta noise as well. The three distributions are shown in Figure 52 and the histograms of simulated data based on the three error assumptions are shown in Figure 69. The data shows the characteristics of each assumptions: the Gaussian distribution is bell-shaped, the uniform distribution has as much probability density near the boundary as its middle value, and the beta distribution is asymmetric. Because the beta distribution is in the range $[0, 1]$, the beta noise is rescaled to the $±3\sigma_ε$ range. Figure 54 shows the posterior distributions of the compressor efficiency and flow scalars with the three error assumptions. Although there is a slight difference in the height and the position of the modes, the shape of the distributions do not change with the error assumption.
Figure 51: Comparison of the True and Full Models, and the BMA Method: the Turbojet Engine with the Degraded Compressor
Figure 52: Three Random Error Assumptions
Figure 53: Noisy Data with the Three Random Error Assumptions in the Degraded Compressor Case of the Turbojet Engine
Figure 54: Posteriors of the Compressor Efficiency and Flow Scalars with the Three Random Error assumptions: the Turbojet Engine with the Degraded Compressor
Table 9: Variable Setting for the Multiple Degraded Components

<table>
<thead>
<tr>
<th>Variables</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{CE}$</td>
<td>0.98</td>
</tr>
<tr>
<td>$X_{CF}$</td>
<td>0.98</td>
</tr>
<tr>
<td>$X_{TE}$</td>
<td>0.98</td>
</tr>
<tr>
<td>$X_{TF}$</td>
<td>0.98</td>
</tr>
<tr>
<td>$X_{CFG}$</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 10: Models with the Eight Highest Posterior Probabilities: the Turbojet Engine with the Degraded Major Components

<table>
<thead>
<tr>
<th>Model No.</th>
<th>Probability</th>
<th>Cumulative Prob.</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>0.2084</td>
<td>0.2084</td>
<td>CE CF BPD TE TF DPD CFG</td>
</tr>
<tr>
<td>112</td>
<td>0.1974</td>
<td>0.4059</td>
<td>CE CF TE TF DPD CFG</td>
</tr>
<tr>
<td>126</td>
<td>0.1710</td>
<td>0.5769</td>
<td>CE CF BPD TE TF CFG</td>
</tr>
<tr>
<td>110</td>
<td>0.1619</td>
<td>0.7388</td>
<td>CE CF TE TF CFG</td>
</tr>
<tr>
<td>46</td>
<td>0.0857</td>
<td>0.8244</td>
<td>CF TE TF CFG</td>
</tr>
<tr>
<td>62</td>
<td>0.0679</td>
<td>0.8923</td>
<td>CF BPD TE TF CFG</td>
</tr>
<tr>
<td>48</td>
<td>0.0574</td>
<td>0.9498</td>
<td>CF TE TF DPD CFG</td>
</tr>
<tr>
<td>64</td>
<td>0.0502</td>
<td>1.0000</td>
<td>CF BPD TE TF DPD CFG</td>
</tr>
</tbody>
</table>

7.1.3 An Extreme Case: the Multiple Degraded Components

In the degraded compress case the true model has two factors, $X_{CE}$ and $X_{CF}$ while the full model has seven factors. Because of the difference in the number of factors, the full model performed poorly and the current method averaging multiple models was better than the full model. What if the full model is close to the true model? Would the current method still perform better than the full model?

Let us examine an extreme case in which compressor, turbine, and nozzle are degraded at the same time. In this case, the true model includes five factors as listed in Table 9.

With the setting fifty simulated data points are generated in the same way as the previous cases. Whereas a half of the models gives rise to the posterior probability above zero in the previous case, only eight models have non-zero probability. The eight models are listed in Table 10 with the variables included in each model. The pressure drop scalars have little effect in the models as in the previous case.

The posteriors of the state variables are shown in Figures 55, 56, and 57 accompanied by the results from the full and true models. In general the current method results in reasonably good approximate of the true model, although the full model does better. The
Table 11: The State Variables and the Observable of the GE 7FA+e Gas Turbine

<table>
<thead>
<tr>
<th>State Variables (X)</th>
<th>The Observable (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressor flow scalar (CF)</td>
<td>Generator output (DW)</td>
</tr>
<tr>
<td>Compressor efficiency scalar (CE)</td>
<td>Compressor discharge temperature (CDT)</td>
</tr>
<tr>
<td>Stage 1 nozzle flow coefficient scalar (TCQ)</td>
<td>Compressor discharge pressure (CDP)</td>
</tr>
<tr>
<td>Turbine efficiency scalar(TE)</td>
<td>Exhaust gas temperature (TEX)</td>
</tr>
<tr>
<td></td>
<td>Fuel flow (WF)</td>
</tr>
<tr>
<td></td>
<td>Air flow (WA)</td>
</tr>
</tbody>
</table>

distributions of the compressor and turbine efficiency scalars have a tall peak at one and 0.97, respectively. The peak at one in the posterior of the compressor efficiency scalar is due to the models that do not include it. The peak at 0.97 in the posterior of the turbine efficiency scalar is due to the strong correlation between the compressor and turbine efficiency scalars. Again, correlations between variables play an important role in analyzing even a univariate posterior distribution.

7.2 Fault Diagnosis of an Industrial Gas Turbine

The second application of the proposed method is the fault diagnosis of an industrial gas turbine. The previous turbojet engine problem takes account for the health parameters of the main components of the turbojet engine. However, the condition of sensors are not considered so that its results are inaccurate when the sensors equipped on the gas turbine are biased.

In this application sensor biases are taken into consideration in addition to the health parameters. The proposed method is tested in three cases, and its results are presented in the following sections.

