Real-Time Monitoring and Fault Diagnostics in Roll-to-Roll Manufacturing Systems

by

Huanyi Shui

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mechanical Engineering)
in The University of Michigan
2018

Doctoral Committee:

Professor Jun Ni, Chair
Professor L. Jay Guo
Assistant Professor Xiaoning Jin, Northeastern University
Professor Jeffrey Stein

Huanyi Shui
huanyis@umich.edu
ORCID iD: 0000-0001-7353-7210

To my family

ACKNOWLEDGEMENTS

Looking back to my journey toward the Doctoral degree, I realized how lucky I am to have so many people help, support, encourage and believe in me. The most precious things I learnt are not about research, but the persistence, humbleness, and self-confidence. I cannot make any achievements without the help of those individuals, to whom I would like to express my deepest respect and gratitude.

First, my foremost gratitude goes to my advisor Professor Jun Ni for offering me the opportunity to pursue my Ph.D. degree at the University of Michigan. During the period of my master and Ph.D. study, he not only provided continuous encouragement, patient guidance, insightful comments on my research work, but also cultivated me to become a better researcher as well as helped me to improve my communication and leadership skills. Without him, I would not think of pursuing my Ph.D. nor accomplishing this dissertation. I would also like to thank Professor Xiaoning Jin, who closely supervised me with my research work and sharpened my research skills. My sincere thanks also go to my dissertation committee members - Professor Jeffrey Stein, Professor and L. Jay Guo, who are willing to spend their precious time serving on my dissertation committee and providing valuable feedbacks to help me improve my research work. I also want to thank Professor Matthew Plumlee for serving on my preliminary exam committee.

Second, I would like to thank those industrial people with whom I have collaborated during my Ph.D., especially Stuart Lebowitz, Steve Lange, Todd Lenser, Clarissa Maldonado, Justin Lawler, Andy Palmer, etc. from Procter & Gamble, Jinhyuk Jung and Ivan Lee from

Samsung Electro Mechanics, and Paul E Krajewski, Yilu Zhang, Shiming Duan, Chaitanya Sankavaram, etc. from General Motor. Their expert knowledge and industrial experience helped me to have a better understanding of real-world problems and inspired ideas in my dissertation.

Third, I am very grateful for my co-workers and friends in the S. M. Wu Manufacturing Research Center for their help and support, especially George Qiao, Dr. Shuhuai Lan, Dr. Yang Li, Dr. Xi Gu, Dr. Xin Weng, Dr. Xun Liu, Dr. Baoyang Jiang, Dr. Xinran Liang, Yangbing Lou, Hao Lei, Kai Chen, Kevin Wilt, Xinjian Lai, Zhiyi Chen, Lei Sun, Tianchen Qiu, Xin Hu, and Yossi Cohen, as well as my other friends who made my life here vibrant and exciting, especially Yuxi Zhang, Yining Lu, Qianning Zhang, Yi Cheng, Qinchun Zhao, Mengtian Zhang, Xiao Jing, Michelle Kang, Wubing Qin, Weiyu Cao, Menglong Duan, Sicheng He, Ben Schoenfeldt Dr. Weisi Li, Dr. Jongsoo Choi, Dr. Xinyi Ge, Dr. Heng Kuang, Dr. Li Jiang, and Dr. Hao Yu.

Finally, I would like to thank my parents Ruilong Shui and Linfei Zhu. They always encourage me and convince me that I don't have to be fear of facing any challenges. With their unconditional support and love, I am able to grow confidently and fight through difficulties over the years. I also appreciate my Uncle Shousong Shui, Aunt Liping Zheng, cousin Huanhuan Shui and her husband Qunbi Zhuge, who provide me another warm home outside China.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

A roll-to-roll (R2R) process is a manufacturing technique involving continuous processing of a flexible substrate as it is transferred between rotating rolls. It integrates many additive and subtractive processing techniques to produce rolls of product in an efficient and cost-effective way due to its high production rate and mass quantity. Therefore, the R2R processes have been increasingly implemented in a wide range of manufacturing industries, including traditional paper/fabric production, plastic and metal foil manufacturing, flexible electronics, thin film batteries, photovoltaics, graphene films production, etc. However, the increasing complexity of R2R processes and high demands on product quality have heightened the needs for effective real-time process monitoring and fault diagnosis in R2R manufacturing systems.

This dissertation aims at developing tools to increase system visibility without additional sensors, in order to enhance real-time monitoring, and fault diagnosis capability in R2R manufacturing systems. First, a multistage modeling method is proposed for process monitoring and quality estimation in R2R processes. Product-centric and process-centric variation propagation are introduced to characterize variation propagation throughout the system. The multistage model mainly focuses on the formulation of process-centric variation propagation, which uniquely exists in R2R processes, and the corresponding product quality measurements with both physical knowledge and sensor data analysis. Second, a nonlinear analytical redundancy method is proposed for sensor validation to ensure the accuracy of sensor measurements for process and quality control. Parity relations based on nonlinear observation

matrix are formulated to characterize system dynamics and sensor measurements. Robust optimization is designed to identify the coefficient of parity relations that can tolerate a certain level of measurement noise and system disturbances. The effect of the change of operating conditions on the value of the optimal objective function – parity residuals and the optimal design variables – parity coefficients are evaluated with sensitivity analysis. Finally, a multiple model approach for anomaly detection and fault diagnosis is introduced to improve the diagnosability under different operating regimes. The growing structure multiple model system (GSMMS) is employed, which utilizes Voronoi sets to automatically partition the entire operating space into smaller operating regimes. The local model identification problem is revised by formulating it into an optimization problem based on the loss minimization framework and solving with the mini-batch stochastic gradient descent method instead of least squares algorithms. This revision to the GSMMS method expands its capability to handle the local model identification problems that cannot be solved with a closed-form solution.

The effectiveness of the models and methods are determined with testbed data from an R2R process. The results show that those proposed models and methods are effective tools to understand variation propagation in R2R processes and improve estimation accuracy of product quality by 70%, identify the health status of sensors promptly to guarantee data accuracy for modeling and decision making, and reduce false alarm rate and increase detection power under different operating conditions. Eventually, those tools developed in this thesis contribute to increase the visibility of R2R manufacturing systems, improve productivity and reduce product rejection rate.

CHAPTER 1 INTRODUCTION

1.1 Motivation

The roll-to-roll (R2R) process is a family of manufacturing techniques involving continuous processing of a flexible substrate as it is transferred between rotating rolls (shown in Figure 1-1). It can integrate many process techniques to produce rolls of product in an efficient and cost-effective way due to its high production rate and mass quantity. High throughput and low piece cost are main reasons that more industries prefer R2R manufacturing to conventional manufacturing. High-throughput R2R manufacturing provides greater manufacturing economy of scale even though the required initial capital investments are higher than traditional printing/patterning systems (US Dept of Energy, 2013). Moreover, due to its flexibility of integrating additive or subtractive processes in a continuous manner, R2R processes have been widely implemented in manufacturing industries to fabricate paper, plastic, and metal foil with high throughput. In addition, R2R has great potential to achieve energy-efficient, low environmental impact, and cost-effective production in emerging industries such as flexible electronics, thin film batteries, solar panels, and graphene films. It is expected that the market value of those products will reach $44 billion in 2021 and will continue to grow significantly in future years (Das and Harrop, 2011).

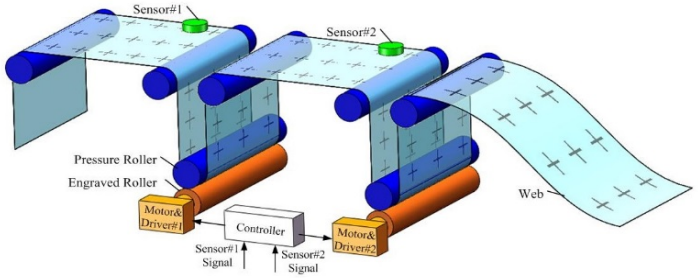


Figure 1-1: An example of a roll-to-roll manufacturing system

The successful deployment of an R2R process depends on the quality of web-based product integration across different upstream supply lines. These web-based products are typically assembled laminates combining several continuous layers with discrete components that must be shaped, placed or printed accurately and repeatedly with tight tolerance at a high speed on flexible and extensible substrates (Department of Energy, 2015). Therefore, any process disruptions, equipment defects or sensor malfunctions can result in a considerable yield loss. In particular, machine or tool failures will cause defects on the product directly and malfunctioning sensors might provide wrong feedback for the control system and lead to plant state variables beyond acceptable limits. Moreover, due to the continuous characteristics of R2R processes, the final product quality is not only affected by the variation at a local operation step but also variations transferred from upstream operations. Therefore, a small quality deviation from an intermediate operation might be built up and eventually generate nonconforming products with downstream operations. Furthermore, unscheduled shutdowns and restarts in an R2R process usually cause a large number of defective products because the system dynamics during ramp up/ramp down periods significantly differs from that during its steady-state operation. Therefore, real-time process monitoring and early diagnosis of failures (e.g., sensor faults) while the system is still under its controllable region are critical to minimize adverse

effects of anomalies, assure product quality, and reduce/prevent productivity loss.

In order to facilitate process and quality control, there are numerous research studies focusing on developing and improving sensor/inspection techniques and related control systems to enhance real-time monitoring capability in high-throughput R2R manufacturing systems. High-resolution and high-speed metrology systems have been developed for web tension and speed monitoring, defect detection (e.g., missing pattern or static buildup errors), surface roughness measurement, registration (pattern position), etc. (Subbaraman et al., 2012). The techniques include but are not limited to load cell sensor, non-contact optical measurements, laser sensing, electroluminescence imaging and light beam induced current mapping (Ulsh, 2014). Advanced sensing and inspection techniques provide rich information for process and quality control. However, they are usually not applicable if misapplied to an entire system due to the cost, processing speed, and the difficulty of physical installation when space is strictly constrained. Therefore, current R2R systems are not fully observable due to the lack of proper cost-effective sensors and inspection systems. Limited system visibility hinders the assessment of intermediate product quality and increases challenges of real-time diagnosis and quality control. In addition, sensors/inspection systems like any other dynamic systems might degrade or fail after a certain time of usage (Jiang, 2011). In this case, unexpected deviations in sensor measurements from actual values will mislead control systems to provide wrong commands and result in non-conforming products. It may also cause unnecessary product waste or system shutdown.

Therefore, there is a need to increase the visibility of R2R systems without additional sensors in order to enhance real-time monitoring, improve fault diagnosis capability, facilitate intermediate product quality control and eliminate wrong decisions induced by defective sensors.

In this thesis, several research issues will be addressed in order to improve quality and productivity in R2R manufacturing systems.

1.2 Research issues

Different from traditional discrete manufacturing systems, an R2R manufacturing system is a continuous process that consists of a flexible substrate with rollers and combines different additive/subtractive operations to complete the product functionality. Due to its continuous manner, an intermediate process failure will cause the shutdown of an entire production line. Moreover, because of high speed and tight quality tolerance, even a minor fault can lead to hundreds or thousands of defective products. Compared with a conventional discrete manufacturing system, an R2R process is more sensitive to process failures/degradation so that it usually has much stricter requirements for real-time fault detection and more precise control to ensure product quality. In the following, several research issues that bring challenges to current real-time monitoring and fault diagnosis in R2R processes are discussed.

1.2.1 Visibility of R2R manufacturing systems

The visibility of a manufacturing system is critical for both operation performance assessment and product quality control. In a high-speed R2R manufacturing system, the lack of visibility hinders corrections of machine failures or degradation, resulting in a large amount of non-conforming products. Current industry practice utilizes sensors, such as temperature, speed and tension sensors to monitor process performance and machine vision systems for quality inspection in R2R processes (Subbaraman et al., 2012). Moreover, in order to cross validate sensor/inspection measurements, additional sensors for one target variable may be added to the system. Nevertheless, additional high-speed sensors and high-resolution in-process inspection

systems increase total production cost and sometimes are not feasible due to technical constraints such as production system configuration or information processing time. In order to provide early warning for defective operations in intermediate steps and reduce non-conforming products, analytical methods that are independent of additional sensors should be employed to estimate intermediate product quality and system performance.

Virtual sensors have been employed in many complex systems such as semiconductor manufacturing systems (Gill, 2011), chemical processes (Kano and Fujiwara, 2013) and building industries (Li et al., 2011). They can estimate operation or quality measurements when physical or hardware sensors are not economic, infeasible or unreliable. Therefore, virtual sensors are an effective solution to increase the system visibility without additional physical sensors. Instead of taking a direct measurement, a virtual sensor predicts the target variable by exploiting the relationship between inputs (other sensed data that can be easily collected) and outputs (target variables). Well-known algorithms for virtual sensing include filtering techniques, state observers or estimators, model-driven (first principle) and data-driven modeling methods (Li et al., 2011; Kano and Fujiwara, 2013).

1.2.2 Complexity of process monitoring and quality control in an R2R process

The complexity of R2R manufacturing systems increases challenges of real-time fault diagnosis and quality control. The first type of complexity resides in massive data from R2R manufacturing systems. R2R processes usually involve a large number of process parameters and input variables (e.g., input material properties and operation commands). Any special-cause variation from inconsistency in input material properties, operational error, equipment degradation or environmental changes can result in considerable yield loss at a high speed. In industry practice, control systems are implemented to regulate the process and provide proper

commands to manufacture products. Also, different sensors and inspection systems are installed to monitor operation performance and obtain measurements of critical quality characteristics (either dimensional or non-dimensional). In this case, different types of data at different scale will be generated. Big data opens the possibility for predicting machine operating conditions and intermediate product quality. Meanwhile, it also brings challenges of effectively and systematically converting those massive data into meaningful information for real-time monitoring and fault diagnosis in R2R manufacturing systems.

The second complexity comes from interactions between operations and product quality. Since an R2R process consists of multiple operations, the final product quality is affected not only by the local variation induced by a current operation but also by variations transferred from upstream operations. A small intermediate variation will be accumulated, and might result in nonconforming products when being processed by downstream operations. Moreover, since an R2R process works on a continuous flexible substrate, workpieces/products usually are all connected and will affect each other during production. Therefore, it is necessary to consider variations that are induced by both operations and the substrate in upstream stations.

1.2.3 Reliability of sensors and inspection systems

In an R2R manufacturing system, sensors (e.g., tension sensor, positioning sensor) and inspection systems are widely implemented for web process control and quality assurance. Those sensors and inspection systems obtain data of key indicators of system status, operation performance and product quality. Based on information from sensors, control systems will send proper signals for production operation (e.g., adjust roller speed to maintain proper tension on a substrate, discard unqualified products) or trigger system shutdown when corrective maintenance is required. Therefore, the performance of those sensors or inspection devices plays a vital role to

ensure the functionality of a production system, and to prevent rejections of qualified products/unnecessary shutdowns.

However, a sensor, like any dynamic systems, will fail if a failure occurs in any of its components such as transducer and signal processor (Jiang, 2011). Sensor degradation (e.g., caused by corroded contacts) or incipient (e.g., caused by deteriorated sensing elements) will generate inaccurate measurements from the target system (Isermann, 1984), and have a severe impact on automation and supervision schemes (Sherry and Mauro, 2014), possibly leading to system instability, loss of information fidelity, wrong decisions and disorientation of remedial actions (Bureau d'Enquêtes et d'Analyses pour la sécurité d, 2012).

In R2R manufacturing systems, intensive workload for sensors/inspection systems makes those devices tend to degrade or fail after operating for a certain of time. For example, a tension sensor installed in a roller often suffers from excessive vibration because of the high rotational speed. Deviations in this tension sensor will mislead the control system in tension regulation, and may result in web breakage if too much tension builds up while wrinkles if insufficient tension is on the substrate. Therefore, evaluating and understanding the current performance of sensor/inspection systems is important to provide accurate information for control systems, operation performance evaluation and quality control.

However, the dynamic behavior of an R2R manufacturing system is often highly nonlinear. It involves both steady state and transient state (e.g., ramp-up & ramp-down, material changeover), and may have quick changeovers to produce different types of products whose size or material may be different. In this case, an R2R manufacturing system often switches from one operating regime to another frequently. Different operating regimes trigger different system behaviors and may involve different levels of measurement noise and system disturbances. Using

a linear model or a simplified nonlinear model presents a challenge to describe the system behaviors under its entire operating range. Moreover, inconsistent model accuracy under different operating regimes brings challenges to differentiate between a sensor failure and measurement noise/system disturbances. Therefore, it is essential to explore sensor fault diagnosis in such complex systems.

1.2.4 Capability of adaptive learning with new operating regimes and failures

Due to the complexity of an R2R manufacturing system, a variety of malfunctions that are involved in components, process parameters and sensors under different operating regimes make complete reliance on human operators in dealing with those system malfunctions difficult and inefficient. To reduce product rejection rate and increase system availability, it is desirable to enable self-awareness in R2R processes, which can promptly detect any anomaly and reliably identify the root cause so that corrections can be taken quickly to restore the system to its normal operation. A great amount of efforts have been made to explore quantitative and qualitative models for process monitoring and fault diagnosis with model-based methods (Gao et al., 2015b) and data-driven methods (Gao et al., 2015a).

Despite the progress made up to date, most conventional fault diagnosis methods only address situations or failures that have been occurred, or can be anticipated. Any deviations away from learnt nominal behaviors are considered to be anomalies so that patterns of residuals between models and system behaviors are recognized to identify the severity or types of faults. Therefore, nominal system behaviors under entire working space and all possible signatures of failures/degradation should be well interpreted in advance, which is a challenging task. The difficulty resides in the fact that residuals might differ from one operating regime to another and the current trained model is not sufficient to describe the entire operating range of the system

accurately. Moreover, it is inapplicable and time-consuming to obtain sufficient data to understand system dynamics and possible failures/degradations under all various operating regimes. Therefore, known and unknown need to be discriminated without the knowledge of unknown behaviors, and an efficient adaptive learning algorithm is required to update models to improve its diagnostic capability.

1.3 Research objectives

This research aims to establish theoretical foundations for real-time monitoring and fault diagnosis in high-speed R2R manufacturing systems, to gain a fundamental understanding of multistage continuous R2R process dynamics and quality variation, to increase system visibility, and to enable self-awareness in R2R manufacturing systems. The following fundamental questions will be addressed in this dissertation:

1. How to characterize variation propagation in an R2R manufacturing system and increase the system visibility via virtual sensor methods so that to enable early detection of operation failure/degradation and improve quality control.
2. How to validate sensors/inspection systems in nonlinear systems so that to guarantee correct and reliable measurements for operation performance evaluation and quality control.
3. How to identify failures in complex R2R manufacturing systems with adaptive learning capability so that to handle new operating regimes and unprecedented failures efficiently.

To address the questions above, first, the characteristics of product variation propagation – called twofold variation propagation in R2R manufacturing systems are investigated. A multistage modeling based on the idea of "Stream of Variation" is proposed to quantify the twofold variation propagation and estimate its associated quality measurements. The modeling

technique employs both physics-based analysis (e.g., web handling system dynamics) and regression methods (e.g., censored regression, and linear/logistic regression) using multi-sensor signals. The estimation results from the model can serve as virtual sensing and virtual metrology tools to increase the system visibility and be applied for process monitoring and error detection in real time.

Second, to ensure accurate information for process monitoring and fault diagnosis, the measurements from sensors/inspection systems are validated. A nonlinear analytical redundancy method is developed to detect sensor faults in a general nonlinear system. Parity relations are formulated to describe system dynamics and sensor measurements and a robust optimization is designed to find the coefficients for the parity relations, which can tolerate the uncertainty from disturbance and measurements noise. The residuals generated from the parity relations (parity residuals) are used for sensor fault diagnosis. Post-processing sensitivity analysis is conducted to evaluate the effect of the change of operating regimes on parity residuals, and provide quantitative information of the effective operating regimes for the designed parity relations in sensor fault diagnosis.

At last, a generalized growing structure multiple model system (GSMMS) is designed for multi-regime anomaly detection and fault diagnosis. Following the idea of "divide and conquer", the multiple model approach is explored. The GSMMS method is revised to develop local models by formulating the model parameter identification problem into an optimization problem based on the loss minimization framework and solving with a gradient descent method. The growing self-organizing map is employed to automatically partition the operating space and to grow the number of local models based on input fed to the system. The proposed method is able to generate input-dependent residuals for anomaly detection and fault diagnosis and improve the

diagnostic capability.