7.2.1 Problem

The system in question is a GE 7FA+e industrial gas turbine. The gas turbine is modeled with a system of linear equations $Y = AX + B + \varepsilon$. The state variable vector $X$ consists of four health parameters, and the observables $Y$ six measurements. The names of the health parameters and measurements are listed in Table 11. $B$ is a vector of six biases in the measurements. The network structure of the health parameters, measurements, and biases are shown in Figure 58.
Figure 55: Posterior Distributions of the Compressor Related Scalars: the Turbojet Engine with the Degraded Major Components
Figure 56: Posterior Distributions of the Turbine Related Scalars: the Turbojet Engine with the Degraded Major Components
Figure 57: Posterior Distribution of the Nozzle Gross Thrust Coefficient: the Turbojet Engine with the Degraded Major Components

Figure 58: Network Structure of the State Variables and the Bias Variables for the GE 7FA+e Gas Turbine
While a typical sensor bias is independent to other sensor measurements, the bias in the compressor discharge pressure sensor and the exhaust gas temperature sensor affects all the measurement because they are related to the control of the gas turbine. The GE 7FA+e gas turbine has a control mechanism that adjusts fuel flow so that its exhaust gas temperature matches a prescribed value given compressor discharge pressure. The locus of the prescribed exhaust gas temperature value over a range of compressor discharge pressure is called a control curve [65]. When either the compressor discharge pressure sensor or the exhaust gas temperature sensor are biased, the gas turbine does not operate along the control curve. It actually operates at a point deviated either vertically or horizontally, depending on which sensor is biased, from the control curve. Thus, all the measurements are affected by the bias in these sensors.

The stage 1 nozzle flow coefficient is defined as

$$C_q = \frac{1}{K} \frac{W \sqrt{RT}}{A \frac{P}{T}}$$

(65)

where $W$ is the flow to the following stage, $A$ stage 1 nozzle area, $R$ the universal gas constant, $T$ inlet temperature, $K$ a constant [3].

The coefficient matrix $A$ in the system equation is determined in the same way as in the

Figure 59: Control Curve of the GE 7FA+e Gas Turbine
### Table 12: Ranges of the State Variables and the Bias Variables of the GE 7FA+e Gas Turbine

<table>
<thead>
<tr>
<th>Variables</th>
<th>Ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{CF}$</td>
<td>[0.92, 1.02]</td>
</tr>
<tr>
<td>$X_{CE}$</td>
<td>[0.92, 1.02]</td>
</tr>
<tr>
<td>$X_{TCQ}$</td>
<td>[0.92, 1.02]</td>
</tr>
<tr>
<td>$X_{TE}$</td>
<td>[0.92, 1.02]</td>
</tr>
<tr>
<td>$B_{DW}$</td>
<td>[-3%, 3%]</td>
</tr>
<tr>
<td>$B_{CDT}$</td>
<td>[-24°F, 24°F]</td>
</tr>
<tr>
<td>$B_{CDP}$</td>
<td>[-2.5%, 2.5%]</td>
</tr>
<tr>
<td>$B_{TEX}$</td>
<td>[-44°F, 44°F]</td>
</tr>
<tr>
<td>$B_{WF}$</td>
<td>[-10%, 10%]</td>
</tr>
<tr>
<td>$B_{WA}$</td>
<td>[-8%, 8%]</td>
</tr>
</tbody>
</table>

previous turbojet problem. A thermodynamic program GTP (Gas Turbine Performance), developed at GE [41], is used for simulating the gas turbine. The ranges of the state variables and the biases are listed in Table 12. All sensor biases are in percentage of the values at the design condition besides the temperature measurements, which are in deviation from the design condition value.

Simulated data needed for test cases are generated similarly to the previous application. Figure 60 shows the process for generating simulated data. To simulate a faulty condition, GTP is run with the health parameters representing the condition. If a sensor bias presents in that condition, the magnitude of the bias is added to the corresponding GTP output. In addition to that, Gaussian random numbers are added to emulate random sensor noise. The variances of the random numbers are obtained from the test data provided by GE Energy. They are not listed here due to their proprietary nature.

Three test cases are designed so that all the health parameters and the sensor biases can be examined. The three case are a degraded compressor with two biased sensors, a degraded turbine with the biased compressor discharge pressure sensor, and three biases sensors.

#### 7.2.2 Case 1: A Degraded Compressor with the Biased Fuel and Air Flow Sensors

The first case is the diagnosis of a gas turbine with the degraded compressor with two biased sensors. The compressor performance often concerns customers because severe degradation in a compressor can cause fouling. Both compressor efficiency and flow are set to 0.96, which
**Figure 60:** Process for Generating Simulated Data: the GE 7FA+e Industrial Gas Turbine

**Table 13:** Variable Setting for Case of the GE 7FA+e Gas Turbine

<table>
<thead>
<tr>
<th>Variables</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{CF}$</td>
<td>0.96</td>
</tr>
<tr>
<td>$X_{CE}$</td>
<td>0.96</td>
</tr>
<tr>
<td>$B_{WF}$</td>
<td>5%</td>
</tr>
<tr>
<td>$B_{WA}$</td>
<td>4%</td>
</tr>
</tbody>
</table>

is 4% below the design condition. In addition to the compressor degradation, the fuel flow sensor has 5% bias from the design condition value. The variables involved in this case and their values are shown in Table 13.

To find a proper number of data points for the this test case, the fault diagnosis method is applied to the case with various numbers of data points. Figure 61 shows the posteriors of $X_{CF}$. As the number of data points increases, the posterior of $X_{CF}$ becomes taller and narrower near the true value, 0.96. There is a little change in the posterior after the 30 data points. The results shown hereafter are from the 30 data points.

With the 30 data points three chains of three thousand Gibbs samples, total nine thousand samples, are drawn. Figure 62 shows the convergence history of the Gibbs samples. The three chains, which are in three different colors, are mixed well.

Figure 63 shows the posterior distributions of all the state variables and the sensor biases. Among the variables involved in this test case, $X_{CF}$, $B_{WF}$, and $B_{WA}$ have a unimodal posterior distribution. The mode of each distribution is peaked near the corresponding true value. In contrast, the posterior of $X_{CE}$ is multimodal. The mode near one consists of
samples from the models that do not include $X_{CE}$. All other variables are fixed at the baseline value.

Using the nine thousand Gibbs samples, correlation coefficients between each state variables are calculated and listed in Table 14. The graphical representation of the correlation coefficients are shown in Figure 64. $X_{CE}$ and $B_{CDT}$ are strongly correlated, and so are $X_{CF}$ and $X_{TCQ}$ as well as $X_{CF}$ and $B_{WA}$. The bivariate scatter plots of these three pairs shown in Figures 65, 66, and 67 confirm the calculated correlation coefficients.

Since $X_{CE}$ has a multimodal posterior, it needs further investigation. According to Figure 67 $X_{CE}$ is almost linearly proportional to $B_{CDT}$. When $X_{CE}$ is 0.96, $B_{CDT}$ should be 0. On the other hand, when $X_{CE}$ is one, $B_{CDT}$ should be around 20°F. It means that either $X_{CE}$ or $B_{CDT}$ is included in the model while the other is fixed at the baseline value.

Figure 68 shows the posterior distributions of $\gamma$ associated with $X_{CE}$ and $B_{CDT}$. While $\gamma_{CDT}$ is likely to be zero as nearly much as it is one, $\gamma_{CE}$ is likely to be one six times as much as it is likely to be zero. It means that $X_{CE}$ is likely to be included in the model while $B_{CDT}$ is hard to be concluded.
Figure 62: Convergence History of the Gibbs Samples in Case 1 of the GE 7FA+e Gas Turbine
Figure 63: Posterior Distributions of the State Variables and the Bias Variables in Case 1 of the GE 7FA+e Gas Turbine; Dashed Line: the True Value
<table>
<thead>
<tr>
<th></th>
<th>CF</th>
<th>CE</th>
<th>TCQ</th>
<th>TE</th>
<th>BCDP</th>
<th>BTEX</th>
<th>BDW</th>
<th>BCDT</th>
<th>BWF</th>
<th>BWA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF</td>
<td>1.0000</td>
<td>-0.2184</td>
<td>0.8749</td>
<td>-0.2070</td>
<td>-0.2033</td>
<td>0.2128</td>
<td>-0.2628</td>
<td>-0.2625</td>
<td>-0.5301</td>
<td>-0.9339</td>
</tr>
<tr>
<td>CE</td>
<td>-0.2184</td>
<td>1.0000</td>
<td>-0.4424</td>
<td>-0.5044</td>
<td>-0.0426</td>
<td>0.2546</td>
<td>-0.3253</td>
<td>0.9681</td>
<td>0.1430</td>
<td>0.1998</td>
</tr>
<tr>
<td>TCQ</td>
<td>0.8749</td>
<td>-0.4424</td>
<td>1.0000</td>
<td>-0.0964</td>
<td>-0.0003</td>
<td>-0.0417</td>
<td>-0.2897</td>
<td>-0.4380</td>
<td>-0.6731</td>
<td>-0.8130</td>
</tr>
<tr>
<td>TE</td>
<td>-0.2070</td>
<td>-0.5044</td>
<td>-0.0964</td>
<td>1.0000</td>
<td>0.0364</td>
<td>0.2178</td>
<td>0.0129</td>
<td>-0.4863</td>
<td>0.2360</td>
<td>0.1969</td>
</tr>
<tr>
<td>BCDP</td>
<td>-0.2033</td>
<td>-0.0426</td>
<td>-0.0003</td>
<td>0.0364</td>
<td>1.0000</td>
<td>-0.0167</td>
<td>0.0269</td>
<td>-0.0003</td>
<td>0.0045</td>
<td>0.1997</td>
</tr>
<tr>
<td>BTEX</td>
<td>0.2128</td>
<td>0.2546</td>
<td>-0.0417</td>
<td>0.2178</td>
<td>-0.0167</td>
<td>1.0000</td>
<td>0.1840</td>
<td>0.2271</td>
<td>0.5299</td>
<td>-0.1942</td>
</tr>
<tr>
<td>BDW</td>
<td>-0.2628</td>
<td>-0.3253</td>
<td>-0.2897</td>
<td>0.0129</td>
<td>0.0269</td>
<td>0.1840</td>
<td>1.0000</td>
<td>-0.3030</td>
<td>0.6431</td>
<td>0.2502</td>
</tr>
<tr>
<td>BCDT</td>
<td>-0.2625</td>
<td>0.9681</td>
<td>-0.4380</td>
<td>-0.4863</td>
<td>-0.0003</td>
<td>0.2271</td>
<td>-0.3030</td>
<td>1.0000</td>
<td>0.15006</td>
<td>0.2403</td>
</tr>
<tr>
<td>BWF</td>
<td>-0.5301</td>
<td>0.1430</td>
<td>-0.6731</td>
<td>0.2360</td>
<td>0.0045</td>
<td>0.5299</td>
<td>0.6431</td>
<td>0.15006</td>
<td>1.0000</td>
<td>0.4970</td>
</tr>
<tr>
<td>BWA</td>
<td>-0.9339</td>
<td>0.1998</td>
<td>-0.8130</td>
<td>0.1969</td>
<td>0.1997</td>
<td>-0.1942</td>
<td>0.2502</td>
<td>0.2403</td>
<td>0.4970</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Figure 64: Graphical Representation of the Correlation Coefficients: Case 1 of the GE 7FA+e Gas Turbine

Figure 65: Bivariate Scatter Plot of the Gibbs Samples in the \((X_{CF}, B_{WA})\) Coordinate: Case 1 of the GE 7FA+e Gas Turbine
Figure 66: Bivariate Scatter Plot of the Gibbs Samples in the \((X_{CF}, X_{TCQ})\) Coordinate: Case 1 of the GE 7FA+e Gas Turbine

Figure 67: Bivariate Scatter Plot of the Gibbs Samples in the \((X_{CE}, B_{CDT})\) Coordinate: Case 1 of the GE 7FA+e Gas Turbine
The proposed method is tested with uniform noise and beta noise as well as the Gaussian noise. Using these distributions, 50 data points are generated. The histograms of simulated data are shown in Figure 69. The larger number of data points are used to ensure generated noise have the characteristic of each distribution: the uniform distribution has as much probability density near the boundary as its middle value and the beta distribution is asymmetric. The posterior distributions of $X_{CF}$, $X_{CE}$, $B_{WF}$, and $B_{WA}$ are shown in Figures 70, 71, 72, and 73. The three random error assumptions do not make a significant difference in the posteriors of the relevant variables. It can be said that, although the proposed method assumes that random sensor noise follows a Gaussian distribution, it is not sensitive to the random error assumption, at least, to those tested in this case.

Figure 74, 75, 76, and 77 show the posterior distributions of the four relevant variables, $X_{CF}$, $X_{CE}$, $B_{WF}$, and $B_{WA}$, from the current method, which uses Bayesian model averaging, in comparison with the true model and the full model. The true model the posteriors that are peaked near the true value. In contrast, the full model results in the posteriors that are widely distributed over the range. The current method results in less uncertain distributions.