1.4 Outline

The rest of this thesis is organized as follows.

Chapter 2 illustrates the characteristics of twofold variation propagation – process-centric and product-centric variation in R2R manufacturing systems. The multistage modeling method is introduced to describe the process-centric variation propagation, which uniquely exists in R2R processes, and its resulting quality measurements. A web handling system is employed to validate the modeling method. This chapter is based on a conference paper "Roll-to-Roll manufacturing system modeling and analysis by stream of variation theory" by H. Shui, X. Jin, J. Ni, published by ASME 2016 11th International Manufacturing Science and Engineering Conference, and a journal paper "Twofold variation propagation modeling and analysis in R2R manufacturing systems", by H. Shui, X. Jin, J. Ni, conditionally accepted by IEEE Transactions on Automation Science and Engineering.

Chapter 3 focuses on sensor fault diagnosis, which aims to validate the sensor/inspection measurements to facilitate system performance monitoring and fault diagnosis. The analytical redundancy approach is extended from linear systems to general nonlinear systems. Local parity structures and coefficients are determined by the nonlinear observation matrix and robust optimization design. The generated residuals are employed for sensor fault diagnosis. A post-processing sensitivity analysis is conducted to evaluate the effect of changing operating conditions on the residual generation of the proposed methods. The proposed method is validated with data from an R2R registration process.

Chapter 4 introduces a multiple regime modeling approach for anomaly detection and fault

diagnosis. The GSMMS network is employed and revised. The operating regimes are automatically learnt and partitioned by the growing self-organizing map. The local model identification problem is formulated as an optimization problem based on the loss minimization framework and solved with the mini-batch stochastic gradient descent method instead of the least squares method. The proposed method is demonstrated with the application of multiple regime sensor fault diagnosis and validated with data collected from an R2R registration process.

Chapter 5 summaries the contributions of this dissertation and proposes possible future works.

CHAPTER 2 TWOFOLD VARIATION PROPAGATION MODELING AND QUALITY ESTIMATION IN ROLL-TO-ROLL MANUFACTURING SYSTEMS

2.1 Introduction

Roll-to-Roll (R2R) manufacturing systems with high-volume and high-speed production capability require reliable system performance and precise quality control. The growth of new technological development of R2R processes in various applications increases the challenges faced by manufacturers in quality control and improvement. The difficulties reside in the limited understanding of complex interactions among sequential operations and product quality variations, and the restrictions of sensors/inspection systems implementation in R2R manufacturing systems.

This chapter aims to investigate a multistage modeling method for R2R processes to enhance the capability of quality monitoring and fault detection when in-situ sensing and inspection systems are limited. The challenges of achieving the objective entail two aspects: (1) the complexity of R2R processes and the lack of high-fidelity process model, and (2) the limited availability and accessibility of in-situ sensing and inspection capability. To meet these challenges, the variation propagation mechanism is investigated and a hybrid modeling approach – physics-based models and data-driven methods that complement each other – will be

developed to quantify the propagation phenomenon and the associated quality measures.

In the followings, the state of the art of quantitative models in R2R processes for quality monitoring and fault detection purposes is presented in Section 2.2. The novelty of this research work is also summarized in this section. Section 2.3 characterizes the twofold variation propagation model in R2R processes with an extended form of the Stream-of-Variation (SoV) model. Section 2.4 describes how to develop a multistage model that characterizes the variation propagation and the product quality evolution in R2R processes. In Section 2.5, a case study is provided to demonstrate the model validity and effectiveness. Discussion and conclusions are presented in Section 2.6 and 2.7, respectively.

2.2 Literature review

Literature that is mostly relevant to this work includes: R2R process modeling and quality control, and multistage modeling. The state-of-the-art of these research areas is reviewed below.

2.2.1 The complexity of R2R process modeling

An R2R process involves many operational inputs and parameters (e.g., operational settings, material properties, environment conditions, etc.). In particular, the large variability of input material properties (e.g., thickness, density, modulus, etc.) could lead to process variation and thus nonconforming products. Machine/tool degradation can also cause defects on products directly. Researchers have investigated non-ideal effects such as roller oscillation, temperature and moisture changes on web tension experimentally (Whitworth and Harrison, 1983; Branca et al., 2012) The mechanism of roll eccentricity under various roll speeds has been investigated in hot rolling mill (Lee et al., 2005). However, there is a lack of system-level R2R process modeling methods to quantify the relationship between process conditions and product quality,

and integrity such as printed pattern alignment, dimensional and positional accuracy.

Moreover, an R2R process consists of multiple operation steps in a continuous manner with substrates that are all connected, which is different from discrete manufacturing systems. Therefore, in addition to the dimensional and non-dimensional variations generated at each operation stage, and transformed and propagated as the substrate moves to the next operation, the quality of substrate is also affected by the upstream disturbance (e.g., changes of tension) that will instantaneously travel further downstream along the substrate. Such phenomenon introduces an additional spatial variation propagation in R2R processes.

Most previous research works focus on process modeling and controller design for a target subsystem without considering the effect of upstream change on downstream product quality. The quantitative modeling methods have been underdeveloped to understand the variation propagation mechanism and quality evolution in R2R processes, which are critical for process and quality control.

2.2.2 Limited in-situ measurement for process monitoring and quality control

There have been growing research interests in developing and improving sensing and inspection techniques to provide more information for real-time quality control and improvement (Subbaraman et al., 2012; Ulsh, 2014). Some data-driven process monitoring and quality control methods have been developed to improve the performance of R2R processes (Qiang Liu et al., 2011; Xiao et al., 2016; Dong et al., 2017). However, it is not possible to obtain measurements from all individual components (e.g., rollers, guides) in an R2R manufacturing system, due to the constraints of budget for instrumentation, data acquisition window, and space for sensor installation. Therefore, the performance and condition of an R2R process are usually not effectively observable due to the limited cost-effective sensors and in-process metrology tools

available in the system. The lack of visibility in the R2R process hinders the timely fault detection and delays the response for adjustment and recovery, which could result in hundreds of nonconforming products even when a small error occurs in a high-speed production. Thus, how to increase the system visibility by taking advantages of analytical methods without using additional sensors/inspection systems needs to be investigated.

The development of virtual sensing and virtual metrology by analytical models is envisioned to provide a complementary tool for quality control and fault diagnosis when physical sensors are limited. Analytical models incorporate empirical system knowledge to estimate product quality during the process. However, in R2R processes, most analytical models are developed to characterize the web transmission behavior for controller design. For example, dynamic models and empirical models have been well studied for web tension regulation in terms of torque and velocity control (Shelton, 1986), and lateral and longitudinal web dynamics (Young and Reid, 1993). These studies contribute to design precise and stable tension control systems to prevent unexpected web damage. Mathematical models are formulated with web tension, mold thickness, misalignment angle and substrate bending, and quantify effects of those process parameters in a patterning process for positioning/locating error compensation (Hwang et al., 2015). Causal Bayesian networks are employed to discover the causal relationships between product quality and process variables in a rolling process for process control (Li and Shi, 2007). There are few studies on quantitative modeling method to characterize quality variation propagation (e.g., dimensional errors) by given operational inputs and estimate the induced quality measurement in R2R processes.

2.2.3 Multistage modeling in discrete manufacturing systems

In the literature, both physical models and data-driven methods have been employed to

build multistage models to illustrate and analyze the interactions among stages and process variation propagation along sequential stages in discrete manufacturing systems. Those models serve as a foundation for quality analysis, root cause identification and process variation reduction (Shi and Zhou, 2009). Physics-based models are based on engineering knowledge and require comprehensive understandings of system dynamics of multiple process operations. The Stream of Variation (SoV) is a mainstream modeling technique that has been widely studied for multistage system modeling, especially for multistage modeling in rigid part assembly processes (Jin and Shi, 1999; Huang et al., 2007), compliant-part assembly processes (Camelio et al., 2003; Xie et al., 2007) and machining processes (Huang et al., 2002; Zhou et al., 2003; Djurdjanovic and Ni, 2003; Huang and Shi, 2004; Abellán-Nebot et al., 2013). The traditional SoV models utilize a stage-index state space model to describe variation propagation in discrete manufacturing systems.

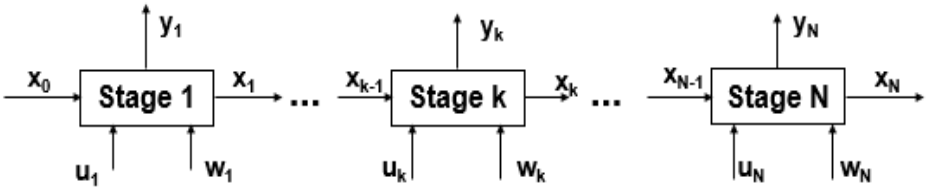


Figure 2-1: An example of a multistage manufacturing system

Figure 2-1 describes an N-stage discrete manufacturing system, and a generic state space model to describe the system is shown in Eq. (2.1):

$$\text{System Equation: } \boldsymbol{X}_k = \boldsymbol{A}_k \boldsymbol{X}_{k-1} + \boldsymbol{B}_k \boldsymbol{U}_k + \boldsymbol{W}_k$$

$$\text{Observation Equation: } \boldsymbol{Y}_k = \boldsymbol{C}_k \boldsymbol{X}_k + \boldsymbol{V}_k \tag{2.1}$$

where $\boldsymbol{X}_k$ is the state vector that represents a set of key quality characteristics of the product at stage k ($k = 1, \ldots, N$). N is the total number of stages. In Eq. (2.1), $\boldsymbol{X}_k$ denotes the

dimensional errors accumulated up to operation stage k, $\boldsymbol{U}_k$ is the system input vector that represents process error sources, $\boldsymbol{Y}_k$ is the measurement vector. $\boldsymbol{W}_k$ and $\boldsymbol{V}_k$ denote the system noise and the measurement noise, respectively. $\boldsymbol{A}_k$ is the transform matrix that describes how errors that accumulated up to stage $k-1$ are transformed to stage k while matrix $\boldsymbol{B}_k$ represents how new errors are introduced into the intermediate product at stage k. $\boldsymbol{C}_k$ links the quality characteristics to the quality measurements.

However, physics-based models are often difficult to develop due to the complexity of manufacturing systems. Data-driven methods, as alternatives to the physics-based models, have been employed to identify interactions among stages, such as factor analysis (Liu et al., 2008), graphical models (Zeng and Zhou, 2007) and Bayesian approaches (Djurdjanovic and Ni, 2004). By manipulating the system equation and observation equation into a generic model in the form of Eq. (2), data-driven methods estimate the transform matrix $\boldsymbol{\Gamma}$ by making effective use of historical sensing data. Those methods require a large amount of reliable and representative sensing data and inspections instead of physical knowledge to explore the relationship between inputs and outputs, which is not applicable in many manufacturing environments due to its high cost and low speed.

$$\boldsymbol{Y} = \boldsymbol{\Gamma f} + \boldsymbol{\epsilon} \tag{2.2}$$

where $\boldsymbol{\Gamma}$ is a coefficient matrix to illustrate the relationship between process and product; $\boldsymbol{f}$ represents the operation error sources and $\boldsymbol{\epsilon}$ is the random noise.

To date, no comprehensive modeling method has been developed to take advantages of both physics-based models and data-driven methods to analyze the variation generation and propagation in continuous web handling systems.

2.3 Summary of contribution

To fully realize the potential advantages of R2R processes and address these existing barriers, this thesis will investigate the variation propagation mechanism and the associated printing quality issues in the continuous R2R process. A twofold variation model describing how variations are introduced and transformed as the substrate goes from one operation stage to another (**product-centric variation model**), and how variation propagate instantaneously to the downstream substrate (**process-centric variation model**) will be built based on the web-and-roller dynamics of the continuous R2R process (see Figure 2-2).

This research work is also novel in developing a *hybrid modeling* approach to characterize the twofold variation propagation by given operational inputs, and to estimate the product quality. The modeling integrates physics-based models (torque equilibrium, and Hooke's law) with data-driven methods (e.g., censored regression, and linear/logistic regression) to address the complex variation propagation phenomenon. Specifically, sensor data analytics will complement the lack of full physical knowledge of an R2R process dynamics, while the physical knowledge will minimize the requirement for sensing and inspection at each stage. A multistage model based on the formulation of SoV will be developed to quantify the twofold variation propagation and the induced quality measurements for R2R processes. The estimates of the state variables (e.g., web tension) serve as a virtual sensor, while the outputs of the observation equations (e.g., printed product length or elongation) serve as a virtual metrology tool for intermediate product quality measurements based on system inputs (e.g., material properties and operational variables). As a result, the model can be a foundation for process diagnosis/prognosis, and quality control and improvement in R2R manufacturing systems. To the best of our knowledge, this is the first work that uses a SoV model to characterize the process-centric variation propagation and product

variability in a continuous system, whose mechanism is very different from the one in discrete manufacturing systems.

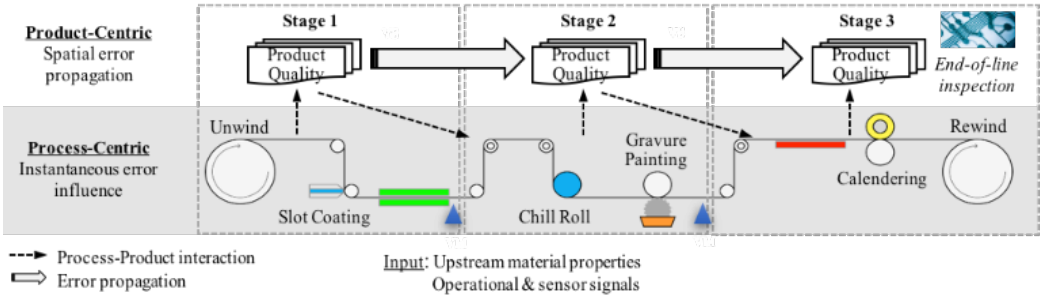


Figure 2-2: The schematic of the twofold variation model in an R2R system

2.4 Characterization of variation propagation in R2R manufacturing systems

This section presents the model framework, illustrates the formulation of each sub-model, and explains how they are developed separately and then integrated for product quality monitoring. The *product-centric variation* represents both dimensional and non-dimensional variations that are induced by sequential operations when the substrate moves from one operation stage to another, while the *process-centric variation* describes tension variation that is generated by the instantaneous effects from upstream substrate tension change. The formulation of the SoV model is extended to characterize the twofold variation mechanisms and the quality estimates in R2R manufacturing systems. The details are shown as follows.

Product-Centric Variation Model: An R2R process consists of a series of sequential operations to complete the product functionality. Unlike discrete manufacturing systems, there is no clear boundary between stages in R2R processes. Thus, to facilitate modeling, the continuous process is segmented into multiple stages according to the functional or dimensional features that rollers and operations act together to achieve. There is no rule of thumb, but it is based on the

requirement of the model fidelity (how much visibility the system would like to have). More segmentation may require more efforts of modeling and computation, while rough segmentation might limit the visibility of the system and prevent timely fault detection in earlier operation steps.

Moreover, to facilitate an efficient way of representing the multistage modeling for an R2R process, the continuous substrate can also be *segmented into different tension spans* especially when there are repeated printed patterns on the substrate. Each segment of the substrate is denoted as a 'pitch', which represents the size of a printing product. Particularly, the pitch length can be aligned with dimensional patterns on the substrate but does not exceed the length between two stages.

Product-centric variation is built upon the same product (i.e., one pitch) and is propagated temporally as it sequentially goes through one stage to another. Take the cold roll forming process of quadrate steel tube as an example (Figure 2-3), the locating error of rollers is introduced, transformed and propagated until the sheet metal is formed into a final shape, which has an adverse impact on the dimensional quality of the product (Zhang et al., 2008). By discretizing the continuous process into multiple stages according to the functionality of the operations and quality outputs of interests (Figure 2-4), the SoV model is used to describe the influence of roller's locating error on the dimensional variations of the product throughout the process.

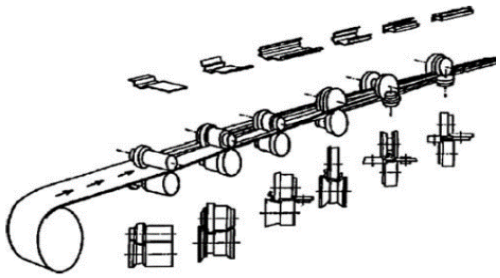


Figure 2-3: Cold roll forming process of quadrate steel tube (Zhang et al., 2008)

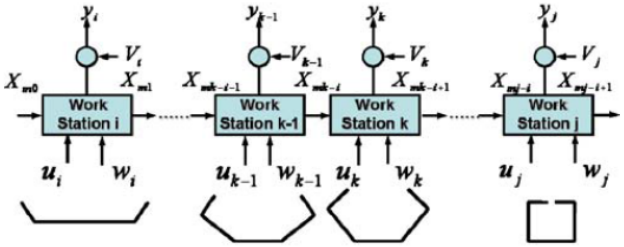


Figure 2-4: Segmentation of the cold roll forming process (Zhang et al., 2008)

In this research work, the SoV model is employed for the modeling of the product-centric variation propagation in the same way as in discrete manufacturing systems to describe variations on one pitch generated and propagated over a sequence of operations. The system equation for the *product-centric variation propagation* of one pitch is written as:

$$\boldsymbol{S}_k^{id} = \boldsymbol{A}_k \boldsymbol{S}_{k-1}^{id} + \boldsymbol{B_k U}_k + \boldsymbol{W}_k \tag{2.3}$$

where $\boldsymbol{S}_k^{id}$ is the state vector that represents deviation of id^{th} pitch in stage k at time $\mathrm{t} = t_k$, and $\boldsymbol{S}_{k-1}^{id}$ presents id^{th} pitch at the previous stage $k-1$ at time $t = t_{k-1}$. $\boldsymbol{U}_k$ is the system input vector. $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ are the state and input matrices, which can be determined by mathematical tools (Zhou et al., 2003; Camelio et al., 2003), computer-aided engineering (Li et al., 2007; Li and Shi, 2007), and/or data-driven methods (Zeng and Zhou, 2007; Liu et al., 2008). Those methods have been proven to be effective for characterizing variation propagation in discrete manufacturing systems. Figure. 2-5 shows the schematic of the product-centric variation stream

in a three-stage R2R process.

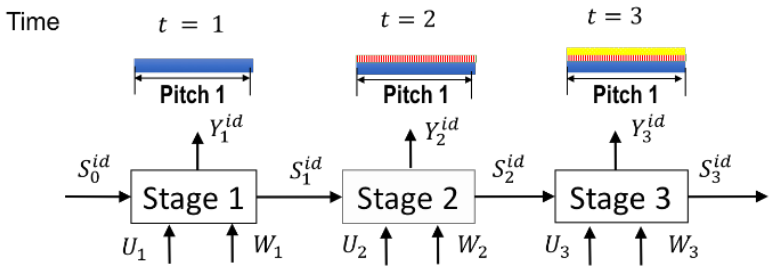


Figure 2-5: Schematic of the product-centric variation propagation in a three-stage R2R process

Process-Centric Variation Propagation: In addition to the product-centric variation, there exists a unique variation propagation phenomenon in R2R processes that is defined as *process-centric variation propagation* in this thesis. Process-centric variation is propagated specially at each fixed time, which means a change on the upstream substrate can affect the downstream substrate **instantaneously**. In this study, we only concern web tension with the *process-centric variation propagation* as it simultaneously occurs at different locations of the substrate. In general, the web tension is controlled by driven rollers to avoid undesirable over-stretching or under-stretching problems. However, web tension is sensitive to the change of roller performance, wrap angle, material properties, and environment conditions (e.g., temperature, humidity). The variation in web tension will be transformed into improper pitch length and thickness of the printing deposition, as well as generate wrinkles, scratches and breaks. Figure 2-6 shows an example of improper tension from the upstream stage instantaneously affecting the tension on the downstream substrate and resulting in misalignment of downstream pitches. Figure 2-7 provides an example of the printing misalignment problem caused by excessive tension in the substrate.