![Figure 68: Posterior Distributions of $\gamma_{CE}$ and $\gamma_{CDT}$ in Case 1 of the GE 7FA+e Gas Turbine](image)
Figure 69: Noisy Data Following Gaussian, Uniform, and Beta Distributions in Case 1 of the GE 7FA+e Gas Turbine
Figure 70: Posterior Distributions of $X_{CF}$ with the Three Random Error Assumptions in Case 1 of the GE 7FA+e Gas Turbine

Figure 71: Posterior Distributions of $X_{CE}$ with the Three Random Error Assumptions in Case 1 of the GE 7FA+e Gas Turbine
**Figure 72:** Posterior Distributions of $B_{WF}$ with the Three Random Error Assumptions in Case 1 of the GE 7FA+e Gas Turbine

**Figure 73:** Posterior Distributions of $B_{WA}$ with the Three Random Error Assumptions in Case 1 of the GE 7FA+e Gas Turbine
Figure 74: Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $X_{CF}$ in Case 1 of the GE 7FA+e Gas Turbine

than the full model does, although it results in a multimodal distribution for $X_{CE}$.

7.2.3 Case 2: An Underfiring Unit with the Degraded Turbine

The gas turbine in this test case is equipped with the degraded turbine and the biased compressor discharge pressure sensor. Due to the turbine degradation the stage 1 nozzle flow coefficient and turbine efficiency are below its design values. In addition to that, the compressor discharge pressure sensor reads value higher than actual pressure.

How the bias in the compressor discharge pressure affect on the operation of a gas turbine is shown in Figure 78. If a positive bias is present in compressor discharge pressure measurements, the gas turbine is controlled to a temperature lower than it is demanded. Consequently, the gas turbine produces less generator output than it is supposed to do. This is called underfiring [49]. On the other hand, if a negative bias is present, the gas turbine is controlled to operate at a temperature higher than it is demanded. This is called overfiring.

The gas turbine in this test case is underfiring because the compressor discharge pressure sensor is positively biased.
Figure 75: Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $X_{CE}$ in Case 1 of the GE 7FA+e Gas Turbine

Figure 76: Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $B_{WF}$ in Case 1 of the GE 7FA+e Gas Turbine
This test case involves $X_{TCQ}$, $X_{TE}$, and $B_{CDP}$, and their values are set as shown in Table 15. With the variable setting 30 simulated data points are generated using GTP and the Gaussian random error assumption.

The 30 data points are analyzed using the proposed method. Figure 79 shows the posterior distributions of the state variables and the sensor biases. All the variables not relevant to this case have a peaked distribution at the baseline value. While the posterior distributions of $X_{TCQ}$ and $B_{CDP}$ are unimodal and are centered near their true value, $X_{TE}$ has a multimodal posterior. One mode is near the true value, 0.96, and the other at the baseline value, one. The mode at the baseline value constitutes of the samples from the models that exclude $X_{TE}$ and fix it at the baseline value. According to the posterior probability of $\gamma_{TE}$ in Figure 80, $X_{TE}$ is likely to be included in a model five times as much.
Figure 78: Effect of the Bias in the Compressor Discharge Pressure Sensor on the GE 7FA+e Gas Turbine Control
Table 16: Variable Setting for Case 3 of the GE 7FA+e Gas Turbine

<table>
<thead>
<tr>
<th>Variables</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_{DW}$</td>
<td>1.5</td>
</tr>
<tr>
<td>$B_{CDT}$</td>
<td>10</td>
</tr>
<tr>
<td>$B_{TE}$</td>
<td>-20</td>
</tr>
</tbody>
</table>

as it is not. The mode at 0.96 is five times as substantial as the other mode at the baseline value.

The current method is compared with the true model and the full model. Figure 81, 82, and 83 show the posterior distributions of the relevant variables, $X_{TCQ}$, $X_{TE}$, and $B_{CDP}$, from the current method accompanied with those from the true and full models. The full model results in the posteriors widely distributed over the range for $X_{TCQ}$ and $X_{TE}$ but the posterior of $B_{CDP}$ fairly close to that of the true model. In contrast, the current method approximates the true model fairly well for all the three variables, although it give rise to an extra mode at the baseline value for $X_{TE}$.

7.2.4 Case 3: An Overfiring Unit with Two Biased Sensors

The gas turbine in this test case is equipped with three biased sensors. These sensors are the generator output, compressor discharge temperature, and exhaust gas temperature sensors. The variables involved in this case are $B_{DW}$, $B_{CDT}$, and $B_{TE}$. The value of each variable is set as shown in Table 16. The exhaust gas temperature measurement is related to the gas turbine control so that bias in the measurements cause underfiring or overfiring of a gas turbine. Figure 84 shows how negative or positive bias in exhaust gas temperature measurements affect the gas turbine control. In this test case the exhaust temperature sensor is biased in the negative direction so that the gas turbine is overfiring. Overfiring of a gas turbine during peak load operation leads to increased component metal temperatures and can impact on hot gas path parts life[5].

Thirty simulated data points are generated similarly to the previous test cases. Give the data points the current method results in the posterior distributions of the state variables and the sensor biases as shown in Figure 85. The probability of $B_{TE}$ is distributed near 25°F, which is 5°F higher than the true value. The probability distributions of $B_{DW}$ and
Figure 79: Posterior Distributions of the State Variables and the Bias Variables in Case 2 of the GE 7FA+e Gas Turbine; Dashed Line: the True Value
Figure 80: Posterior Distribution of $\gamma_{TE}$ in Case 2 of the GE 7FA+e Gas Turbine

Figure 81: Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $X_{TCQ}$ in Case 2 of the GE 7FA+e Gas Turbine
Figure 82: Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $X_{TE}$ in Case 2 of the GE 7FA+e Gas Turbine

Figure 83: Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $B_{CDP}$ in Case 2 of the GE 7FA+e Gas Turbine
(a) Positive Bias in the Exhaust Gas Temperature

(b) Negative Bias in the Exhaust Gas Temperature

**Figure 84**: Effect of the Bias in the Exhaust Gas Temperature Sensor on the GE 7FA+e Gas Turbine Control
$B_{CDT}$ have two modes: one at the true value and the other at the baseline value. The decision on which mode is more likely can be made using the posterior of the corresponding $\gamma$. Figure 86 shows the posterior probabilities of $\gamma_{DW}$ and $\gamma_{CDT}$. Both of $B_{DW}$ and $B_{CDT}$ are likely to be included in a model more than four times as much as they are not.