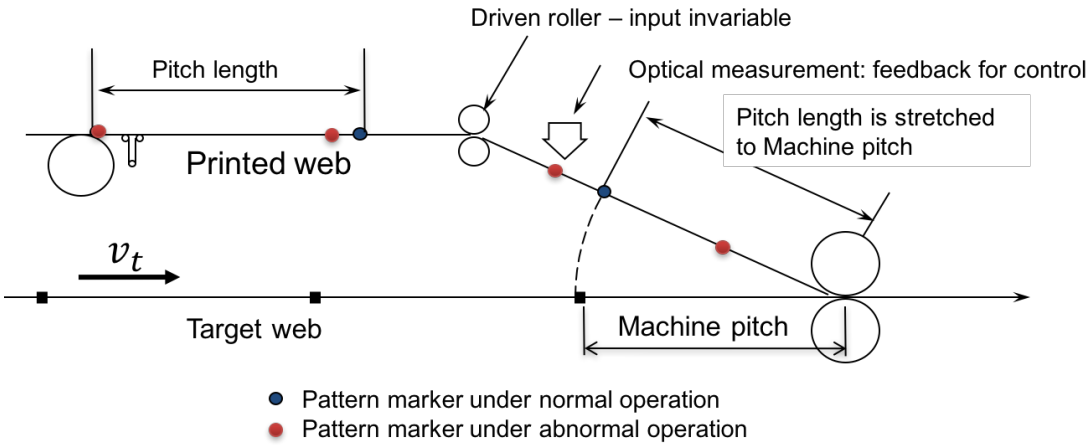


Figure 2-6: An example of tension variation propagation in a registration process

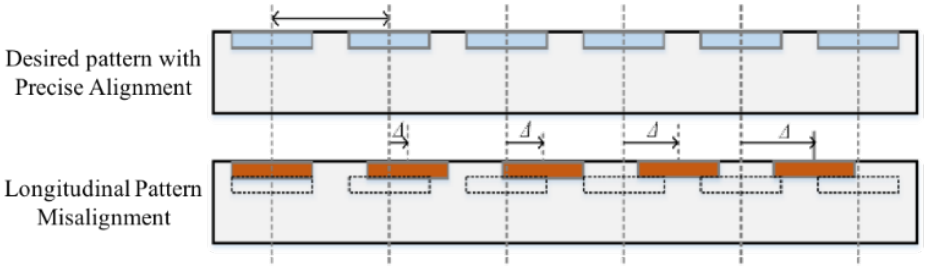


Figure 2-7: An example of tension variation propagation that induces misalignment

In this research work, the web tension is assumed to be *uniformly* distributed on the substrate within one stage. The tension on two sides of a pair of rollers within one stage might be different. However, since the modeling focuses on the input and output quality measures from each stage, the tension calculation aggregates all the intermediate rollers within one stage and only outputs one tension for the one stage, neglecting the tension difference within the stage. Also, there might be multiple pitches at one stage, the web tension of those pitches is treated as identical (Figure 2-8). This assumption is for notational convenience. In practice, there can be multiple rollers at one stage, resulting in non-uniform web tension on pitches within one stage (e.g., three pitches in Stage 2).

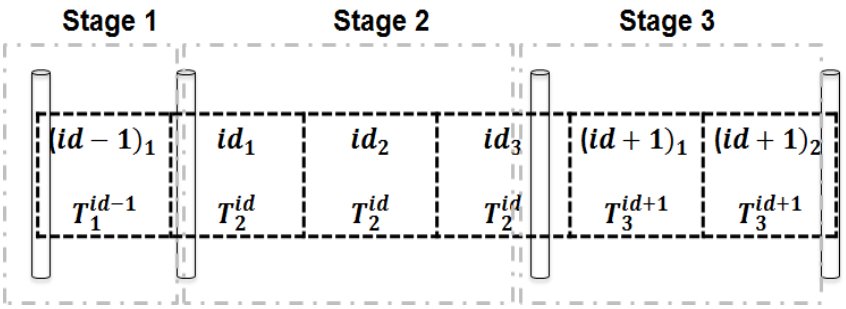


Figure 2-8: Segmentation of a substrate

At time *t*, a state-space form of the *process-centric variation propagation* is written as follows:

$$\boldsymbol{T}_k^{id}(t) = \boldsymbol{E}_k \boldsymbol{T}_{k-1}^{id-1}(t) + \boldsymbol{F}_k \boldsymbol{R}_k(t) + \boldsymbol{Z}_k \tag{2.4}$$

where $\boldsymbol{T}_k^{id}(t)$ is the state vector describing the web tension on the id^{th} pitch at stage k. $\boldsymbol{T}_k^{id}(t)$ is instantaneously affected by the web tension on $(id-1)^{\text{th}}$ pitch at the preceding stage, $\boldsymbol{T}_{k-1}^{id-1}(t)$. $\boldsymbol{R}_k(t)$ is the input vector related to roller operations, which affects the tension $\boldsymbol{T}_k^{id}(t)$. Matrices $\boldsymbol{E}_k$ and $\boldsymbol{F}_k$ denote how the upstream web tension and the current operation affect the tension at the current stage. The approach for determining those two matrices is detailed in Section 2.6.

Table 2-1 illustrates how the product-centric and process-centric models work together to characterize the variation evolution in an R2R process (shown in Figure. 2-9). In this example, the process is segmented into three stages. At time t, S_k^{id} is calculated for the id^{th} pitch on the substrate which is transferred to downstream stages at each time $t + \Delta t$ ($\Delta t = t_2 - t_1 = t_3 - t_2$), while T_k^{id}'s represent the web tension on different pitches that are located at all previous stages.

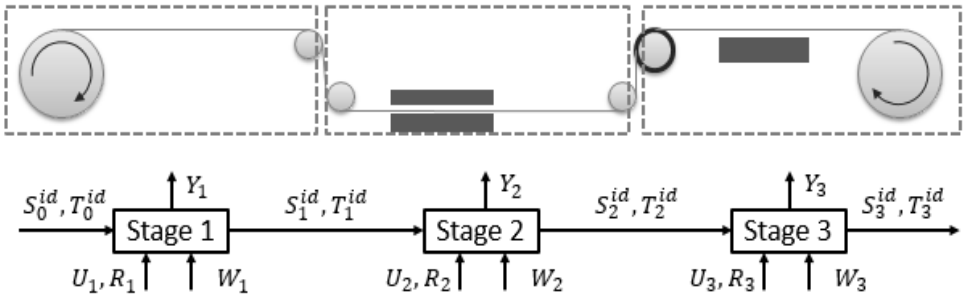


Figure 2-9: A multistage model for an example R2R process

Table 2-1: An example of the twofold variation propagation modeling

t	Stage 1	Stage 2	Stage 3
$t = t_1$	$S_1^{id} = A_1 S_0^{id} + B_1 U_1 + W_1$ $T_1^{id}(t_1) = E_1 T_0^{id-1}(t_1) + F_1 R_1(t_1) + Z_1$		
$t = t_2$	$T_1^{id-1}(t_2) = E_1 T_0^{id-2}(t_2) + F_1 R_1(t_2) + Z_1$	$S_2^{id} = A_2 S_1^{id} + B_2 U_2 + W_2$ $T_2^{id}(t_2) = E_2 T_1^{id-1}(t_2) + F_2 R_2(t_2) + Z_2$	
$t = t_3$	$T_1^{id-2}(t_3) = E_1 T_0^{id-3}(t_3) + F_1 R_1(t_3) + Z_1$	$T_2^{id-1}(t_3) = E_2 T_1^{id-2}(t_3) + F_2 R_2(t_3) + Z_2$	$S_3^{id} = A_3 S_2^{id} + B_3 U_3 + W_3$ $T_3^{id}(t_3) = E_3 T_2^{id-1}(t_3) + F_3 R_3(t_3) + Z_3$

At last, the quality measurements $\boldsymbol{Y}_k$ at stage k are characterized with both $\boldsymbol{S}_k^{id}$ and $\boldsymbol{T}_k^{id}(t)$. Considering the combinatorial effects of product-centric and process-centric variations on the product quality, at time t, the product quality measurement at stage k is represented as:

$$\boldsymbol{Y}_k = h(\boldsymbol{S}_k^{id}, \boldsymbol{T}_k^{id}) + \boldsymbol{V}_k \tag{2.5}$$

where the notation of $\boldsymbol{T}_k^{id}(t)$ is simplified as $\boldsymbol{T}_k^{id}$. The function $h(\boldsymbol{S}_k^{id}, \boldsymbol{T}_k^{id})$ links product-centric variation and process-centric variation to the product quality measurement. The function h describes the mapping between the state variables and the output variables. The explicit expression of h can be determined by either physics-based models or data-driven methods, which is detailed in Section 2.5. Altogether, the system Eqs. (2.3) and (2.4), and the observation Eq. (2.5) form a complete multistage model for the R2R manufacturing system, where the unmolded

errors $\boldsymbol{W}_k$ and $\boldsymbol{V}_k$ are *independent and identically distributed* (*i.i.d.*) with $N(0, \sigma_w)$ and $N(0, \sigma_v)$, respectively. The measurement noise $\boldsymbol{Z}_k$ is *i.i.d.* with $N(0, \sigma_z)$. Lastly, σ_w, σ_v and σ_z can be estimated from samples.

Table 2-2: Overview of quantitative modeling methods in multistage manufacturing systems

System Type	Product-centric error propagation	Process-centric error propagation
Discrete	SoV modeling: ● Physical models: (Camelio et al., 2003; Djurdjanovic and Ni, 2003a, 2006; Huang et al., 2007a, 2007b; Jin and Shi, 1999; Liu et al., 2010; Loose et al., 2010; Xie et al., 2007; Zhou et al., 2003) ● Data-driven methods: (Djurdjanovic and Ni, 2003b; Liu et al., 2008; Zeng and Zhou, 2007)	N/A
Continuous (R2R)	SoV modeling: Physical models: (Zhang et al., 2008)	The problem studied in this paper with a hybrid model that integrates both physical models and data-driven method.

Table 2-2 provides a summary of the existing literature relevant to the quantitative modeling of multistage manufacturing systems in different categories based on the type of systems (i.e., discrete or continuous) and the type of variation propagation mechanisms in the system (i.e., product-centric or process-centric). This thesis is the first work investigating the modeling method for both types of variation propagation in R2R processes and the influence of the variations on product quality characteristics.

2.5 Multistage modeling for R2R manufacturing systems

This section focuses on the *process-centric variation propagation* model (i.e., tension propagation) in R2R manufacturing systems. The web tension at all individual operation stages are modeled as the system states and their variation propagations are formulated as state-space form equations. The relationship between the state and measurable quality variables is described as observation equations.

The overall modeling approach is a mix of physics-based and data-driven methods (see Figure 2-10). For each operation stage, a physics-based model will be first investigated to formulate a preliminary state-space model. However, some parameters or comprehensive physical knowledge might not be available to describe the variation propagation and quality evolution from one stage to another. Therefore, data-driven methods are embedded to identify those unknown parameters and explore the inexplicit relationship between inputs and outputs. The preliminary physical models give insights to extract features for data-driven models so that the model can be trained effectively. For simplicity, only the model application for *longitudinal tension* variation propagation is described here. The lateral tension can be estimated using the same approach.

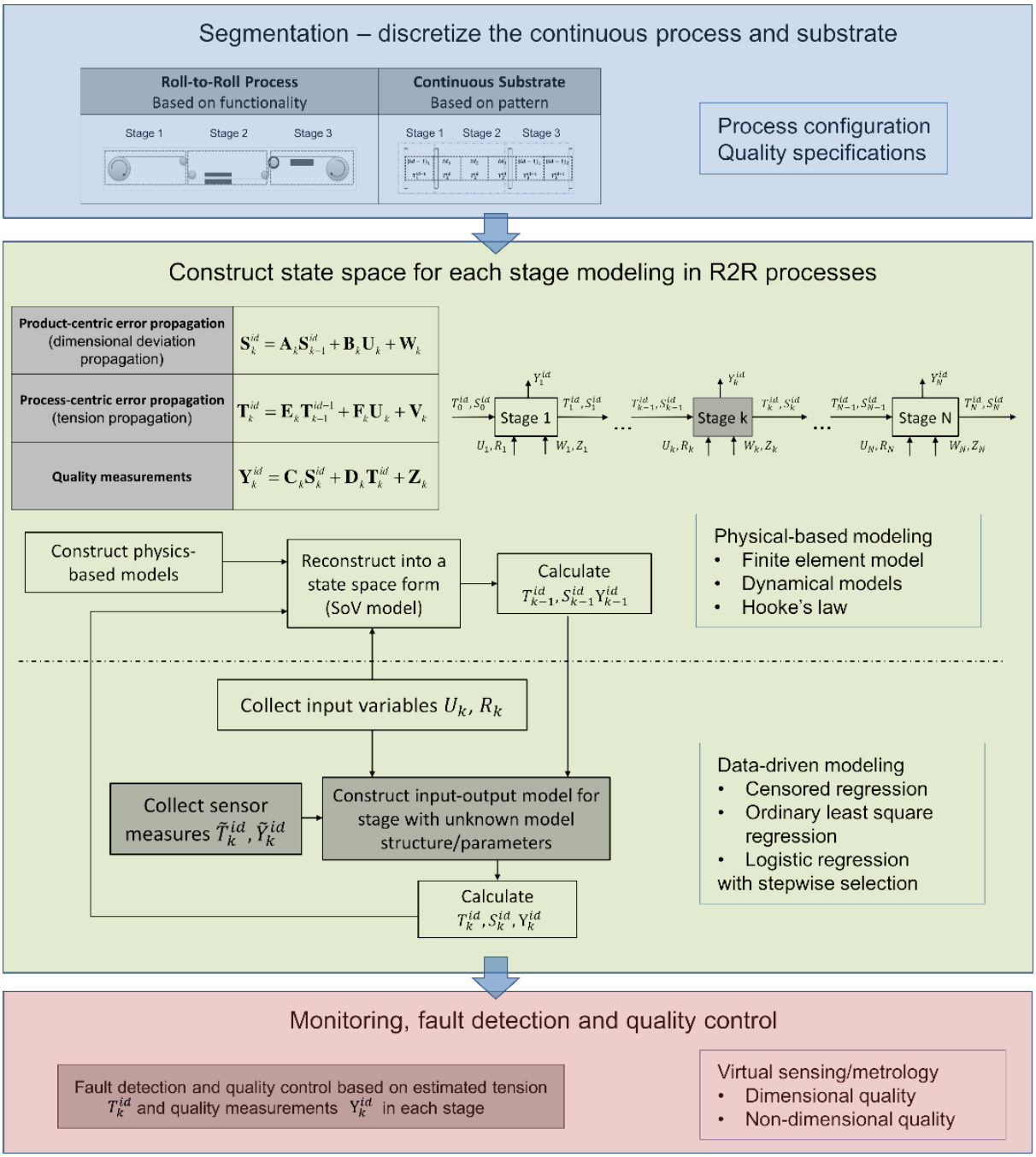


Figure 2-10: The framework of multistage modeling in R2R manufacturing systems

2.5.1 Web dynamic models for roll-substrate motions

The web dynamics of an R2R manufacturing system have been well studied primarily for tension control and speed regulation. In general, an R2R manufacturing system consists of

several critical components such as (a) material rolls, (b) driven rollers, (c) idle rollers and (d) a dancer as shown in Figure 2-11. A material roll is a rewinding roll or an unwinding roll that delivers substrate in/out of the production system. A driven roller is driven by a motor to regulate the roller speed, and control the tension on the substrate and its transportation speed. An idle roller is free from the motor and rotates due to the movement of the substrate over it. The main function of an idle roller is to support web transmission and guide web direction.

To characterize the tension variation propagation, the dynamic equations of material rolls, driven rollers, and idle rollers that are derived from torque equilibrium analysis (Appendix I) are reconstructed into the form of the SoV model. The system equations that illustrate the *process-centric variation propagation* in Eq. (2.4) therefore can be obtained for material rolls, driven rollers and idle rollers as shown in Eqs. (2.6), (2.7), and (2.8), respectively.

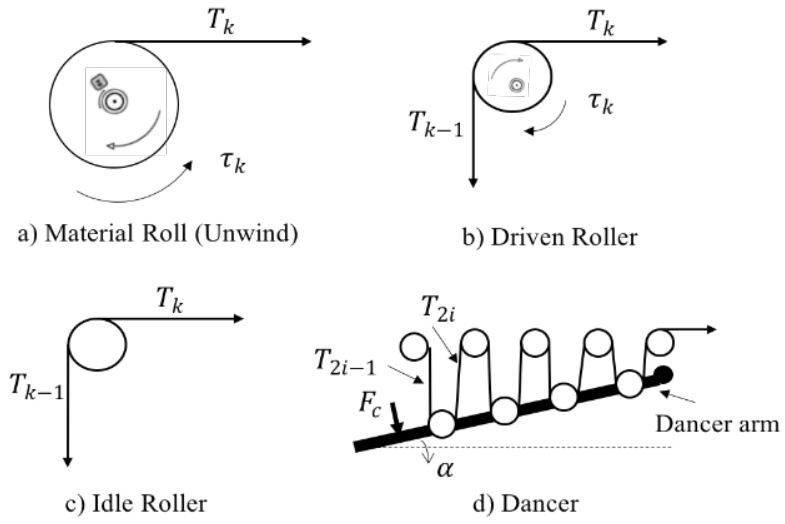


Figure 2-11: Primary rollers in R2R manufacturing systems – (a) material roll, (b) driven roller, (c) idle Roller, (d) dancer

Material roll:

$$T_k^{id}(t) = T_{k-1}^{id-1}(t) + \frac{J_0\dot{\omega}_k(t)}{R_k(t)} + \frac{\pi}{2R_k(t)}\rho_w w_w (R_k^4(t) - R_c^4)\dot{\omega}_k(t) - \rho_w w_w t_w R_k^2(t)\omega_k^2(t) +$$

$$\frac{\tau_k(t)+\tau_{fk}}{R_i(t)} + Z_k \tag{2.6}$$

Driven roller*:*

$$T_k^{id}(t) = T_{k-1}^{id-1}(t) + \frac{1}{R_k}\left(J_k\dot{\omega}_k(t) + \tau_{fk} - \tau_k(t)\right) + Z_k \tag{2.7}$$

Idle roller:

$$T_k^{id}(t) = T_{k-1}^{id-1}(t) + \frac{1}{R_k}\left(J_k\dot{\omega}_i(t) + \tau_{fk}\right) + Z_k \tag{2.8}$$

The left-hand side of each equation represents the web tension going out of the current stage. The right-hand side consists of the tension generated by operations at the current stage and the upstream tension $T_{k-1}^{id-1}(t)$. This indicates that the web tension at one stage is not only determined by the current operation but is also instantaneously affected by the upstream operations, hence the SoV model is well adapted.

2.5.2 Censored regression model for system equation

For some critical components in an R2R process, the system equation cannot be explicitly obtained from torque equilibrium analysis. Two main reasons for the difficulty in using physics-based governing equations are as follows: (1) due to the complexity and nonlinear dynamics of the components, the model structures and parameters cannot be directly obtained from physical or engineering knowledge; (2) the dynamic behavior may change dramatically when the system is in its transients (e.g., material changeover, ramp-up and ramp-down). An example in R2R processes is the modeling of the dynamic of the dancer motion. A dancer is a critical component that functions as a buffer to store and release materials. The classical design of a dancer is to use multiple idle rollers and a rocker arm of the dancer. Eq. (2.9) shows the governing equation for a dancer (Frechard et al., 2013).

$$J_c\frac{d^2\alpha}{dt^2} = \sum r_i(T_{2i} + T_{2i-1}) - F_cL_c - M_cg(sin\alpha)L_d \tag{2.9}$$

where J_c is the inertia of the dancer, α is the angle between dancer arm and horizontal, F_c is the force generated by a spring, L_c is the distance between the force applied and dancer arm rotational axis, M_c is the mass of the dancer arm, L_d is the distance between dancer arm rotational axis and dancer arm gravity center, r_i is the distance between i^{th} idler and dancer arm rotational axis.

However, an explicit state-space model for describing the relationship between the output tension from a dancer with the process variables is difficult to derive from Eq. (2.9) because during a transient state (e.g., material changeover), the dancer's arm moves up and down leading to a changing r_i, which is not measurable and difficult to estimate. Therefore, a data-driven method is more applicable to obtain the correlation among the web tension from the upstream stage, the dancer operation stage and the stage after.

In this new modeling framework, a data-driven method is employed to identify the input-output relationship when it cannot be obtained from the physical analysis. From another perspective, the physical models developed in Section 2.5.1 provides a relaxation of the requirement of having sensors/inspection systems installed at each stage, rendering useful insights on feature selection for model.

In this research, a ***censored regression model*** and ***forward stepwise selection method*** (CR-FSS) with *F*-tests are employed to establish the system Eq. (2.4). Censored regression introduced by Tobin is originally employed for censored observations in econometric and biometric applications (Tobin, 1958). Censored observations are the variables that can only be observed under certain conditions. In this case, a traditional ordinary least square regression that provides a biased estimator is no longer appropriate. Web tension is modeled as a censored variable, which can only be observed when the substrate is stretched, otherwise, it remains zero. Therefore,

the system Eq. (2.4) is revised by the censored regression model as follows:

$$T_k^{id} = \begin{cases} T_k^{id^*} & if\ T_k^{id^*} > L \\ 0 & if\ T_k^{id^*} \leq L \end{cases} \tag{2.10}$$

where T_k^{id} is a real measurement (longitudinal tension) and $T_k^{id^*}(t) = E_k T_{k-1}^{id-1}(t) + \boldsymbol{F}_k \boldsymbol{R}_k(t) + Z_k$ is a latent variable, $L = 0$. Parameters E_k and $\boldsymbol{F}_k$ can be obtained by the maximum likelihood estimation method (detailed in Appendix II).

For those stages with unknown parameters, the system equations that are derived from physics-based models in Section 2.5.1 can be revised as Eq. (2.10) to obtain unknown parameters. For those stages with unknown model structures, the *forward stepwise selection method* (FSS) is employed to determine the dominant inputs for the system equations. FSS is a variable selection method that iteratively adds predictive variables among many candidates to identify critical variables according to *F*-test. The procedure of an FSS is described as follows:

(1) Select the operating variables (input variables) that may affect the tension variations at individual stages based on the engineering knowledge.

(2) At each iteration, add one operating variable in the input vector $\boldsymbol{R}$ and form the system equation as Eq. (2.10) to estimate parameters E_k and $\boldsymbol{F}_k$.

(3) Conduct an *F*-test to determine whether to drop or include the newly added variable.

Take the dancer component as an example, the dancer angle α, the external force F_c, the angular velocity and acceleration of an idle roller in the dancer ω_i and $\dot{\omega}_i$ might affect the tension on the exit of the dancer. After the FSS is processed for offline modeling using historical data, the critical operating variables are determined. In this example, they are dancer angle α, angular velocity and acceleration of idle rollers in the dancer ω_i and $\dot{\omega}_i$, which will be further

demonstrated in the case study in Section 2.6.

2.5.3 Physical analysis for observation equation

The observation equations describe the relationship between the state variables and quality measurements. In this thesis, the deviation of key quality measurements at each operation stage is represented as a combinational effect of both product-centric variation and process-centric variation. This section discusses how the observation equation is formulated to describe the product quality deviation accumulated across multiple stages induced by the process-centric variation propagation, which exists uniquely in R2R processes.

In an R2R process, tension anomalies directly affect the substrate dimension (i.e., elongation) as well as the material transportation rate. The amount of elongation of the substrate must be kept within an acceptable range for quality assurance. One of the most critical quality measures relevant to the elongation is the dimension of the pattern printed or generated on the substrate because it is often used as a reference for downstream operations such as printing, lamination, registration, and cutting. Therefore, tension needs to be well monitored and controlled to avoid excessive elongation and undesired pattern geometry. To demonstrate the physics-based modeling of observation equations, we formulate the longitudinal length of pattern L_k by using Hooke's law $T = ES\varepsilon$ (ε is the engineering strain $\varepsilon = \frac{\Delta L}{L_0} = \frac{L-L_0}{L_0}$):

$$L_k = (\frac{T_k}{ES} + 1)L_0 + V_k \tag{2.11}$$

where L_0 is the relaxed length (no tension on the substrate) and L_k is the elongated length in stage k (tension applied on the substrate). E is the elastic modulus. S is the cross-section area of the substrate. T_k is the tension applied on the substrate in stage k and V_k is the random error. Eq. (2.11) also represents an explicit function to describe the intertwined effect of the product-centric

variation propagation (L_0 is printed by upstream operations) and process centric variation (T_k) on the resulting product quality (L_k).

2.5.4 Logistic regression for observation equations

When the physical analysis is not sufficient to construct the observation equations, data-driven methods such as linear regression and logistic regression can be employed to correlate the twofold variation propagation with dimensional and non-dimensional quality measurements, respectively. Since the modeling procedures of linear and logistic regression are similar, this thesis will describe the detailed modeling for non-dimensional variations with logistic regression. In the R2R processes, non-dimensional variations could be caused by abnormal web tension. Too much tension could cause the substrate to be stretched beyond its elastic limit, leading to a damage. However, insufficient tension could cause slippage between the substrate and roller or wrinkles on the substrate, leading to defective products. Here, the prediction of non-dimensional defects is formulated as a classification problem with the logistic regression model.

Without the loss of generality, the *logistic regression* (LG) model for multiclass classification is shown below. For multiple classes, the probability of a defect (e.g., wrinkle) is conditioning on both process-centric and product centric variations. In this case, the conditional probability of the observation output Y_k can be defined as follows (Bishop, 2006):

$$\left(Y_k = o \middle| S_k^{id}, T_k^{id}\right) = \frac{e^{C_k^{(o)} S_k^{id} + D_k^{(o)} T_k^{id}}}{\sum_{o=1}^{O} e^{C_k^{(o)} S_k^{id} + D_k^{(o)} T_k^{id}}} \tag{2.12}$$

The maximum likelihood estimator (MLE) for the parameters $C_k^{(o)}$ and $D_k^{(o)}$ can be derived as follows:

$$\left(\hat{C}_k^{(o)}, \hat{D}_k^{(o)}\right) = \underset{C_k^{(o)}, D_k^{(o)}}{arg\,max} \left\{ \sum_m^M \sum_o^O {Y_k}^j \ln p\left({Y_k}^j = h \middle| {S_k^{id}}^j, {T_k^{id}}^j, C_k^{(o)}, D_k^{(o)}\right) \right\} \tag{2.13}$$

where h denotes the severity level of a defect ($o = 1, \ldots, O$). O is the total number of

severity levels.

2.6 Case study

In this section, the proposed multistage modeling framework for R2R manufacturing systems is demonstrated and validated by a web handling process from a real R2R manufacturing system. Due to the confidentiality, all results are normalized, and data identifiers are removed. The simplified configuration of this process is shown in Figure 2-12. The overall function of this process is to deliver pre-patterned substrate and ensure accurate registration at stage 5. Therefore, the most critical dimensional measure is the pattern longitudinal length at the registration – stage 5. With this case study, we illustrate how the proposed model based on the SoV theory is formulated for web tension estimation and pitch length prediction throughout of the five-stage R2R process.

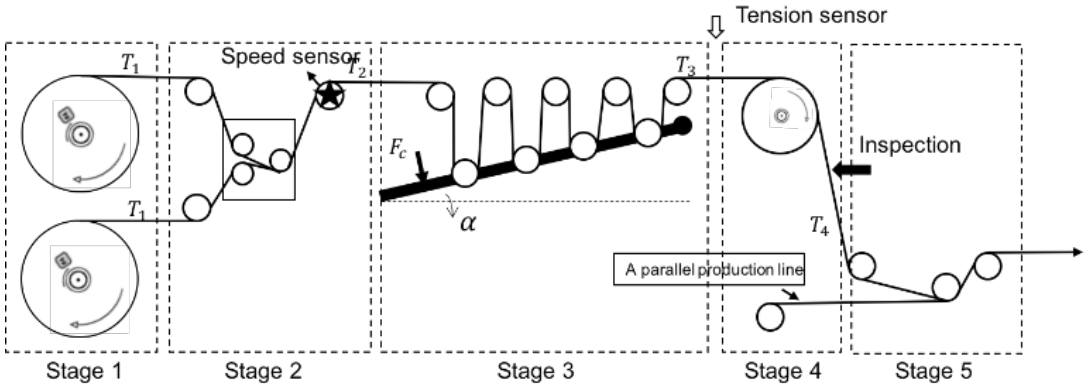


Figure 2-12: Testbed layout

First, the web handling process is segmented into *five* stages according to their functions – (1) unwinding, (2) splicing, (3) dancer, (4) driven roller and (5) registration. There are two material rolls in the *first* stage – one is called running roll and the other is called back-up roll which is used when the running roll is used up. The main function of this stage is to pull the substrate from its wound roll and feed it into the production line. The *second* stage consists of

several idle rollers and a splicing device. The splicing device will be automatically engaged in when the main material roll expires. It cuts and tapes the leading edge of the new material roll with the old roll to maintain the continuous process. The *third* stage is a dancer consisting of several idle rollers and a dancer arm. An external force from an air cylinder drives the movement of the dancer arm. The dancer is used to accumulate substrate before the running roll is expired and release substrate for downstream operation during splicing to eliminate stoppage in production. The *fourth* stage is a driven roller and the last stage has several idle rollers for registration. At stage *four*, a two-dimensional (2D) machine vision system is installed to capture the pitch length information. A closed-loop control system is embedded to regulate the speed of the driven roller to control the tension according to the feedback from the 2D machine vision system.

To identify and validate the hybrid multistage modelling method, a tension sensor is installed after stage 2 and a speed sensor is installed on the idle roller located at stage 2. The radius of the material roll is monitored by an ultrasonic sensor. For the material roll and driven roller, the velocity and torque that are generated by motors are collected from the closed-loop control system. There are seven sets of data being collected for model training (4 cycles of data) and validation (3 cycles of data). Both tension and idler speed signals are preprocessed by wavelet de-noising and outlier removal. In addition, it is assumed that each idler has the same speed with their adjacent idlers and all data are collected under an ideal environment (i.e., consistent humidity and temperature, no degradation of rollers).

2.6.1 Multistage modeling for an unwinding process

After segmenting the R2R process, a preliminary multistage model is first established based on the physical models in Section 2.5.1. The system equations and the observation

equations in Eq. (2.1) for the web handling process can be written as:

System equations:

$$T_{si} = \begin{cases} T_{si}^* & if\ T_{si}^* > 0 \\ 0 & if\ T_{si}^* \leq 0 \end{cases} \tag{2.14}$$

where T_{si}^* for stage 1 to 5 is formed as below:

$$T_{s1}^* = T_{s0}(R_1(t)) + \frac{J_1\dot{\omega}_1(t)}{R_1(t)} + \frac{\pi}{2R_1(t)}\rho_w w_w (R_1^4(t) - R_c^4)\dot{\omega}_1(t) - \rho_w w_w t_w R_1^2(t)\omega_1^2(t) + \frac{\tau_1 + \tau_{f1}}{R_1(t)} + W_{s1} \tag{2.15}$$

$$T_{s2}^* = \sum_i^{N_{s2}} \frac{1}{R_i}\left(J_i\dot{\omega}_i(t) + \tau_{fi}\right) + T_{s1} + W_{s2} \tag{2.16}$$

$$T_{s3}^* = f\left(T_{s2}, \alpha, \omega_1(t), \dot{\omega}_1(t)\right) + W_{s3} \tag{2.17}$$

$$T_{s4}^* = \frac{1}{R_i}\left(J_i\dot{\omega}_i(t) + \tau_{fi} - \tau_i\right) + T_{s3} + W_{s4} \tag{2.18}$$

$$T_{s5}^* = \sum_i^{N_{s5}} \frac{1}{R_i}\left(J_i\dot{\omega}_i(t) + \tau_{fi}\right) + T_{s4} + W_{s5} \tag{2.19}$$

Observation equations:

$$Y_{s4} = \left(\frac{T_{s4}}{SE_{s4}} + 1\right)L_0(R_1(t)) + V_{s4} \tag{2.20}$$

$$Y_{s5} = \left(\frac{T_{s5}}{S[(1-\mathrm{d})\mathrm{E_p} + \mathrm{dE}_q]} + 1\right)L_0(R_1(t)) + V_{s5} \tag{2.21}$$

where N_{s2} and N_{s5} present the total number of the idlers in stages 2 and 5, respectively. $T_{s0}(R_1(t))$ is the initial tension from the wound roll, and $L_0(R_1(t))$ is the relaxed pitch length.

In the above equations, an explicit model for stage 3 - the dancer $f\left(T_{s2}, \alpha, \omega_1(t), \dot{\omega}_1(t)\right)$, and the initial tension $T_{s0}(R_1(t))$ cannot be derived from physical analysis directly. Therefore, a censored regression model and a forward selection method are employed to determine those unknown structures and parameters. Similarly, the function of the relaxed length $L_0(R_1(t))$ is

obtained by a linear regression model with a forward selection method. The resulting models are shown as follows.

$$T_{s0}(R_1(t)) = p_0 + p_1 R_1(t) + p_2 R_1^2(t) \tag{2.22}$$

$$T_{s3}^* = d_0 + d_1 T_{s2} + d_2 \alpha + d_3 \dot{\omega}_1(t) + W_{s3} \tag{2.23}$$

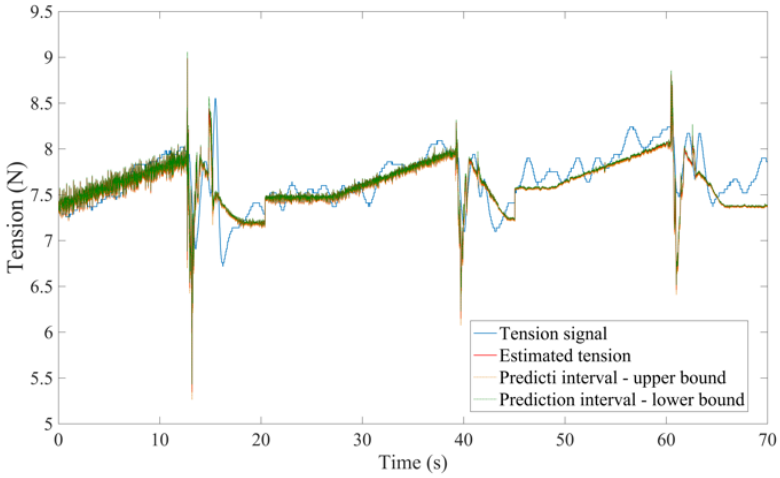
$$L_0(R_1(t)) = q_0 + q_1 R_1^3(t) \tag{2.24}$$

2.6.2 Model validation

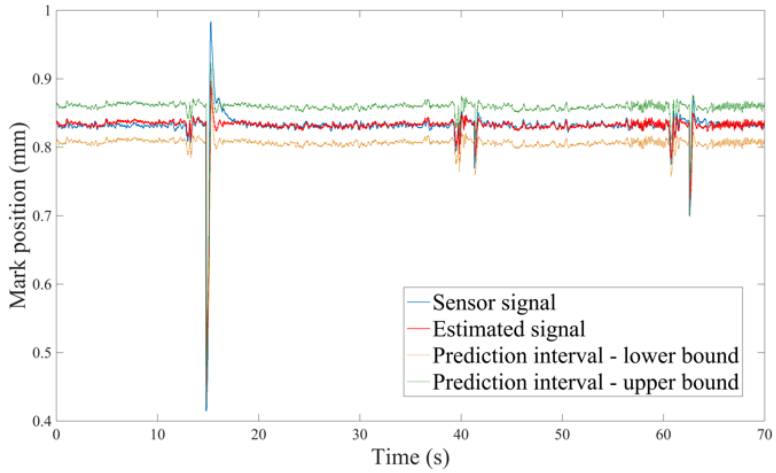
This section presents the model validation and model performance analysis. The estimated web tension and predicted pitch length from the model are compared with the real measurements, i.e., tension signals from strain gage transducer load cells, and pitch length measurements captured by a 2D machine vision system.

Figure 2-13 shows the results of three operation cycles, starting from the beginning of a new material roll and ends when it's depleted and changed over by a new roll. During the material changeover, the web speed at stages 1 and 2 is reduced to zero to facilitate the splicing process, while the dancer releases material to downstream operation stages so that the web speed at stages 4 and 5 remains constant to ensure the process smoothness. Therefore, spikes are generated by the abrupt splicing events at stage 2, causing tension variations on the downstream substrate.

The comparison of the estimated intermediate tension T_{s2}^* from stage 2 with real-time tension signals is shown in Figure 2-13 (a), while the comparison of the estimated pitch length Y_{s4} and its measurement via the 2D machine vision system at stage 4 is shown in Figure 2-13(b). Both the estimated intermediate tension and pitch length can capture same trends as the sensor measurements.

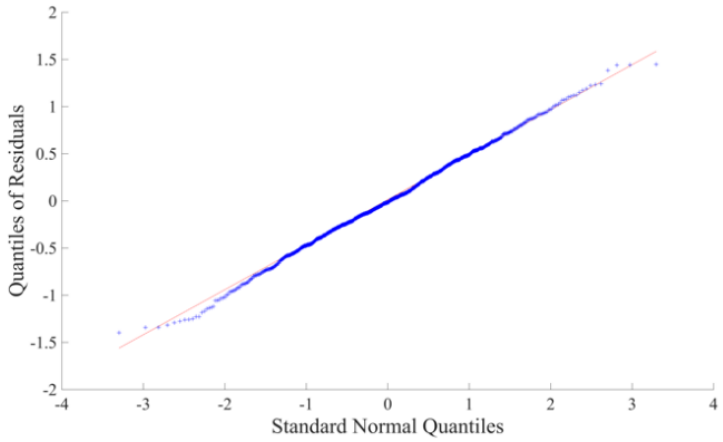


a) Estimated tension in stage 2 vs 1st tension signal

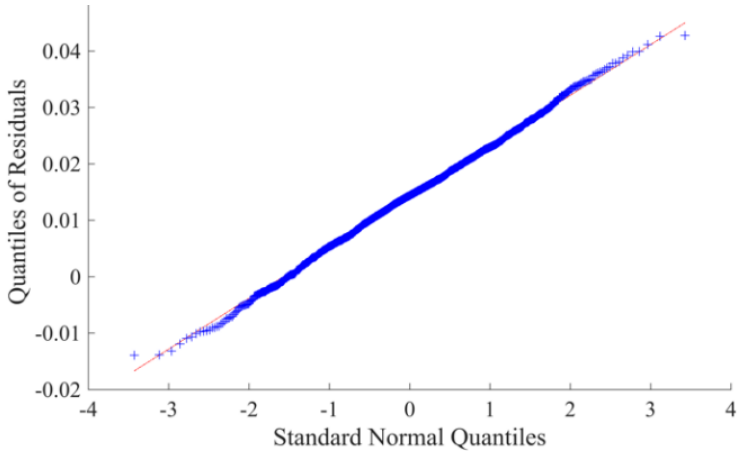


b) Estimated pitch length in stage 3 vs real measurement

Figure 2-13: Multistage model validation results



a) QQ plot of residuals for 1st tension versus standard normal



b) QQ plot of residuals for pitch length versus standard normal

Figure 2-14: Residual analysis for multistage model

In Table 2-3, the hybrid model that integrates physics-based and data-driven methods is compared with the regression-based data-driven methods. Furthermore, the tension propagation modeling results from censored regression is compared with linear regression. The root-mean-square errors (RMSE) from validation results indicate that the proposed hybrid model is more accurate than the data-driven model in estimation of the tension (improved by 15%) and pitch length (improved by 70%). AIC criteria is employed to quantify the goodness of fit (Burnham

and Anderson, 2002). The AIC results shown in Table 2-3 imply that the hybrid model can capture more information than the pure data-driven method. Moreover, the comparison of RMSE that are generated from a steady state and a transient state is shown in Table 2-4. It shows during the steady state, the state estimates by the hybrid model is smaller than those during the transient state (i.e., spikes). This is because during the transient state, more noise and disturbance are introduced into the system.