Figures 87, 88, and 89 show the posterior distributions of $B_{TEX}$, $B_{DW}$, and $B_{CDT}$ resulting from the current method, denoted as “BMA” in the figures, in comparison with the posteriors from the true and full models. In the three figures the true model results in a peaked distribution around the true value of each variable. In contrast, the full model gives rise to the probability density distributed over the whole range of each variable so that they do not provide much information to the analyst. In addition, their peak point, which is a maximum a posterior (MAP) estimate of the corresponding variable, does not quite match that of the true model. The last, but not the least, the current method results in more informative distributions than the full model does. Furthermore, if a MAP estimate is made for each variable, it is much accurate than the MAP estimate of the full model.
Figure 85: Posterior Distributions of the State Variables and the Biase Variables in Case 3 of the GE 7FA+e Gas Turbine; Dashed Line: the True Value
Figure 86: Posterior Probabilities of $\gamma$ in Case 3 of the GE 7FA+e Gas Turbine
**Figure 87:** Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $B_{TEX}$ in Case 3 of the GE 7FA+e Gas Turbine

**Figure 88:** Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $B_{DW}$ in Case 3 of the GE 7FA+e Gas Turbine
Figure 89: Comparison of the True and Full Models, and the Bayesian Model Averaging Method: $B_{CDT}$ in Case 3 of the GE 7FA+e Gas Turbine
Chapter VIII

CONCLUSIONS

In this dissertation an engineering problem has been posed that requires fundamental understanding of underlying statistical and mathematical principles. As a result, a variety of topics has been discussed, including fault diagnosis, inverse problems, and statistical approaches for inverse problems. This discussion has led to several philosophical notions in statistics such as the model uncertainty and the principle of insufficient reasons in fault identification. This chapter summarizes the suggested answers to the research questions, contributions to the current work to the general understanding of the underlying phenomena, along with the limitations of the proposed approach and the directions of the future work.

8.1 Review of Research Questions

In Chapter 4 the three research questions were raised and attempted to be answered by means of illustrating the proposed method in two applications. Let us revisit each research question and discuss how the results from the two applications are relevant to each question.

The first question is:

How can we identify faults with data containing error?

The data we used in the first application contain random noises and the data in the second application contain sensor biases as well. In spite of these errors, the proposed method successfully identified the implanted faults with limited data in all test cases. In particular, in the case of industrial gas turbines, the proposed method successfully separated physical faults in the components of the gas turbine from the sensor biases.

In addition, the proposed method was tested with the three random error assumptions in a test case in each application. In the test cases, random noises are generated from three different distributions: a Gaussian, uniform, and beta distributions. In spite of a
large variance of noises from the uniform distribution and asymmetric noises from the beta
distribution, the performance of the proposed method was not substantially deteriorated.

How can we identify multiple solutions of fault identification if they are present?

In the test case of the turbojet engine with the degraded compressor, the proposed method
successfully identified the two solutions that give rise to nearly the same measurement pat-
tern. Furthermore, the method reported one solution is more likely to the other, and the
former was the implanted fault in the data.

How can we take account for the ambiguity in selecting variables that need to
be identified?

The proposed method accounts for the model uncertainty by using multiple models and
averaging them. In all the test cases from the two applications, the Bayesian model averaging
(BMA) method approximated the true model reasonably and performed better than the full
model: the results were less uncertain. The BMA method can robustly analyze data when
the true model is not known, which is often the case in the real world problems.

8.2 Summary of Contributions

A primary academic contribution of this dissertation is the introduction of the model un-
certainty in fault identification. Fault identification is formulated as a statistical inference
problem. In typical inference problems a statistical model is built, and the probabilities of
unknown variables are updated once observable variables are instantiated. The more is the
statistical model close to the true mechanism with which the data is produced, the more
accurate is the inference. However, in most cases how the data is produced is hardly known
beforehand. When the statistical model is a complex model including a large number of
variables, it will not leave out necessary variables. On the other hand, unnecessary variables
included in the model gives rise to an unfavorable trend in solutions, the smearing effect.
In order for a fault identification algorithm to be robust as well as to be accurate in various
fault situations, multiple models with various complexity have to be considered and logically
incorporated. For this dissertation results from multiple models are averaged according to how much each model is likely to be true.

This general principle is implemented in a specific method described in Chapter 6. The method assumes that observable variables follow a multivariate normal distribution. The mean of the multivariate normal distribution is a linear function of unknown state variables and sensor biases. The precision matrix of the multivariate normal distribution is also treated as a variable which uses a Wishart distribution as the prior distribution. The unknown state variables and sensor biases use a uniform distribution as their prior distribution. When the observable variables are instantiated, the probabilities of the state variables and biases are updated. The updated probabilities are estimated using Gibbs sampling.

The last, but not the least, contribution is the various ways in which the results of the method are interpreted. Since results of this method are statistical and complex, it is critical to provide means for their visual representation. One of the examples is a graph representation of correlations. In the graph the state variables and biases are plotted as nodes. A correlation between two variables is mapped to the opacity of the line between the two nodes so that a line is in a fainter color as the corresponding correlation is weaker. This graph helps understand the univariate posterior probability of each variable when it is multimodal.

8.3 Comparison with Other Methods

The characteristics of the method is shown in Table 17 compared with other methods. Each method as a rule has multiple variations For example, the original Kalman filter conceived by Kalman led to the development of extended Kalman filter and unscented Kalman filter. Each variation uses different assumptions and is applicable to different problems so that it is hard to generalize the characteristics of all the variational forms. The table is a compilation of commonly accepted characterization collected from literature [78, 59].

The current method is based on mathematically sound probability theory. Unlike the weighted least squares (WLS) method and Kalman filters, which result in maximum a posterior estimates, the current method deals with probability distributions. Thus, it has a
### Table 17: Comparison of the Current Method with Others

<table>
<thead>
<tr>
<th></th>
<th>The Current Method</th>
<th>Kalman Filters</th>
<th>Genetic Algorithm</th>
<th>Weighted Least Squares</th>
<th>Fuzzy Logic</th>
<th>Neural Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Solutions</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solution Confidence</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonlinearity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Time Dependency</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Discretization</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

* When nonlinear Kalman filters or a nonlinear weighted least squares method is used.
capability to identify multiple solutions in the solution space. Genetic algorithm (GA) also has that capability only if it keeps track of population in each iteration. In general, GA gives rise to the solution with the highest fitness. Since the current method is probabilistic, the confidence of a solution can be obtained naturally.