The results of residuals diagnostics for the hybrid model are listed in Table 2-5 and Figure 2-14 (Q-Q plot), showing that the residuals are not auto-correlated, have constant variance and follow normal distribution. Therefore, the model is proven to be adequate.

Table 2-3: Root Mean Squared Error (RMSE) Comparison

RMSE		1st Tension			Pitch Length		
		Training	Validation	AIC	Training	Validation	AIC
Physics-based + data-driven	Physics-based + Linear regression	0.2136	0.2228	N/A	**0.0140**	**0.0102**	**-4229**
	Physics-based + Censored regression	**0.2137**	**0.2214**	**7096**	N/A	N/A	N/A
Data-driven	Linear regression	0.1493	0.2526	N/A	0.0138	0.0344	-4210
	Censored regression	0.1354	0.2612	7797	N/A	N/A	N/A

Table 2-4: Model Selection Test

	RMSE comparison		
	Steady state (without spikes)	Transient state (only spikes)	One cycle
1st Tension	0.1277	0.3229	0.2214
Pitch Length	0.0037	0.0260	0.0102

Table 2-5: Residual Statistics Test

	Residual Auto-correlation test (Ljung-Box Q-test)		Residual Heteroscedasticity (Engle's ARCH test)	
		P-value		P-value
1st Tension	Not reject	0.4076	Not reject	0.5590
Pitch Length	Not reject	0.1001	Not reject	0.2818

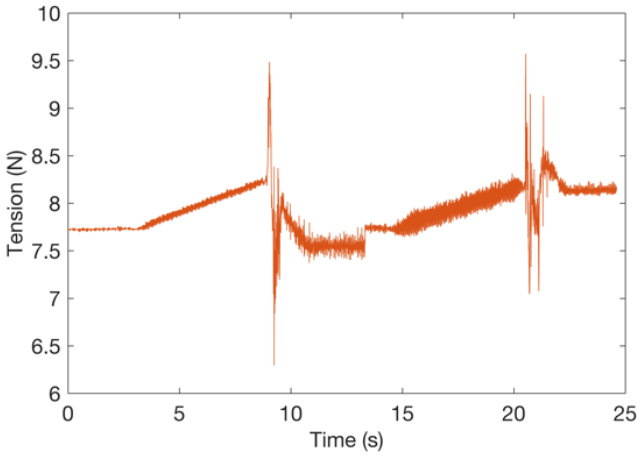
2.7 Discussion

Current R2R manufacturing systems suffer from a lack of sufficient visibility of system performance and quality monitoring because of limited in-situ measurements. Such lack of visibility hinders the timely detection of product defects and correction actions, leading to significant wastes including materials and capital spending when a small error occurs. The proposed hybrid approach for variation propagation modeling in R2R processes is capable of devicing a new virtual sensing tool and a virtual metrology by estimating the intermediate tension and quality measurements, respectively. Therefore, the visibility of the system and process performance can be improved and the additional information for system performance evaluation will be available for the quality control. In the following, an example – dancer failure in an R2R print registration unit is presented to demonstrate how the proposed method can be applied for effective fault detection and quality monitoring in R2R manufacturing systems.

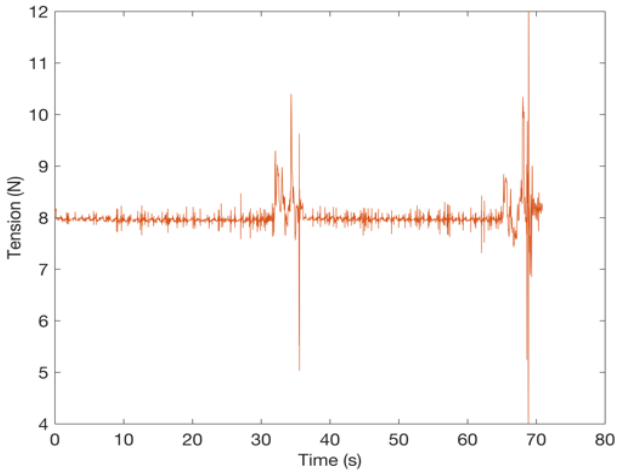
We particularly address the longitudinal tension variation problem. Too much tension can cause the web to be stretched beyond its elastic limit, hence the web is damaged and cannot be

recovered in the later stages. Insufficient tension could cause wrinkles generated on the web or bubbles during lamination. In addition, a slack web may be gathered or stretched around the rollers and cause catastrophic machine failure. Therefore, web tension is critical for the detection of web damage or slippage, and its trend can also indicate faults and degradation in operations.

For example, by comparing the two estimated tension profiles over time in Figure 2-15, we observed a different tension trend. The tension profiles from two production cycles in Figure 2-15 (a) are from normal operation while in Figure 2-15 (b), a web breakage occurs during another two production cycles. Under a normal operation, the tension will gradually increase before the material roll is changed over. However, there is no such a climbing pattern in the tension profiles under an abnormal operation. By checking the system fault events log, the breakage occurred during the material changeover, justifying our conclusion. The dancer arm failed to move to accumulate materials so that there was no enough material for downstream operation during the material changeover. In this case, too much tension was suddenly generated on the web and caused the web breakage. The proposed model is effective in sensing such fault during the process in real-time and provides insights on the possible physical reasoning.



a) Normal operation



b) Web breakage

Figure 2-15: Tension profiles under normal and abnormal operating conditions

2.8 Conclusion

In this chapter, a novel modeling method is presented for the twofold variation propagation and its induced quality deviation in R2R manufacturing systems. Based on the formulation of the SoV theory, a hybrid modeling approach is proposed to formulate a multistage model to characterize both product-centric and process-centric variation propagation, as well as their associated quality measures in R2R processes. This chapter mainly focuses on the systematic modeling of process-centric variation propagation - tension propagation since it uniquely exists in R2R processes, and its induced quality variation by coupling physics-based models with data-driven methods. First, the continuous R2R manufacturing system is segmented into multiple stages and converted to fit the formulation of SoV model. Second, the physics-based models based on torque equilibrium and the Hooke's law are reconstructed, which releases the requirement of full inspections after each stage, and well describes how the tension propagates across multiple stages and how critical product quality variations evolves from one stage to

another. At last, the censored regression, and the linear/logistic regression are employed as a complementary approach to explore complex interactions between rolls and substrate, and identify immeasurable parameters in the system. The case study – a web handling process demonstrates the performance of the proposed modeling method by comparing the estimated results with real-time sensor measurements, as well as the results estimated from pure data-driven method.

The result shows that the proposed hybrid modeling method outperforms the data-driven method – the tension estimation is improved by 15% and the pitch length estimation is improved by 70%. With the proposed multistage model, the visibility of the R2R manufacturing systems is successfully improved by estimating intermediate tension and measures of the key product characteristics at each stage. The state estimates and the predicted model outputs can serve as virtual sensing and virtual metrology, respectively to improve the system performance monitoring and product quality control. In the discussion, we've shown that by monitoring the intermediate tension, undesired web damage or slippage can be avoided. The tension trend can provide effective information for detection and identification of operational faults. The intermediate and final product quality can also localize abnormal operation by checking the deviation of product quality at each stage. As for the future work, process-centric variation other than tension can be investigated. Moreover, since the development of the multistage model partially relies on data from sensor and inspection systems, methods for sensor validation (Chapter 3) as well as optimal sensor placement are worthy of exploration.

In summary, the hybrid modeling method can effectively represent the operation performance and product quality in an R2R manufacturing system and reveal the errors in real time. There is a great potential for a wide range of applications with its benefits including the

reduction of product rejection rate and unexpected downtime, and improvement of system reliability and productivity.

CHAPTER 3 A GENERALIZED NONLINEAR ANALYTICAL REDUNDANCY METHOD FOR SENSOR FAULT DIAGNOSIS

3.1 Introduction

The benefit of automated monitoring and control procedures advances the research development and usage of sensing techniques in engineering systems. For example, in a typical automotive vehicle, there are 60-100 sensors on board and is projected to reach 200 sensors per car. GE launched a new factory for battery production with more than 10,000 sensors spread across 180,000 square feet of manufacturing space. Those sensors play important roles in data collection to make commands for system operation, supervise system performance, and diagnose and accommodate faults for ensuring system reliability and safety. Similarly, in a high-throughput R2R manufacturing system, an ever-increasing number of sensors and inspection systems are installed to make informed decisions for facilitating operation monitoring and quality control during the production to manufacture qualified products. The actions executed based on sensor measurements include but are not limited to adjust operation variables (e.g., tension or speed control of the substrate), reject unqualified products, and request corrective maintenance.

However, in R2R manufacturing systems, sensors often work in a severe and fast-changing environment with high pressure, high temperature or strong vibration. Like any dynamic systems,

those sensors are very vulnerable to fail or degrade over time. Both abrupt (e.g., caused by corroded contacts) or incipient (e.g., caused by deteriorated sensing elements) can generate non-permitted deviations from characteristic properties in sensors and result in inaccurate measurements from monitored target variables (Isermann, 1984). Consequently, a sensor malfunction can lead to wrong control efforts, unnecessary product rejections/system shutdowns, and equipment failures. For example, within a tension control system, deviations in a tension sensor may cause wrong control commands and lead to excessive or insufficient tension on the substrate, which will result in a breakage or wrinkles.

To validate sensor measurements, both hardware redundancy and analytical redundancy approaches have been developed for sensor fault diagnosis. Hardware redundancy usually requires a high cost of extra sensor installation and maintenance, and often time is restricted by space and weight concerns. Analytical redundancy is more cost-effective and has been developed for many engineering applications. It employs mathematical models to describe the systems and generate residuals for fault diagnosis. However, mathematical models are often inaccurate because of modeling errors or disturbances. Although there have been many robust fault diagnosis techniques developed to handle such uncertainty as well as ensure the sensitivity to faults, the sensor fault diagnosis in nonlinear systems still remains challenging. The modeling in an R2R manufacturing system is a typical example, whose system dynamic behavior is often highly nonlinear. It involves steady state and transient state (e.g., ramp-up & ramp-down, material changeover), and may have quick changeovers to produce different types of products whose size or material various. In this case, an R2R manufacturing system might switch from one operating regime to another frequently. Different operating regimes triggers different system behaviors and may involve different levels of measurement noise and model uncertainties. It is

usually not sufficient to use a linear model or a simplified nonlinear model to describe the system behavior under its entire operating range. The inconsistent model accuracy under different operating regimes brings challenges to differentiate between a sensor failure and measurement noise/model uncertainties.

To conquer the challenge, this chapter aims at developing a real-time sensor fault diagnosis method in such complex systems by proposing a model-based nonlinear analytical redundancy approach. Nonlinear observation matrix is employed to derive input-output equations to describe system dynamics and sensor measurements. Robust optimization is designed to obtain best coefficients so that the generated residuals will be robust to noise and uncertainties, but sensitive to sensor failures.

In the following, the literature review of sensor fault diagnosis is shown in Section 3.2 and the problem formulation of nonlinear analytical redundancy is presented in Section 3.3. In Section 3.4, the proposed modeling method for parity residual generation in a nonlinear system is proposed. The robust optimization and post-processing sensitivity analysis are given to obtain model coefficients and evaluate the detectability of the designed model. A case study that demonstrates and validates the proposed method is detailed in Section 3.5 and the conclusion is in Section 3.6.

3.2 Literature review of sensor fault diagnosis

A general solution to validate sensor measurements in a real-time environment is to add redundancy in the system. Hardware redundancy is the most intuitive approach that has been applied to many quality/safety-critical systems. It adds additional sensors to measure critical targets in a system and check the consistency among redundant sensors to detect if any sensor is

faulty. However, additional sensors require extra cost, weight, and space. Even though the recent evolution of micro-technology has contributed to reducing the size and cost of sensors, the hardware redundancy approach is still not applicable in many industrial applications. Moreover, redundant sensors may fail or degrade in the same way of the primary sensor since they all work under the similar operating environment (Patton et al., 1989). At last, when multiple sensors fail, the hardware redundancy approach tends to be infeasible to detect sensor failures under majority voting scheme (Broen, 1974).

Analytical redundancy provides a promising solution that is independent of redundant sensors. Both qualitative and quantitative models have been developed to add analytical redundancies in systems to check the accuracy of sensor measurements. Qualitative models are mainly built based on qualitative and heuristic reasoning or causal relationship between observations and system performance. Quantitative models employ mathematical expressions to represent system dynamics and estimate sensor measurements under fault-free conditions. With a nominal model and real-time sensor measurements, residuals are generated to detect and isolate sensor failures using various methodologies. The successful deployment of analytical approaches highly relies on model accuracy, which is affected by different levels of measurement noise and model uncertainties under different operating conditions.

The state-of-the-art modeling methods to achieve analytical redundancy for sensor fault detection and isolation can be further categorized into three: model-based methods, knowledge-based expert systems and data-driven methods (Jiang, 2011). The model-based methods such as parity relations (Chow and Willsky, 1984), Luenberger observers and Kalman filtering, (Clark, 1978; Tesheng Hsiao and Tomizuka, 2005; Du and Mhaskar, 2014) and parameter estimators (Upadhyaya and Kerlin, 1987) can develop quantitative models to generate for sensor fault

diagnosis. However, those methods require thorough knowledge of target system dynamics to formulate high fidelity models, which is often not applicable for complex systems. Knowledge-based expert systems (Betta et al., 1995; Kim, 1997) such as lookup table, fault tree and fuzzy logic based methods require comprehensive engineering domain knowledge of system behaviors under various normal conditions and faulty conditions. It has limited capability to handle dynamic systems especially during the initial development phase due to its rule-based mechanisms. Data-driven methods such as artificial neural networks (Mathioudakis and Romessis, 2004; Elnokity et al., 2012) and multivariate statistical methods (Negiz and Cinar, 1992; Dunia et al., 1996; Huang et al., 2000) are able to handle complex systems but require sufficient data to learn data patterns or trends to represent system performance, and usually lack physical insights. In real practice, those methods can be integrated to leverage their own advantages and disadvantages so that to obtain representative symptoms for diagnosis/prognosis. Nevertheless, it is usually preferable to start with model-based methods when physical knowledge is available.

In literature, the standard model-based analytical methods have been developed for linear systems (Chow and Willsky, 1984; Qin and Li, 2001; Li and Shah, 2002). However, many engineering systems are nonlinear. Nonlinearity does not obey superposition principles, and tends to introduce discontinuity and unpredictable output into systems. Those properties make the implementation of nonlinear analytical redundancy difficult. Many efforts have been made to linearize nonlinear systems in order to apply linear analytical redundancy methods (Nguang et al., 2007). However, nonlinear systems suffer considerably from linearization since it may introduce errors and reduce model accuracy. The performance of model-based analytical redundancy methods is sensitive to inconsistencies between nominal models and actual system behaviors.

Therefore, those errors will affect the effectiveness of linear analytical redundancy methods for sensor fault diagnosis in nonlinear systems. Other researchers explored nonlinear analytical redundancy methods for some special cases of nonlinear systems whose nonlinear observation matrix can be formulated into a linear form in terms of inputs and outputs (Yu and Shields, 2001; Shumsky, 2008; Leuschen et al., 2005). However, those nonlinear analytical methods are only feasible for specific system types, therefore, lack generality. A general form of nonlinear analytical redundancy approach is still not available.

To fill the gap, this chapter will investigate a model-based analytical redundancy method for the sensor fault diagnosis problem in general nonlinear systems in which, both input and output equations are nonlinear functions of states and inputs. Following the idea of the linear analytical redundancy method that utilizes observation matrix to construct input-output relations to describe the relationships between system behaviors and sensor measurements, this study will employ nonlinear observation matrix that is derived from system dynamic equations in the control theory to build the input-output relations with a parity space approach. The number of available analytical redundancies that can be added for sensor fault diagnosis will be determined by the rank of the nonlinear observation matrix.

3.3 Problem formulation

Consider a general nonlinear system with N states, Q measurable inputs and M sensors:

$$\begin{aligned} \dot{x} &= f(x,u) + \varepsilon \\ y &= h(x,u) + \delta \end{aligned} \tag{3.1}$$

where $x \subseteq R^N$ is the state vector, $u \subseteq R^Q$ is the measurable input vector, $y \subseteq R^M$ is the sensor measurements, ε is the system disturbance and δ is the measurement noise. Different from

existing works, the system here is a general form, in which both the input equation and the output equation are nonlinear functions of states and inputs. Thus, the formulation of nonlinear analytical redundancy should be flexible to represent the dynamic behavior of the system by considering its dependence on both states and inputs.

A model-based analytical redundancy method exploits the null-space of the state space observation matrix to generate residuals for fault diagnosis. Those residuals contain the complete information from sensor data and actuator inputs to detect any deviations from the nominal behavior of sensors. Existing works for nonlinear systems only have addressed those system models that can be linearized (Nguang et al., 2007) or use a simplified nonlinear observation matrix to formulate a structure for residuals generation (Leuschen et al., 2005). Those approximations will introduce a considerable amount of model errors so that, for general nonlinear systems, they sometimes can hardly provide effective solutions for accurate sensor fault diagnosis.

In order to formulate effective analytical redundancy for general nonlinear systems, the notion of observability is employed. Observability is a fundamental measure in control theory, which reflects the possibility for estimating intermediate states based on input and output signals.

Definition 3.1: The system is **locally observable** at x_0 if there exists a neighborhood of x_0 such that every x in that neighborhood other than x_0 is distinguishable from x_0. The mathematical expression for checking local observability is:

$$Rank(\nabla O(x_0, u^*)) = N \tag{3.2}$$

where N is the rank of x. For each output y_i, ($i = 1,...,M$)

$$\begin{bmatrix} y_i \\ \dot{y}_i \\ \ddot{y}_i \\ \vdots \end{bmatrix} = \begin{bmatrix} L_f^0 \\ L_f^1 \\ L_f^2 \\ \vdots \end{bmatrix} h(x_0, u^*) = O(x_0, u^*) \tag{3.3}$$

$$L_f^k = L_f(L_f^{k-1} h_i) \quad \text{and} \quad L_f^0 h_i = h_i \tag{3.4}$$

$$L_f^1 h_i = \frac{\partial h_i}{\partial x} f + \frac{\partial h_i}{\partial u} \frac{du}{dt} \tag{3.5}$$

where O is a nonlinear observation matrix, and $L_f^1 h_i$ is the Lie derivative of h_i with respect to f. In nonlinear systems, nonlinear observability is only feasible in its local space that is close to x_0 and u^*. It also requires the system to be smooth so that the Lie derivative $L_f^k h_i$ exists. It is assumed that the major operating regime meets this requirement in this research.

From the definition, it is seen that the second term $\frac{\partial h_i}{\partial u} \frac{du}{dt}$ on the right-hand side of Eq. (3.5) is the main difference between this work and others. Existing methods do not consider the situation that the output y_i is related to inputs. As a result, the effect of this term is neglected. In this research, a complete form of analytical redundancy structure for general nonlinear systems is proposed.

Moreover, the locally observable property indicates that the observability in a nonlinear system is valid within a certain working region that is near the current states and inputs. In real practice, when various operating conditions exist, an operating regime with the same level of observability has to be identified, and the entire working space has to be partitioned according to different operational and environmental conditions. Chapter 4 will propose a multiple operating regimes modeling method, which can automatically partition the working space so that local

analytical redundancies can be extended to global space, while this chapter will employ robust optimization method to identify the coefficients for local analytical redundancies around an operating point (x_0 and u^*) and apply post-sensitivity analysis to evaluate the effect of the change of the operating point on the optimal objective function and coefficients.

To sum up, Chapter 3 focuses on the formulation of the nonlinear analytical redundancy in general nonlinear systems. A complete form of analytical redundancy structure is proposed for sensor fault diagnosis. A robust optimization is designed to identify the model coefficients so that the generated residuals are projected to a space where they are sensitive to sensor failures but robust to noise. Moreover, post-processing sensitivity analysis is applied to the optimization problem to evaluate how inputs affect the optimal solution. The resulting residuals generated from the proposed analytical redundancies can be utilized to identify sensor failures or quantify its degradation status. The diagnostic capability of the proposed method in nonlinear systems is demonstrated and validated with simulation data from an R2R registration process. The systematic approach that can automatically partition operating regimes and learn local models will be proposed in Chapter 4, while this chapter will demonstrate the necessity and how different operational inputs affect the diagnosis result.

3.4 Methodology of nonlinear analytical redundancy

The model-based analytical redundancy method exploits the basic concept of observability, which contains key information of the system behaviors that can be inferred from the observation space. Also, a proper design of analytical redundancy is able to generate linearly independent residuals so that all detectable deviations from the system models can be accounted. In this section, the parity space method is employed to add analytical redundancies in the system. To address the sensor fault diagnosis problem in nonlinear systems, the traditional parity space

model is extended from linear to nonlinear systems so that input-output equations can be formulated to generate residuals. Furthermore, robust optimization design is employed to obtain optimal coefficients for the parity space model so that parity residuals will be close to zero when only measurement noise and model uncertainty involve, while will have large magnitudes when sensor failures occur in the system. Post-processing sensitivity analysis is conducted for the optimization problem to test the effects of different operating inputs on the optimal objective function – the magnitude of parity residuals.

3.4.1 Parity residual generation in general nonlinear systems

The parity space captures key information of system dynamics from the observation space. As shown in Figure 3-1, parity relations characterize relationships among measurable inputs and sensor outputs so that a set of residuals can be obtained (Chow and Willsky, 1984). In addition, the linear independence property in parity space guarantees that every parity residual generated contains at least some information that has not been covered by other residuals. On the other hand, each observable deviation from the system model is covered by at least one of the parity residuals.

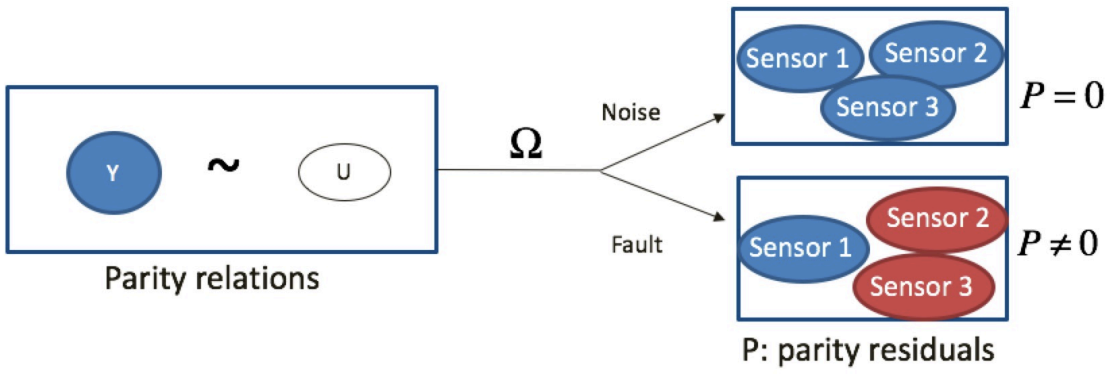


Figure 3-1: Sensor fault diagnosis with the parity space approach

The nonlinear observability imposes a prerequisite to construct the parity space that allows analytical redundancy for general nonlinear systems. In the following, based on **Definition 3.1**, the complete parity relations are derived, and the parity space is formed with

$$\Omega^{\perp} O(x_0, u^*) = 0 \tag{3.6}$$

where Ω is a vector of nonzero coefficients that transfers residuals from original space to parity (null) space.

A straightforward approach to formulate the input-output equations for analytical redundancies in general nonlinear systems is to follow the methods in linear systems, which formulate analytical redundancy structures with linear functions of parameters (Chow and Willsky, 1984). However, due to the nonlinear form of both states and inputs, it is difficult to directly adapt the linear analytical redundancy structure for the general nonlinear systems. Also, since the output *y* depends on both states and inputs, it is also not applicable to simplify the nonlinear observation matrix, or to formulate the input-output equations as shown in (Leuschen et al., 2005).

Instead of manipulating the observability matrix, the nonlinear observation matrix *O* is decomposed into two parts: one is with respect to *x*, denoted as O_{NL} while the other one is with respect to $\begin{bmatrix} u & \dot{u} & \ddot{u} & \cdots \end{bmatrix}$ denoted as *EU*. The observation matrix in Eq. (3.3) is revised as

$$O(x_0, u^*) = \begin{bmatrix} h_j \\ \frac{\partial L_f^0 h_j}{\partial x} f \\ \frac{\partial L_f^0 h_j}{\partial x} f \\ \vdots \end{bmatrix}_{x_0, u^*} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & \frac{\partial L_f^0 h_j}{\partial u} & 0 & 0 & 0 & \cdots \\ 0 & \frac{\partial L_f^1 h_j}{\partial u} & \frac{\partial L_f^1 h_j}{\partial \dot{u}} & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & 0 & \vdots \end{bmatrix}_{x_0, u^*} \begin{bmatrix} u \\ \dot{u} \\ \ddot{u} \\ \vdots \end{bmatrix} = O_{NL} + EU \tag{3.7}$$

The first part O_{NL} in Eq. (3.7) is used to determine Ω in Eq. (3.6) so that the parity (null) space can be formulated. According to the nonlinear observability of general nonlinear systems, the Eq. (3.6) is re-derived as:

$$\Omega^{\perp}\nabla_x O_{NL}(x_0, u^*) = 0 \tag{3.8}$$

where for the j^{th} sensor

$$\nabla_x O_{NL}(x_0, u^*) = \nabla_x \begin{bmatrix} h_j \\ \frac{\partial L_f^0 h_j}{\partial x} f \\ \vdots \\ \frac{\partial L_f^{m_j-1} h_j}{\partial x} f \end{bmatrix}_{x_0, u^*} \tag{3.9}$$

The number of redundancies in this system is determined by the rank of each $\nabla O(x_0, u^*)$. For the j^{th} sensor, the rank is determined as:

$$m_j = rank\left(\nabla_x O(x_0, u^*)\right) \tag{3.10}$$

Here, the system can be either observable or unobservable.

The number of independent analytical redundancies is denoted as:

$$n - N = \sum_{j=1}^{M} (m_j + 1) - N \tag{3.11}$$

n-N sets of independent Ω is determined to formulate the parity space i.e., the dimension of Ω is $\left(\sum_{j=1}^{M}(m_j + 1) - \mathrm{N}\right) \times \sum_{j=1}^{M}(m_j + 1)$.

The second part - *EU* in the Eq. (3.7) mainly represents the information of inputs, and is used to construct analytical redundancy structure in parity space. *EU* is moved to the side of *y* in

Eq. (3.3) so that parity relations for the general nonlinear systems can be formulated as:

$$P = \Omega^{\perp}\left\{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{bmatrix} - \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_M \end{bmatrix} U\right\} = 0 \tag{3.12}$$

where P is parity residuals, which has the dimension of $\sum_{j=1}^{M}(m_j + 1) \times 1$ and ideally, will be nonzero only if a failure presents and the E_j matrix is derived as

$$E_j = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & \cdots \\ 0 & \frac{\partial L_f^0 h_j}{\partial u} & 0 & 0 & 0 & \cdots & \cdots \\ 0 & \frac{\partial L_f^1 h_j}{\partial u} & \frac{\partial L_f^1 h_j}{\partial \dot{u}} & 0 & 0 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \frac{\partial L_f^{m_j-1} h_j}{\partial u} & \frac{\partial L_f^{m_j-1} h_j}{\partial \dot{u}} & \cdots & \frac{\partial L_f^{m_j-1} h_j}{\partial u^{m_j-1}} & 0 & \cdots \end{bmatrix} \tag{3.13}$$

where $E = \begin{bmatrix} E_1 & E_2 & \cdots & E_M \end{bmatrix}$ is a $\sum_{j=1}^{M}(m_j+1) \times \max(m_j) Q$ matrix when the rank of j^{th} sensor is m_j. $U = = \begin{bmatrix} U_1 & U_2 & \cdots & U_M \end{bmatrix}$ is a matrix with dimension $\max(m_j) Q \times 1$ and $U^q = [\ u_q \quad \frac{du_q}{dt} \quad \cdots \quad \frac{d^{\max(m_j)} u_q}{dt^{\max(m_j)}}\ ]^T$ for the q^{th} input.

Essentially, the parity relations can be viewed as a weighted combination of sensor outputs and actuator inputs. The structure of a parity relation defines what should be included, while the coefficients of the parity relation determine the weights.

Table 3-1 lists the comparison of parity residuals generation in linear systems and general nonlinear systems. It shows that the notion of nonlinear analytical redundancy with the parity relation method for general nonlinear systems is analogous to the notion of linear analytical

redundancy.

Table 3-1: Parity relations in linear systems vs. general nonlinear systems

	Linear System	Nonlinear System
System	$\dot{x} = Ax + bu$ $y = cx + du$	$\dot{x} = f(x,u)$ $y = h(x,u)$
	$\mathrm{Y} = \begin{bmatrix} y \\ \dot{y} \\ \vdots \\ y^{m_j} \end{bmatrix} = \begin{bmatrix} c \\ cA \\ \vdots \\ cA^{m_j} \end{bmatrix} x - \begin{bmatrix} d & 0 & \cdots 0 \\ cb & d & \cdots 0 \\ \vdots & \vdots & \ddots 0 \\ cA^{m_j-1}b & cA^{m_j-2}b & \cdots 0 \end{bmatrix} u$	$\mathrm{Y} = \begin{bmatrix} y \\ \dot{y} \\ \vdots \\ y^{m_j} \end{bmatrix} = \begin{bmatrix} h \\ L_f^1 h \\ \vdots \\ L_f^{m_j} h \end{bmatrix}, \quad L_f^1 h = \frac{\partial h}{\partial x} f + \frac{\partial h}{\partial u}\frac{du}{dt}$
Number of P (redundancies)	$\sum_{j=1}^{M} (m_j + 1) - N$	$\sum_{j=1}^{M} (m_j + 1) - N$
Null Space	$\Omega^{\perp} O_L = 0 \qquad O_L = \begin{bmatrix} c \\ cA \\ \vdots \\ cA^{m_j} \end{bmatrix}$	$\Omega^{\perp} \nabla_x O_{NL} = 0 \qquad O_{NL} = \begin{bmatrix} h_j \\ \frac{\partial L_f^0 h_j}{\partial x} f \\ \vdots \\ \frac{\partial L_f^{m_j-1} h_j}{\partial x} f \end{bmatrix}$
Parity residual structure	$\mathrm{P} = \Omega^{\perp} \left\{ \begin{bmatrix} Y_1 \\ \vdots \\ Y_{\mathrm{M}} \end{bmatrix}_{\sum_{j=1}^{\mathrm{M}} \mathrm{m_j} \times 1} - \begin{bmatrix} B_1 \\ \vdots \\ B_{\mathrm{M}} \end{bmatrix} U \right\}$	$\mathrm{P} = \Omega^{\perp} \left\{ \begin{bmatrix} Y_1 \\ \vdots \\ Y_{\mathrm{M}} \end{bmatrix}_{\sum_{j=1}^{\mathrm{M}} \mathrm{m_j} \times 1} - \begin{bmatrix} E_1 \\ \vdots \\ E_{\mathrm{M}} \end{bmatrix} U \right\}$
	$B_j = \begin{bmatrix} d_j & 0 & 0 & 0 & \cdot & \cdot & \cdots \\ c_j B & d_j & 0 & 0 & \cdot & \cdot & \cdots \\ c_j AB & c_j B & d_j & 0 & \cdot & \cdot & \cdots \\ \vdots & \vdots & \vdots & \ddots & \cdot & \cdot & \cdots \\ c_j A^{m_j-1} B & c_j A^{m_j-2} B & \cdots & c_j B & d_j & 0 & \cdots \end{bmatrix}$	$E_j = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & \frac{\partial L_f^0 h_j}{\partial u} & 0 & 0 & 0 & 0 & \cdots \\ 0 & \frac{\partial L_f^0 h_j}{\partial u} & \frac{\partial L_f^1 h_j}{\partial \dot{u}} & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \frac{\partial L_f^{m_j-1} h_j}{\partial u} & \frac{\partial L_f^{m_j-1} h_j}{\partial \dot{u}} & \cdots & \frac{\partial L_f^{m_j-1} h_j}{\partial^{m_j-1} u} & 0 & \cdots \end{bmatrix}$

3.4.2 Parity structure and coefficient design

Given the general formulation of parity relations in the previous section, one is faced with the problem of finding the best structure and coefficients so that parity residuals are projected to a space where they are robust to noise but sensitive to sensor failures. The local property of nonlinear observability may lead to different parity structures under different operating points. Moreover, the statistical characteristics of parity residuals may be affected significantly from one operating regime to another because of different system disturbances and measurement noise. In order to provide accurate sensor fault diagnosis, the working space needs to be partitioned to determine local parity structures and coefficients. The automatic working space partition method

will be presented in Chapter 4. While in this section, the robust optimization is designed to find the best choice of coefficients that can make the candidate parity relations close to zero under no-failure conditions, and the resulting parity residuals can provide significant failure signatures to indicate anomalies or sensor failures. At last, post-processing sensitivity analysis is applied to evaluate how different inputs and states affect the optimal solution.

In this study, it is assumed that local operating regimes are well defined. After constructing a set of parity structures for each local operating regime following the methods proposed in Section 3.4.1, it is proceeding to determine the coefficients in parity relations. In many applications, system disturbance and measurement noise within each operating regime will lead to the difficulty of selecting Ω such that $P = 0$. In this case, a robust optimization design is employed. The Ω is determined so that the value of P is minimized in the existence of noise and model uncertainty. The optimization formulation is shown as below:

$$J^* = \min_{\Omega} \max_{\varepsilon,\delta} P^2 \tag{3.14a}$$

$$P^2 = \{\Omega^{\perp}(Y - EU)\}^2 \tag{3.14b}$$

Here, the quantity of $\max_{\varepsilon,\delta} P^2$ is the worst case effect of noise and model uncertainty on the parity relations. A conservative choice is attempted to find the parity coefficients by minimizing the worst case.

However, it has a trivial solution that all coefficients are zero. To provide a meaningful solution, the coefficients Ω are constrained to have unit magnitude. Moreover, based on the mechanism of parity relation, another constraint is that an optimal set of Ω projects the observation matrix to a null space. Therefore, the complete formulation of a robust optimization design is shown as following:

$$\begin{aligned} &\min_{\Omega}\max_{\varepsilon,\delta} P^2 = \{\Omega^{\perp}(Y - EU)\}^2 \\ &s.t.\ \Omega^{\perp}\Omega = 1 \\ &\Omega^{\perp}\nabla O_{NL}(x_0, u^*) = 0 \end{aligned} \tag{3.15}$$

Based on the number of analytical redundancies, n-M sets of independent Ω will be selected.

Notably, the quantity of $\max_{\varepsilon,\delta} P^2$ is dependent on state x and input u, which indicates that the coefficients should be computed at each time step when state x and input u are changing over time. However, it is not desirable to obtain new coefficient all the time. A more applicable approach is to schedule the coefficients based on the operating regimes since a set of coefficients is usually effective for a range of x and u. This indicates that when the state and the inputs are varying at a certain range, the corresponding coefficients are likely to perform closely to the optimum. In this case, appropriate coefficients will be learnt for each operating regime which is characterized by some nominal state x and input u during the training process and can be retrieved for use at corresponding operating regimes. The autonomous process of partitioning the operating regimes with state and input variables and identifying coefficients in each operating regime efficiently will be addressed in Chapter 4.

3.4.3 Post-processing sensitivity analysis

The nonlinear analytical redundancy utilizes the local observability of nonlinear systems to formulate the parity relations, which is only feasible in a local space that is close to x_0 and u^*. To understand how the change of operating conditions affect the effectiveness of each designed analytical redundancy for sensor fault diagnosis, the effect of the change of state x and input u on the optimal objective function - the square of parity residuals is evaluated with a post-processing sensitivity analysis in this section.

Without the loss of generality, the objective function in Eq. (3.15) is first represented by

$$J(\Omega, p) \tag{3.16a}$$

and constraints active at the optimum point are

$$g_1(\Omega) \text{ and } g_2(\Omega, p) \tag{3.16b}$$

where considering the design variables $\{\Omega = [1, \dots, \Omega_K], \mathrm{K} = \sum_{j=1}^{M}(m_j + 1)\}$ to be an implicit differentiable function of p, i.e., $\Omega_k(p)$, and $p = [x, u]$ are design parameters.

To determine how the optimum design will change as a result of changing operating conditions, the total derivative of the objective function with respect to the design parameters of interest (i.e., inputs u and states x) $\frac{dJ}{dp}$ as well as the rates of change of the optimum values of the design variables $\frac{\partial \Omega_k}{\partial p}$ will be determined. Those derivatives are referred to as sensitivity derivatives. Follow the equations of sensitivity derivatives from (Armacost and Fiacco, 1974; Armacost and Fiacco, 1974; Vanderplaats and Yoshida, 1985) that were developed for a constrained optimum regardless of the type of optimization algorithms, the Kuhn-Tucker conditions at the optimum are employed to predict the required derivatives, i.e., for any given design parameters, the first-order necessary condition (Lagrange Multiplier) satisfies when Ω^* is (local) optimal with the following equations:

$$\frac{\partial J}{\partial \Omega_k}(\Omega^*) + \frac{\partial g_1}{\partial \Omega_k}(\Omega^*)\lambda_1 + \frac{\partial g_2}{\partial \Omega_k}(\Omega^*)\lambda_2 = 0 \tag{3.17a}$$

$$g_1(\Omega_k) = 0 \text{ and } g_2(\Omega_k, p) = 0 \tag{3.17b}$$

Eq. (3.17) holds under some general assumptions (Fiacco, 1976) when the design parameters p are changing so that:

$$\frac{d\left(\frac{\partial J}{\partial \Omega_k}+\frac{\partial g_1}{\partial \Omega_k}\lambda_1+\frac{\partial g_2}{\partial \Omega_k}\lambda_2\right)}{dp}=0 \tag{3.18}$$

With the chain-differentiation rule for the composite functions along with the functional relationships in Eq. (3.16), the differentiations in Eq. (3.18) can be expressed as:

$$\frac{\partial^2 J}{\partial \Omega_k \partial p}+\lambda_1\frac{\partial^2 g_1}{\partial \Omega_k \partial p}+\lambda_2\frac{\partial^2 g_2}{\partial \Omega_k \partial p}+\sum_{l=1}^{K}\left(\frac{\partial^2 J}{\partial \Omega_k \partial \Omega_l}+\lambda_1\frac{\partial^2 g_1}{\partial \Omega_k \partial \Omega_l}+\lambda_2\frac{\partial^2 g_2}{\partial \Omega_k \partial \Omega_l}\right)\frac{\partial \Omega_k}{\partial p}+\frac{\partial \lambda_1}{\partial p}\frac{\partial g_1}{\partial \Omega_k}+$$

$$\frac{\partial \lambda_2}{\partial p}\frac{\partial g_2}{\partial \Omega_k}=0 \tag{3.19a}$$

$$\frac{\partial g_1}{\partial \Omega_k}+\sum_{l=1}^{K}\frac{\partial g_1}{\partial \Omega_k}\frac{\partial \Omega_k}{\partial p}=0 \tag{3.19b}$$

$$\frac{\partial g_2}{\partial \Omega_k}+\sum_{l=1}^{K}\frac{\partial g_2}{\partial \Omega_k}\frac{\partial \Omega_k}{\partial p}=0 \tag{3.19c}$$

which can be rewritten with matrix notation as:

$$\begin{bmatrix} S & Z \\ Z^T & 0 \end{bmatrix}\begin{bmatrix} \delta\Omega \\ \delta\lambda \end{bmatrix}+\begin{bmatrix} v \\ w \end{bmatrix}=0 \tag{3.20}$$

where

$$S_{kl}=\frac{\partial^2 J}{\partial \Omega_k \partial \Omega_l}+\lambda_1\frac{\partial^2 g_1}{\partial \Omega_k \partial \Omega_l}+\lambda_2\frac{\partial^2 g_2}{\partial \Omega_k \partial \Omega_l}, \qquad v_k=\frac{\partial^2 J}{\partial \Omega_k \partial p}+\lambda_1\frac{\partial^2 g_1}{\partial \Omega_k \partial p}+\lambda_2\frac{\partial^2 g_2}{\partial \Omega_k \partial p}$$

$$Z_k=\left[\frac{\partial g_1}{\partial \Omega_k},\frac{\partial g_2}{\partial \Omega_k}\right], \qquad w_k=\left[\frac{\partial g_1}{\partial p},\frac{\partial g_2}{\partial p}\right]$$

$$\delta\Omega=\begin{bmatrix} \frac{\partial \Omega_1}{\partial p} \\ \vdots \\ \frac{\partial \Omega_K}{\partial p} \end{bmatrix}, \qquad \text{and} \quad \delta\lambda=\begin{bmatrix} \frac{\partial \lambda_1}{\partial p} \\ \frac{\partial \lambda_2}{\partial p} \end{bmatrix}$$

and the dimension of matrix S is $\sum_{j=1}^{M}(m_j+1)\times\sum_{j=1}^{M}(m_j+1)$, matrix Z is $\sum_{j=1}^{M}(m_j+1)\times 2$ vector v and $\delta\Omega$ are $\sum_{j=1}^{M}(m_j+1)\times 1$, and vectors w and $\delta\lambda$ are 2×1:

After obtaining the solutions of $\frac{\partial \Omega}{\partial p}$ and $\frac{\partial \lambda}{\partial p}$ from Eq. (3.20), the sensitivity derivative of

the objective function can be determined as the total derivative of the composite function J

$$\frac{dJ}{dp} = \frac{\partial J}{\partial p} + \sum_{k=1}^{K} \frac{\partial J}{\partial \Omega_k} \frac{\partial \Omega_k}{\partial p} \tag{3.21}$$

and the sensitivity derivatives of the design variables are

$$\frac{d\Omega}{dp} = \frac{\partial \Omega}{\partial p} \tag{3.22}$$

It is noted that in Eq. (3.20), if the Lagrange multipliers are not available as a by-product of the optimization solution, they can be estimated with Eq. (3.23) (Livesley, 1971).

$$\lambda_i = -\left(\frac{\partial g_i}{\partial \Omega}^T \frac{\partial g_i}{\partial \Omega}\right)^{-1} \frac{\partial g_i}{\partial \Omega}^T \frac{\partial J}{\partial \Omega} \tag{3.23}$$

3.5 A case study

In this section, the proposed nonlinear analytical redundancy via parity relations for sensor fault diagnosis is demonstrated and validated in an R2R registration process, which aims to obtain accurate alignment of successful print patterns on a substrate. As shown in Figure 3-2, the upper layer has been printed in previous operations and will be laminated on the top of the bottom layer with the printed pattern matched with the pattern on the bottom layer. This process is called registration, which requires automatic control of registration errors to maintain its accuracy. Two sensors are installed – a tension sensor and an optical sensor to provide feedback for control systems so that the speed of the driven roller can be adjusted to ensure registration accuracy.