The current method uses linear models so that it cannot capture nonlinearity. The original Kalman filter is basically a linear model. An extended Kalman filter and unscented Kalman filter are two extensions of the original one to take nonlinearity into consideration. The WLS method can be used for either linear or nonlinear problems.

Like most other methods the current method can analyze data from a steady state. In contrast, Kalman filters can be interpreted as a special case of dynamic Bayesian networks. It takes into consideration a state change of a system with time.

The current method, as well as Kalman filters and the WLS method, does not require the discretization of continuous variables, which can cause several issues described in Section 5.1.1.

8.4 Potential Applications

During the development of the proposed method, it has been tried to maintain the generality as much as possible. For instance, a diffuse prior distribution is used for the covariance matrix of the multivariate normal distribution, which observable variables follow. The covariance matrix, which is hard to define in many cases, do not have to be specified beforehand. As a consequence of this effort, the proposed method is envisioned to solve a wide range of problems, which can be modeled with or, at least, can be reasonably approximated with a system of linear equations. The object in such problems are, but not limited to, mechanical systems, and industrial processes or plants.

One specific example is the leakage detection of industrial plants. Most industrial plants consist of several components. The mass of the plants is conserved as long as it does not leak out of the plants to the environment. The law of mass conservation is expressed with a system of linear equations. Consider a chemical plant consisting of five components as
Figure 90: An Industrial Plant Comprising Five Components

shown in Figure 90. The mass conservation equations are

\[
\begin{align*}
  m_1 - m_2 - m_3 - m_4 &= \varepsilon_1 \\
  m_1 - m_5 &= \varepsilon_2 \\
  m_3 - m_4 &= \varepsilon_3 \\
  m_2 + m_3 + m_4 - m_5 &= \varepsilon_4
\end{align*}
\]

where \( m_i \) is the mass measured at the component \( i \) and \( \varepsilon_j \) is a random error in the equation \( j \). The leakages between each component can be added to Equation 66 as follows:

\[
\begin{align*}
  (m_1 + b_1) - (m_2 + b_2) - (m_3 + b_3) - (m_4 + b_4) &= \varepsilon_1 \\
  (m_1 + b_1) - (m_5 + b_5) &= \varepsilon_2 \\
  (m_3 + b_3) - (m_4 + b_4) &= \varepsilon_3 \\
  (m_2 + b_2) + (m_3 + b_3) + (m_4 + b_4) - (m_5 + b_5) &= \varepsilon_4
\end{align*}
\]

where \( b_i \) is a bias of the sensor \( i \). The equation can be written in the matrix form as follows:

\[
\mathbf{m} + \mathbf{b} + \varepsilon = 0
\]

It should be noted that in this problem a leakage cannot be differentiated from the bias in the corresponding sensor measurements.
8.5 Future Work

There is a room for improvement in many directions. One of the directions is the identifiability of variables. Throughout the two applications, it was noticed that some variables are easier to identify than others. However, no attempts were made to improve the identifiability. The coefficient matrix is one of the factors determining the identifiability. The coefficient matrix depends on sensor selection. Different sets of observable variables give rise to different coefficient matrices. It may be possible to find the best coefficient matrix among the matrices. The optimization requires a criterion that defines the identifiability quantitatively.

There are several topics that are not directly regarding to the current work but are supplementary in a sense that the progress in those areas will broaden the application of the proposed method.

The current method can analyze data produced from a system in a steady state. If mixed data from more than one state is fed into the current method, it will identify a mixed state, which may be none of the states. Therefore, a pre-process is needed to sort out and prepare data for the current method.

As well as the pre-process, an intelligent post-process will be beneficial to improve the practicality of the current method. The results from the current method is so highly statistical that it is hard to translate into an engineering language or to use in decision making. The intelligent post-process will fill the gap between a statistician and an engineer.

8.6 Closing Remarks

In the early phase of this work the author was overwhelmed by the width of fault diagnosis. It is a broad topic consisting of various types of small problems. The author has pursued a depth in one problem rather than cover several problems and has tried to achieve an incremental improvement from the existing state of the art.

This dissertation is effectively a bridge between statistics and engineering. It was amusing to find that surprisingly many techniques in the engineering field are rooted in statistics. The author has a strong belief that there are still many engineering problems where statistics
can help and would like to encourage other engineers to look for an opportunity to marry statistics and engineering.
Appendix A

GAUSSIAN BAYESIAN NETWORK FOR THE QUADRUPLE TANK PROCESS

% Quadruple-Tank Process
% Gaussian Bayesian Network

clear all; close all

A1 = 28; A2 = 32; A3 = A1; A4 = A2;
a1 = 0.071; a2 = 0.057; a3 = a1; a4 = a2;
k1 = 3.33; k2 = 3.35;
g = 981; g2 = (2*g)^0.5;
gamma1 = 0.7; gamma2 = 0.6;
h10 = 12.4; h20 = 12.7; h30 = 1.8; h40 = 1.4;
T1 = A1/a1*(2*h10/g)^0.5;
T2 = A2/a2*(2*h20/g)^0.5;
T3 = A3/a3*(2*h30/g)^0.5;
T4 = A4/a4*(2*h40/g)^0.5;

w = zeros(4,2);
w(1,:) = [T1*gamma1*k1/A1 T1*(1-gamma2)*k2/A1];
w(2,:) = [T2*(1-gamma1)*k1/A2 T2*gamma2*k2/A2];
w(3,:) = [0 T3*(1-gamma2)*k2/A3];
w(4,:) = [T4*(1-gamma1)*k1/A4 0];
U1 = 1; U2 = 2; B1 = 3; B2 = 4; B3 = 5; B4 = 6;
Y11 = 7; Y21 = 8; Y31 = 9; Y41 = 10; Y51 = 11;
Y61 = 12; Y71 = 13; Y81 = 14; Y91 = 15; Y101 = 16;
Y111 = 17; Y121 = 18; Y131 = 19; Y141 = 20; Y151 = 21;
Y161 = 22; Y171 = 23; Y181 = 24; Y191 = 25; Y201 = 26;
Y12 = 27; Y22 = 28; Y32 = 29; Y42 = 30; Y52 = 31;
Y62 = 32; Y72 = 33; Y82 = 34; Y92 = 35; Y102 = 36;
Y112 = 37; Y122 = 38; Y132 = 39; Y142 = 40; Y152 = 41;
Y162 = 42; Y172 = 43; Y182 = 44; Y192 = 45; Y202 = 46;
Y13 = 47; Y23 = 48; Y33 = 49; Y43 = 50; Y53 = 51;
Y63 = 52; Y73 = 53; Y83 = 54; Y93 = 55; Y103 = 56;
Y113 = 57; Y123 = 58; Y133 = 59; Y143 = 60; Y153 = 61;
Y163 = 62; Y173 = 63; Y183 = 64; Y193 = 65; Y203 = 66;
Y14 = 67; Y24 = 68; Y34 = 69; Y44 = 70; Y54 = 71;
Y64 = 72; Y74 = 73; Y84 = 74; Y94 = 75; Y104 = 76;
Y114 = 77; Y124 = 78; Y134 = 79; Y144 = 80; Y154 = 81;
Y164 = 82; Y174 = 83; Y184 = 84; Y194 = 85; Y204 = 86;

N = 86;

dag = zeros(N,N);
dag(U1, [Y11:Y201]) = 1; dag(U2, [Y11:Y201]) = 1;
dag(U1, [Y12:Y202]) = 1; dag(U2, [Y12:Y202]) = 1;
dag(U2, [Y13:Y203]) = 1; dag(U1, [Y14:Y204]) = 1;
dag(B1, [Y11:Y201]) = 1; dag(B2, [Y12:Y202]) = 1;
dag(B3, [Y13:Y203]) = 1; dag(B4, [Y14:Y204]) = 1;

onodes = [Y11:Y204]; unodes = [U1 U2 B1:B4];

mu = cell(1,N);
sd = cell(1,N);
for i = 1:N,
    mu{i} = 0;
end

sd_u = 0.01;
for i = [U1 U2],
    sd{i} = sd_u;
end

sd_b = 0.01;
for i = [B1 B2 B3 B4],
    sd{i} = sd_b;
end

sd_y = 0.01;
for i = [Y11:Y204],
    sd{i} = sd_y;
end

w = cell(1,N);
w{Y11} = [T1*gamma1*k1/A1 T1*(1-gamma2)*k2/A1 1 0 0 0];
w{Y12} = [T2*(1-gamma1)*k1/A2 T2*gamma2*k2/A2 0 1 0 0];
w{Y13} = [0 T3*(1-gamma2)*k2/A3 0 0 1 0];
w{Y14} = [T4*(1-gamma1)*k1/A4 0 0 0 0 1];

for i = Y21:Y201,
    w{i} = w{Y11};
end
for i = Y22:Y202,
    \( w(i) = w(Y12); \)
end

for i = Y23:Y203,
    \( w(i) = w(Y13); \)
end

for i = Y24:Y204,
    \( w(i) = w(Y14); \)
end

LB = \mu(U1) - 2*sd(U1); \ UB = \mu(U1) + 2*sd(U1);

dx = (UB-LB)/101;
x = LB:dx:UB;
x0 = x;
y0 = \text{normpdf}(x, \mu(U1), sd(U1));

LB = \mu(B1) - 2*sd(B1); \ UB = \mu(B1) + 2*sd(B1);

dx = (UB-LB)/101;
x = LB:dx:UB;
xb0 = x; \ yb0 = \text{normpdf}(x, \mu(B2), sd(B2));

data = [
    0.0702 -0.0672 -0.0232 0.0201 \\
    0.0714 -0.0420 -0.0297 0.0288 \\
    0.0753 -0.0534 -0.0316 0.0329 \\
    0.0727 -0.0698 -0.0464 0.0247 \\
    0.0657 -0.0492 -0.0441 0.0304 \\
    0.0632 -0.0518 -0.0345 0.0269 \\
    0.0633 -0.0602 -0.0304 0.0377
]
ndata = size(data, 1);
evidence = cell(1,N);
names = cell(1,N);
names = {'U1' 'U2' 'B1' 'B2' 'B3' 'B4' 'Y1' 'Y2' 'Y3' 'Y4'};

nunodes = length(unodes); nonodes = length(onodes);

mu1 = zeros(1, nunodes); mu2 = zeros(1, nonodes);

Sigma = zeros(N,N);
Sigma11 = zeros(nunodes, nunodes);
Sigma22 = zeros(nonodes, nonodes);
Sigma12 = zeros(nunodes, nonodes);
Sigma21 = zeros(nonodes, nunodes);
b = cell(1,N);
for i = unodes,
    b{i} = zeros(1, i-1);
end

for j = onodes,
    k = j - nunodes - 1;
    b{j} = [w{j} zeros(1,k)]’;
end

Sigma(1,1) = sd{1}^2;
for j = 2:N,
    beta = b{j}

    for i = 1:j-1,
        sum = 0;

        for k = 1:j-1,
            sum = sum + Sigma(i,k)*beta(k);
        end

        Sigma(j,i) = sum;
        Sigma(i,j) = sum;
    end

    sum = 0;

    for k = 1:j-1,
sum = sum + Sigma(j,k)*beta(k);
end

Sigma(j,j) = sd{j}^2 + sum;
end

Sigma11 = Sigma(unodes, unodes); Sigma22 = Sigma(onodes, onodes);
Sigma12 = Sigma(unodes, onodes); Sigma21 = Sigma(onodes, unodes);

j = 1;
for i = Y11:Y201,
evidence{i} = data(j,1);
j = j + 1;
end

j = 1;
for i = Y12:Y202,
evidence{i} = data(j,2);
j = j + 1;
end

j = 1; for i = Y13:Y203,
evidence{i} = data(j,3);
j = j + 1;
end

j = 1; for i = Y14:Y204,
evidence{i} = data(j,4);
j = j + 1;
end

e   = [evidence{onodes}]';
mu1 = [mu{unodes}]'; mu2 = [mu{onodes}]';
mu1 = mu1 + Sigma12*inv(Sigma22)*(e-mu2);
Sigma11 = Sigma11 - Sigma12*inv(Sigma22)*Sigma21;