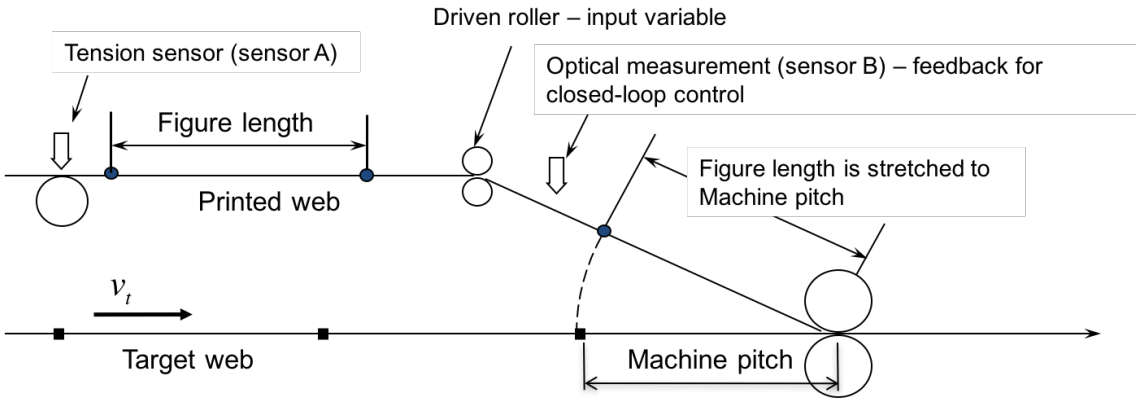


Figure 3-2: A simplified illustration of the R2R registration process

Due to the confidentiality required by our research sponsor, details of the system model are not listed. In the following, a general nonlinear system will be formulated based on the system dynamics of the registration process. The diagnostic capability of the designed analytical redundancies will be demonstrated with normal data from a testbed and abnormal data by introducing simulated sensor faults into the normal data. The effect of the local observability property on parity residuals will be evaluated with the post-processing sensitivity analysis.

3.5.1 Nonlinear analytical redundancies for sensor fault diagnosis

To validate the proposed method, a state space model for a registration process in an R2R manufacturing system is constructed as a general nonlinear system with one state, one actuator input, and two sensor measurements A and B.

$$\begin{aligned} \dot{x} &= f(x,u)+\varepsilon \\ y_A &= h_1(x,u)+\delta_1 \\ y_B &= h_2(x,u)+\delta_2 \end{aligned} \tag{3.24}$$

By examining the observation matrix of y_A and y_B, the rank of each sensor measurement is one so that in total, three analytical redundancies exist in this system. The model estimations of y_A and y_B are compared with real sensor measurements. Figure 3-3 shows that the model

accuracy differs under different operating conditions. Under operating condition 1, the estimation error of sensor A is much smaller than the error generated under operating condition 2. The main reason for this difference is model uncertainty and sensor measurement noise induced by different operating conditions. Here, the entire working space is partitioned into two operating regimes according to the speed of the production system.

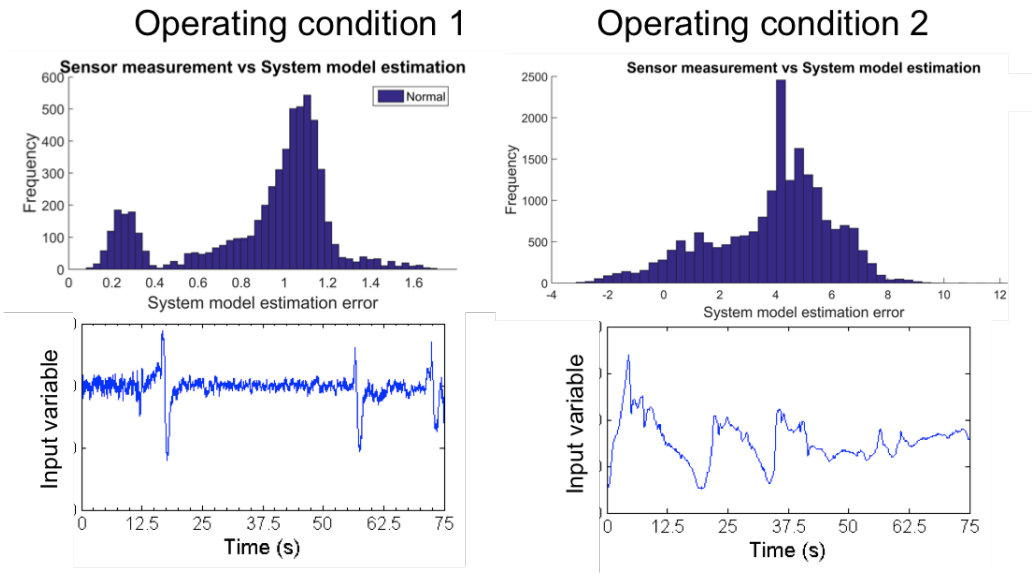


Figure 3-3: State space model estimation error of sensor A under different operating conditions

Based on the observation matrix of the state space model in Eq. (3.15), the parity structure is formulated as:

$$P = \Omega^{\perp}\left\{\begin{bmatrix} y_A \\ \dot{y}_A \\ y_B \\ \dot{y}_B \end{bmatrix} - \begin{bmatrix} 0 \\ \dfrac{\partial h_1}{\partial u} \\ 0 \\ \dfrac{\partial h_2}{\partial u} \end{bmatrix} \dot{u}\right\} \tag{3.25a}$$

and

$$O_{NL}(x_0,u^*)=\begin{bmatrix} h_1 \\ \frac{\partial h_1}{\partial x} f \\ h_2 \\ \frac{\partial h_2}{\partial x} f \end{bmatrix} \tag{3.25b}$$

Based on the rank of the observation matrix of each sensor, there are three independent analytical redundancies available so that three sets of parity coefficients Ω^1, Ω^2, and Ω^3, are determined by the robust optimization design in Eq. (3.15) with nominal data from operating condition 1, shown in Table 3-2.

Table 3-2: Selected parity coefficients under operating condition 1

$\boldsymbol{\Omega}$	y_A	$\dot{y}_A$	y_B	$\dot{y}_B$
$\boldsymbol{\Omega^1}$	-0.2125	-0.0081	0.9771	0
$\boldsymbol{\Omega^2}$	0	0	0.9998	-0.0125
$\boldsymbol{\Omega^3}$	0.9993	0.0369	0	0

3.5.2 Model validation

In order to demonstrate the effectiveness of the designed parity relations based on the proposed method for sensor fault detection in the R2R registration process, different levels of sensor faults are introduced into sensors A and B, respectively. On the left side of Figure 3-3, different amounts of offsets are added to nominal sensor A measurements, while on the right side, different gains are added to sensor B. Those faults are labeled as N05, N15, and P05, P15. The letters N and P present negative and positive offsets/gains that are added to the sensor measurements. The number quantifies the amount of offsets/gains that are introduced into the

sensor measurements. A larger number indicates more severe faults/degradation.

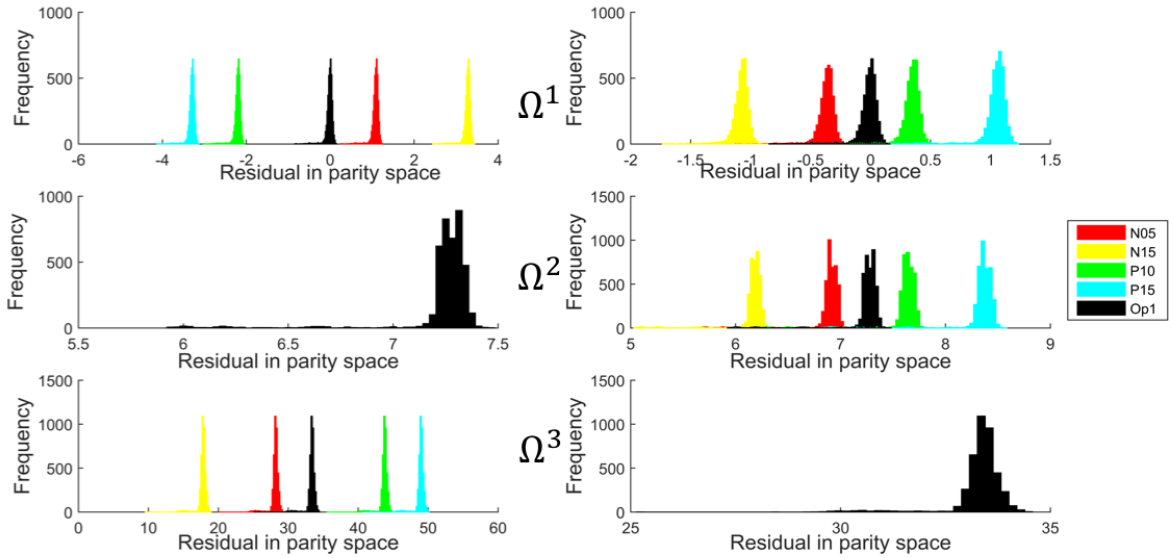


Figure 3-4: Distribution of parity residuals under different levels of sensor faults (Left: offsets added to sensor A; Right: gains added in sensor B)

Figure 3-4 shows that when offsets are added to sensor A, parity residuals governed by Ω^1, and Ω^3 show deviations from their nominal condition, while when gains are added to sensor B, deviations only appear in those parity residuals with Ω^1 and Ω^2. It indicates that under operating condition 1, the parity residuals with Ω^1 contain both information from sensor A and B, while the parity residuals with Ω^3 and Ω^2 only contain the individual information for sensors A and B, respectively. Moreover, when more severe faults are introduced into the sensor measurements, larger deviations can be found in the parity residuals from nominal values. Therefore, those parity residuals can provide effective information for sensor fault detection with the false alarm rate 1.06% under operating condition 1.

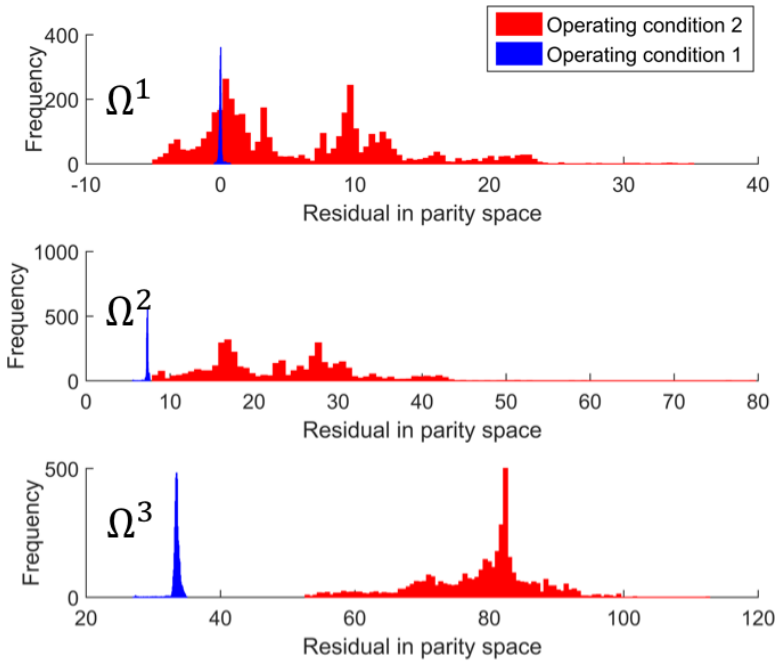
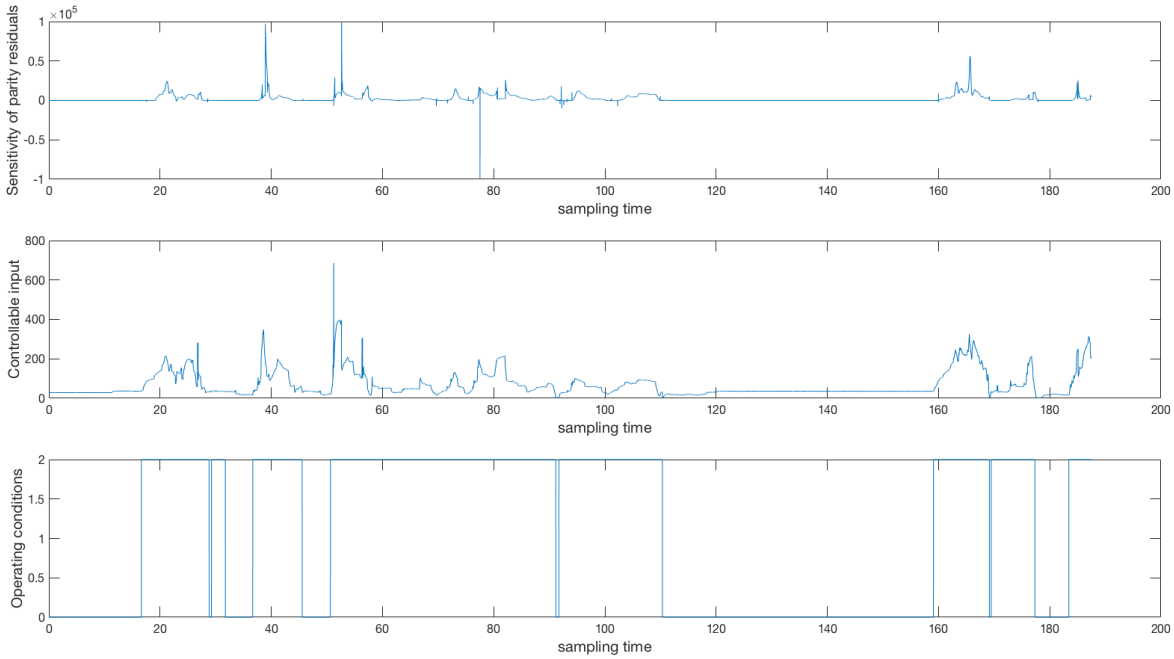


Figure 3-5: Distribution of parity residuals under different operating conditions

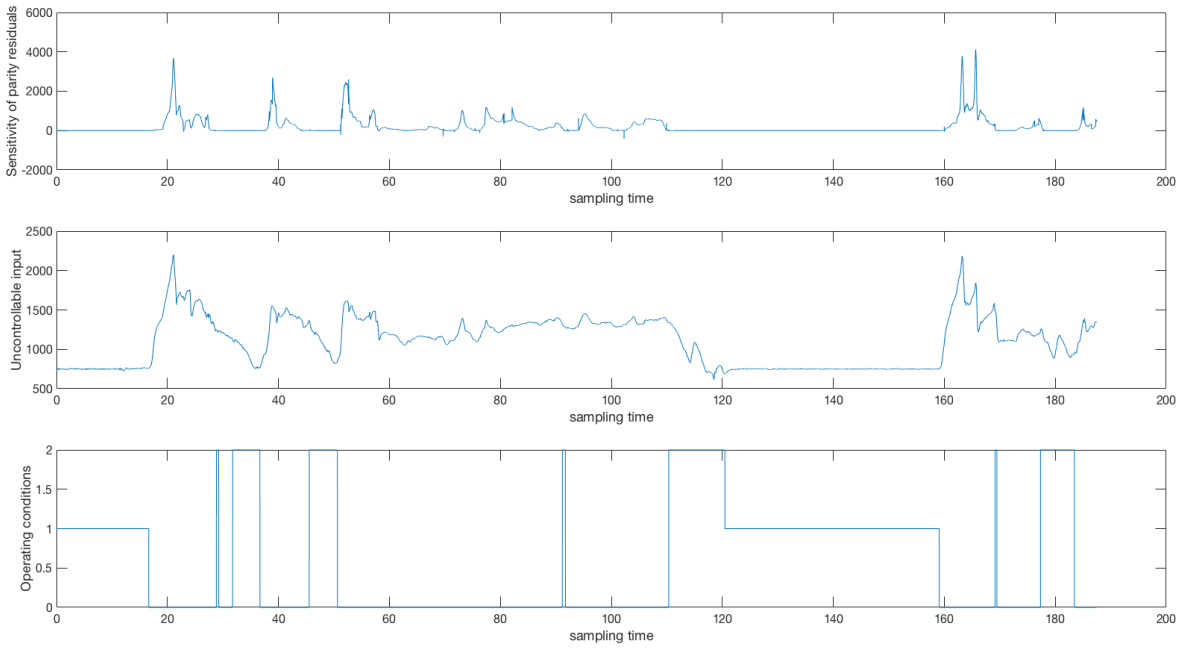
However, when the parity coefficients determined by normal data from operating condition 1 are used to generate parity residuals with nominal data from operating condition 2, the false alarm rate is increased to 91.51%. As shown in Figure 3-5, the deviations will lead us to misjudge system disturbance/sensor noise as a sensor fault. Such high false alarm rate is because the performance of sensor fault detection via parity space is only feasible around a certain range of inputs and state due to the local observability property in nonlinear systems, and is sensitive to the inevitable uncertainty in the knowledge of system dynamics and measurement noise under different operating conditions. Next section shows how the change of operating conditions (i.e., inputs u and state x) affect the parity residuals given the optimal design variables (Ω^*), while the autonomous working space partition and the development of local parity relations will be covered in Chapter 4.