mu{U1} = mu1(1); mu{U2} = mu1(2);
sd{U1} = Sigma11(1,1)^0.5; sd{U2} = Sigma11(2,2)^0.5;

for i = unodes,
    mu{i} = mu1(i);
    sd{i} = Sigma11(i,i)^0.5;
end

for i = unodes,
    fprintf('
%s 	 mean = %f, sd = %f', names{i}, mu{i}, sd{i});
end

x = cell(1,N); y = cell(1,N);
for i = unodes,
    LB = mu{i} - 4*sd{i};
    UB = mu{i} + 4*sd{i};
    dt = (UB-LB)/101;
    t = LB:dt:UB;
    x{i} = t;
    y{i} = normpdf(t, mu{i}, sd{i});
end
% Plot for U1
figure; setplotsize([3.25 3]); setplotfont('Times', 12);
xt = [0.03 0.03]; yt = [0 700];
plot(x0, y0,'-.', x{1}, y{1}, xt, yt, ':')
legend('Prior', 'Posterior', 'True Value', 2);
xlabel('\it{X_1}'); ylabel('Probability Density');
axis([-0.02 0.04 0 700]);
set(gca, 'Box', 'on');
exportfig(gcf,'tank_case2_x1.eps', 'bounds', 'tight', 'color', 'cmyk');
previewfig(gcf,'bounds','tight', 'color', 'cmyk');

% Plot for U2
figure; setplotsize([3.25 3]); setplotfont('Times', 12);
xt = [-0.03 -0.03]; yt = [0 700];
plot(x0, y0,'-.', x{2}, y{2}, xt, yt, ':')
legend('Prior', 'Posterior', 'True Value', 1)
xlabel('\it{X_2}'); ylabel('Probability Density');
axis([-0.04 0.02 0 700]);
set(gca, 'Box', 'on');
exportfig(gcf,'tank_case2_x2.eps', 'bounds', 'tight', 'color', 'cmyk');
previewfig(gcf,'bounds','tight', 'color', 'cmyk');

% Plot for B2
figure; setplotsize([3.25 3]); setplotfont('Times', 12);
xt = [0.03 0.03]; yt = [0 700];
plot(xb0, yb0,'-.', x{4}, y{4}, xt, yt, ':');
legend('Prior', 'Posterior', 'True Value', 1);
xlabel('\it{B_2}'); ylabel('Probability Density')
set(gca, 'Box', 'on' );
exportfig(gcf,'tank_case2_b2.eps', 'bounds', 'tight', 'color', 'cmyk');
previewfig(gcf,'bounds','tight', 'color', 'cmyk');
Appendix B

MATHEMATICA CODE FOR CORRELATION GRAPHS

Needs["MultivariateStatistics'"]; Needs["HypothesisTesting'"]; Needs["Combinatorica'"]; Needs["GraphUtilities'"]; rawData = Import["D:\Program Files\Wolfram Research\Mathematica\6.0\Work\data.txt", "Table"]; {rowCount, columnCount} = Dimensions[rawData];

n = 9000; data = rawData; corr = Table[ SetAccuracy[N[Correlation[data[[All, i]], data[[All, j]]]], 5], {i, columnCount}, {j, columnCount}];

m = MatrixForm[corr, TableHeadings -> {{CE, CF, BPD, TE, TF, DPD, CFG}, {CE, CF, BPD, TE, TF, DPD, CFG}}]; Export["corr_table.eps", m, "EPS"]; adjMatrix = {{0, 1, 1, 1, 1, 1, 1}, {1, 0, 1, 1, 1, 1, 1}, {1, 1, 0, 1, 1, 1, 1}, {1, 1, 1, 0, 1, 1, 1}, {1, 1, 1, 1, 0, 1, 1}, {1, 1, 1, 1, 1, 0, 1}, {1, 1, 1, 1, 1, 1, 0}}; g = FromAdjacencyMatrix[adjMatrix]; e = Edges[g];
vertexLabel = {"CE", "CF", "BPD", "TE", "TF", "DPD", "CFG"};
opacity = Table[0, {i, 1, M[g]}];

For[k = 1, k <= M[g], k++, e1 = e[[k]]; i = e1[[1]]; j = e1[[2]]; opacity[[k]] = Abs[corr[[i, j]]]]

edgeColor = Table[Hue[0.25, 1, 0, opacity[[i]]], {i, 1, M[g]}];

g = ShowGraph[g, Table[{e[[i]], EdgeColor -> edgeColor[[i]]}, {i, 1, M[g]}], VertexLabel -> vertexLabel, VertexColor -> LightGray, VertexStyle -> Disk[0.1], VertexLabelPosition -> Center]

Export["corr_graph.eps", g, "EPS"]
Appendix C

THE TURBOJET ENGINE MODEL

The linear regression equations created from the NPSS simulations are

\[ Y = \beta_0 + \beta X + \varepsilon \]

where \( Y = \begin{bmatrix} WF \\ EPR \\ A8 \\ EGT \\ FN \end{bmatrix} \), \( X = \begin{bmatrix} CE \\ CF \\ BPD \\ TE \\ TF \\ DPD \\ CFG \end{bmatrix} \), and \( \beta_0 = \begin{bmatrix} -0.0685263 \\ -4.8122023 \\ 319.314058 \\ 2394.81151 \\ -15991.802 \end{bmatrix} \).

See the next page for \( \beta \).

The variance of the model fit error of each equation is

\[
\sigma_\varepsilon = \begin{bmatrix}
7.386811E-3 \\
3.105501E-3 \\
4.1291905E-2 \\
2.4555748E-1 \\
1.00863885
\end{bmatrix}
\]
\[ \beta = \begin{bmatrix} 
0.74831779 & -7.5593516 & -0.6745633 & -0.0008271 & 10.5581276 & 8.69E - 6 & 0.00027539 \\
402.161866 & -5673.5021 & -362.28258 & -0.4764514 & 5672.39095 & -0.0297208 & 0.128401 \\
4170.35025 & -9549.6964 & -1429.0348 & 3323.71422 & 19534.5949 & -58.443077 & 10317.6387 
\end{bmatrix} \]
REFERENCES