3.5.3 Post-processing sensitivity analysis

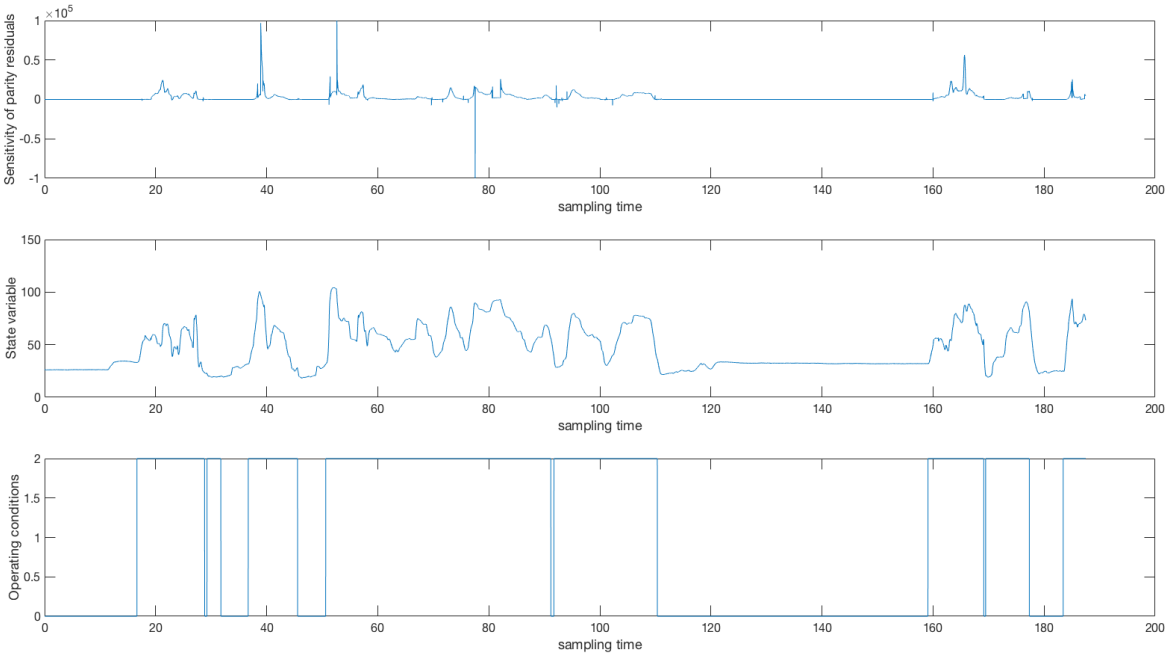
To quantify the effect of the changing operating conditions (i.e., inputs u and state x), the post-processing sensitivity analysis is conducted in this section. The sensitivity derivatives that yield the value of derivatives of the optimal objective function ($\frac{dJ}{dp}$) and design variables ($\frac{d\Omega}{dp}$) with respect to the design parameters of interests are calculated with Eqs. (3.21) and (3.22). The optimal solutions (Ω^*) that are obtained from operating condition 1 with the robust optimization in the last section are employed for the optimum sensitivity analysis. This section only shows the results of the optimal sensitivity analysis with $\Omega^1 = [\Omega_1, \Omega_2, \Omega_3, \Omega_4]$.



a) Sensitivity analysis with respect to the controllable input u_c



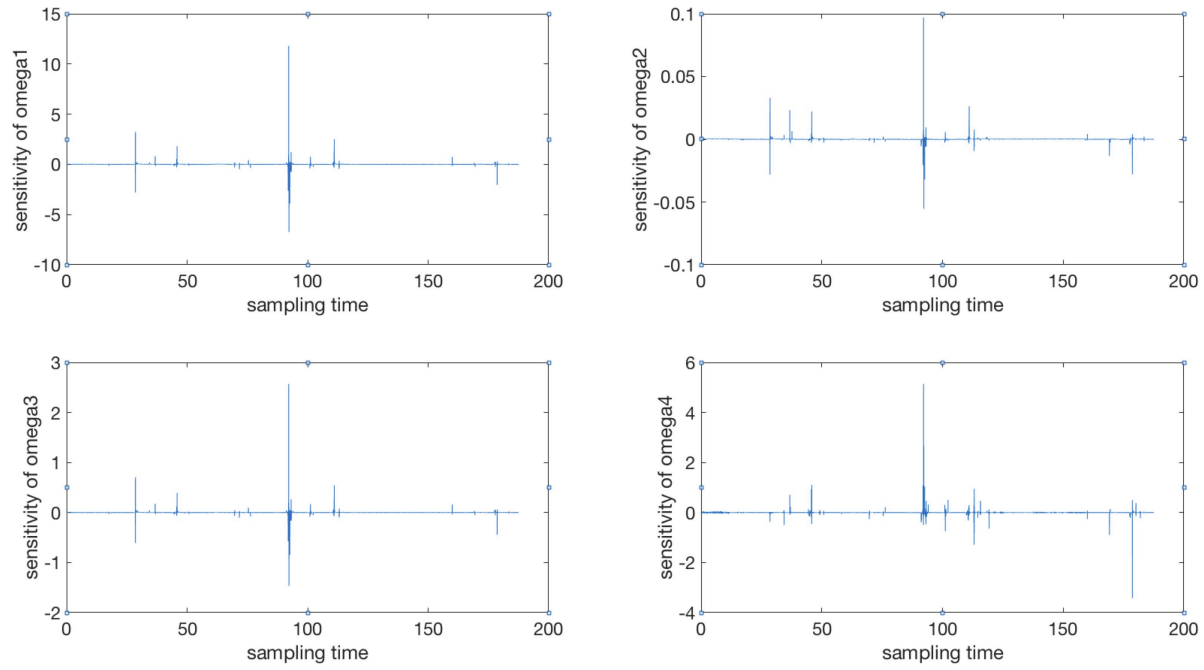
b) Sensitivity analysis with respect to the uncontrollable input u_u



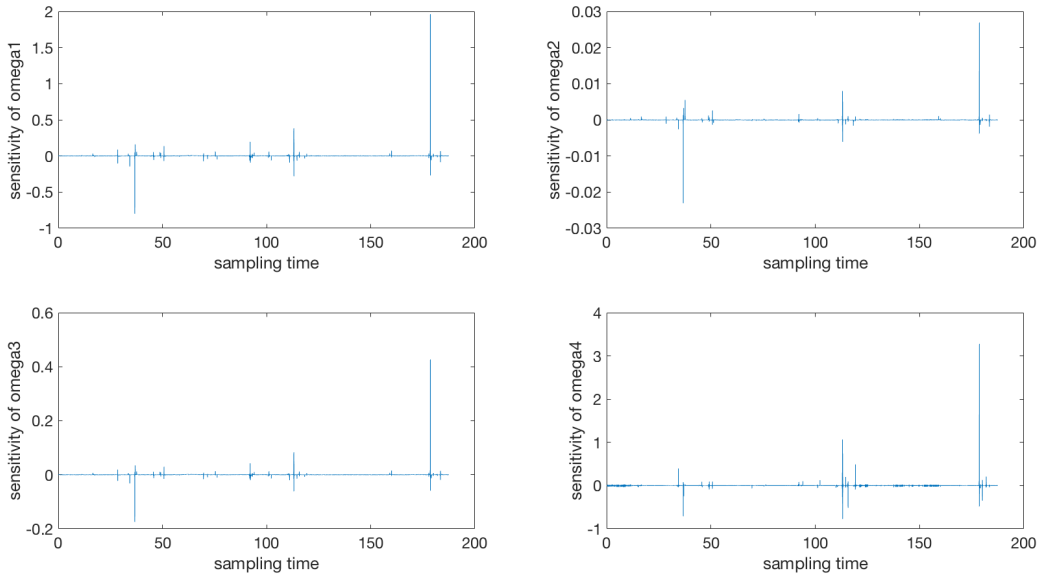
c) Sensitivity analysis with respect to the state variable x

Figure 3-6: The effect of changing operating conditions on parity residuals

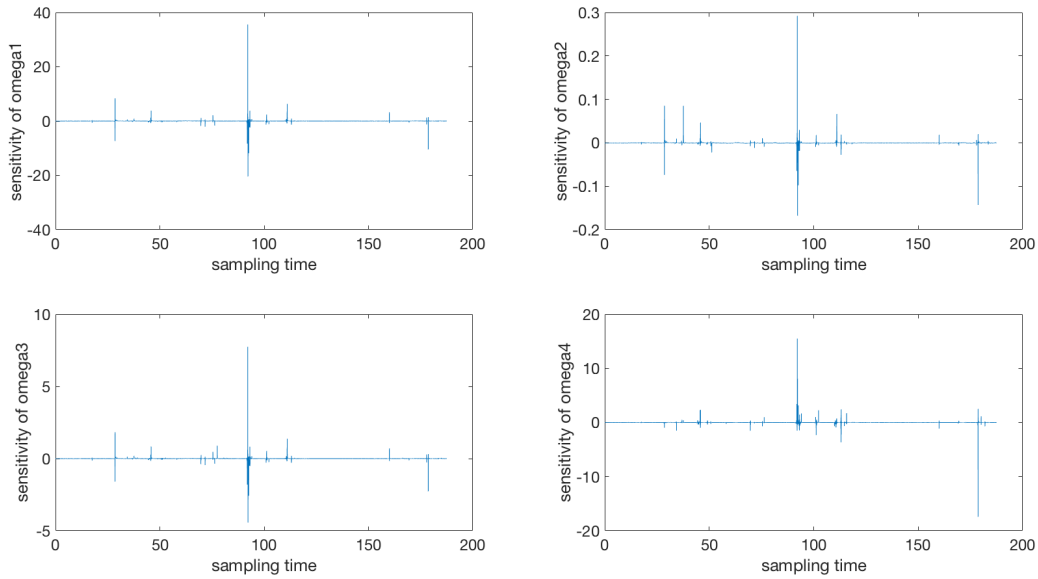
Figures 3-6 a), b), and c) show the sensitivity analysis results of optimal objective function – parity residuals with respect to three design parameters – a) controllable input u_c, b) uncontrollable input u_u and c) state variable x. The change of inputs and state variables leads to the loss function - the parity residuals change dramatically. Such change aligns with the operating conditions (subplots in the third row), which provides the operating range that the designed parity relations are valid for sensor fault diagnosis. Figure 3-7 a), b) and c) are the sensitivity analysis results of optimal parity coefficients with respect to a) controllable input u_c, b) uncontrollable input u_u and c) state variable x. The value of optimal parity coefficients are relatively constant under operating condition 1 while fluctuating under operating condition 2. These observations align with the sensitivity analysis results shown in Figure 3-6.



a) Sensitivity analysis with respect to the controllable input u_c



b) Sensitivity analysis with respect to the uncontrollable input u_u



c) Sensitivity analysis with respect to the state variable x

Figure 3-7: The effect of changing operating conditions on parity coefficients

3.6 Conclusion

This chapter extends the model-based analytical redundancy from linear systems to general nonlinear systems, which fills the gap of analytical redundancy approaches for sensor fault diagnosis in general nonlinear systems, where input and output equations are nonlinear functions of both state variables and input variables. The notion of parity relations is derived based on the nonlinear observation matrix from system dynamic equations, which construct a parity space for sensor fault diagnosis in nonlinear systems. A robust optimization problem is formulated to find the best coefficient for the parity relations against the sensor measurement noise and model uncertainty under a certain operating condition. A post-processing sensitivity analysis is employed to evaluate how the change of the inputs and state affect the optimal objective function – parity residuals and the optimal design variables – parity coefficients. The case study validates the proposed method with data from an R2R registration process. The result shows that the proposed method is capable of identifying sensor degradation with different severity in a nonlinear system under the operating range it designed. The optimal sensitivity analysis quantifies the effect of the change of operating conditions on the diagnostic capability of the trained parity relations and shows its valid local operating range. It demonstrates the necessity of autonomous working space partition to construct local parity relations for sensor fault diagnosis which will be discussed in Chapter 4.

In summary, the nonlinear analytical redundancies developed with the parity space method are effective to detect sensor degradation/faults in a nonlinear system, whose input and output dynamic equations are nonlinear functions of states and inputs. With the robust optimization design, the generated parity residuals are sensitive to sensor degradation/faults but robust to sensor measurement noise and model uncertainty in a certain operating range. This proposed

nonlinear analytical redundancy method for sensor fault diagnosis contributes to the product quality and system productivity improvement in R2R manufacturing systems by eliminating wrong control commands or management decision induced by faulty sensors. It is also applicable for a wide range of nonlinear dynamic systems that other than R2R processes with its benefits of reducing the cost induced by additional hardware sensors and improving the reliability of sensors.

CHAPTER 4 MULTI-REGIME ANOMALY DETECTION AND FAULT DIAGNOSIS

4.1 Introduction

Real-time fault detection and diagnosis is an essential but challenging task in many engineering systems. Most methods are developed for linear systems, which can provide an acceptable performance in the systems with mild nonlinearity or those only operate at a single operating point. However, many systems like R2R manufacturing systems often involve high nonlinearity and it is naïve to assume that it operates only at a single operating point. In an R2R process, operating point continuous to shift due to a variety of disturbances and operating policy changes such as material changeover. Moreover, startups or shutdowns often cause significant shifts in the operating state of the system. During those transitions, diagnosis methods that are designed for linear systems or one operating point are not capable and will lead to erroneous results. Therefore, a real-time fault diagnosis method that is capable of performing under different operating points for nonlinear systems should be developed.

The major challenges of developing effective models for fault diagnosis in such complex engineering systems involve three factors - a priori knowledge of the system, data quality and completeness, and model accuracy. With the increasing complexity of engineering systems and their working environments, to obtain an accurate global model to represent the system dynamics only based on first principles is difficult. Data-driven approaches are employed for the modeling

of complex dynamics systems since it becomes easier to obtain data from sensors and embedded controllers. Neural networks are one of the most popular methods that are used to develop nonlinear models for complex systems due to its universal functional approximating capability. However, this method often suffers from extrapolation and overfitting problem. Also, even though there have been some practical recommendations for the selection of hyperparameters (Bengio, 2012), the selection of activation functions, number of nodes and hidden layers varies in different applications and always a challenging task in literature. Another downside of data-driven methods is that it requires upfront training data from the entire operation space, which is costly and not feasible especially when the deployment of the system is in its early stage. After the global model is learnt, the residuals between the model and the actual system need to be properly interpreted for fault diagnosis. It is common that the model is not perfect and the residuals have different magnitudes under different operational regions. Decision-making algorithms to cope with modeling uncertainties and time-varying process noise need to be designed. However, due to insufficient physical knowledge of the system and unpredictable external influences, not all possible failures under all possible working conditions can be anticipated in advance, which hinders accurate interpretation of the model residuals for fault diagnosis. Therefore, the ability to detect new operating conditions and adapt to them is essential for real-time fault diagnosis in complex systems.

Instead of employing the global model approach, this chapter aims to explore a multiple model approach for fault diagnosis in complex and nonlinear systems based on a "divide and conquer" strategy, which divides the full range of operation into smaller operating regimes and then models the local dynamics individually. With this approach, the modeling tasks for system dynamics under each small operating regime are easier compared to the modeling of the system

as whole. It also enables input-dependent fault diagnosis, which provides more transparency and simplifies the interpretation of the residual errors between the model output and the actual systems. This research will follow the growing structure multiple model systems (GSMMS) proposed in (Liu et al., 2009), which integrates growing self-organizing map to achieve adaptive learning capability for anomaly detection. The difference between this work and the previous work is that when developing the multiple model system, instead of using the linear least squares algorithm for local model identification, this research generalizes the local model identification problem by formulating an optimization problem based on a loss minimization framework and solving it with the mini-batch stochastic gradient descent method. Therefore, the revised GSMMS proposed in this research can be applied to a wider range of applications.

In the followings, the research works of the multiple model approaches for complex systems anomaly detection and fault diagnosis are reviewed in Section 4.2. The contribution of this research work is also summarized in this section. Section 4.3 describes the integration of growing self-organizing map for operation space partition and gradient descent algorithms for local model identification. The case study to demonstrate and validate the proposed methodology with the sensor fault diagnosis problem is presented in Section 4.4. Discussion and conclusions are given in Section 4.5.

4.2 Literature review

For a complex engineering system, many components are coupled and a wide operating range is involved. It requires a sophisticated mathematical model to describe the system dynamics under its entire operating range, which is often not applicable to obtain in industrial applications. To provide more autonomous, intelligent and user-friendly tools, multiple model and operating regime approaches are studied, in which a global model consists of multiple local

models, and each local model has a simpler structure and a certain range of validity (operating regime) less than its entire operating range. The rationale behind this approach is that the system dynamics in each operating regime will be simpler so that the development of several local models is easier than that of one global model.

Many research works in operating regime decomposition or multiple model approaches have been explored for modeling, monitoring and control of complex and nonlinear systems. In general, this approach consists of two steps. The first step is to divide the entire operation range into multiple either exclusive or overlapping operating regimes based on some characterization variables. The selection of those variables is system dependent, which can include manipulated inputs, measured outputs and auxiliary variables. The second step is to identify a local model for each operating regime. Both physical models and data-driven models have been developed to describe the system dynamics in the literature. In the following, previous research work in this area is reviewed.

4.2.1 Multiple model approaches for complex systems

In many engineering systems, the effective monitoring, control and diagnosis need the use of nonlinear models instead of the standard linear time-invariant models. Most methodologies developed for linear systems can only be satisfactorily used in mildly nonlinear systems under certain circumstances. Some nonlinear models are developed based on a priori knowledge of the system, which is complex, need an intensive understanding of the system itself and are not applicable for many industrial applications so they often require some simplifications. For those complex systems that a priori knowledge is partially or totally unavailable, data-driven methods such as kernel estimators, artificial neural networks and fuzzy models were proposed (Suykens et al., 1996; Babuška, 1998; Nelles, 2002; Chen et al., 2005). There are also some research works

that have explored hybrid approaches to leverage the advantages of both the a priori knowledge and input/output data to build up models for complex systems (Hofleitner, 2013; Shui et al., 2018). Nevertheless, the global model approach, which depends on one model to describe the system behaviors under its entire operating range often suffers from fundamental limitations to address highly nonlinear behaviors or complex interactions.

To address this problem, the *divide and conquer* strategy is proposed. The basic idea is to partition a complex problem into several simpler and solvable sub-problems so that simpler and adequate solutions can be designed. Many research works in nonlinear modeling, identification and control have studied the decomposition of operating regimes and the development of multiple models to approximate system dynamics in complex systems. Takagi and Sugeno proposed a rule-based fuzzy modeling method using a fuzzy partition of input space to form linear input-output relation in each fuzzy subspace with the following format (Takagi and Sugeno, 1985):

$$R_m: if\ x_1\ is\ in\ A_{m1}, \dots x_k\ is\ in\ A_{mk}, \quad then\ y_m = g_m(x_1, \dots, x_k) \tag{4.1}$$

where R_m is a fuzzy implication. $x_1, \dots, x_k$ are variables of the premise and y_m is variable of the consequence in the m^{th} implication. $A_{m1}, \dots, A_{mk}$ are membership functions of the fuzzy sets in the premise of the m^{th} implication. g_m is a local model that describes the input-output relation in the m^{th} implication. The final output y is inferred from n implications and is presented as the weighted average of all y_m so that the fuzzy partition has smoothed connections between different regions.

Another approach is proposed by Johansen and Foss in the following (Johansen and Foss, 1993,1995, 1997):

$$y = \sum_{m=1}^{M} v_m(x) g_m(x) \tag{4.2}$$

$$v_m(x) = \frac{\rho_m(x)}{\sum_{m=1}^{M} \rho_m(x)} \tag{4.3}$$

where g_m is the local model and y is the global output. ρ_m is the validation function that indicates the validity of the local model in each operating regime.

Nevertheless, a proper partitioning of the operating range is one key element for the successful deployment of the multiple model approach, which is one of the major challenges in this area of research. There have been different methods developed to decompose the operating range and can be categorized based on deterministic or stochastic assumption (Blom and Bar-Shalom, 1988; Jordan and Jacobs, 1994; Petridis and Kehagias, 1996), soft (Takagi and Sugeno, 1985; Johansen and Foss, 1993) or hard (Sanger, 1991; Barton and Pantelides, 1994; Bencze and Franklin, 1994) partition, and homogeneous or heterogeneous partition (Orjuela et al., 2013). However, the early work of operating regime partition is heuristic, often through offline or trial-and-error approaches. To enable an autonomous partitioning process, the fuzzy adaptive resonance theory algorithm was proposed (Tzafestas and Zikidis, 2001) to determine both local model parameters and model structures by specifying fuzzy rule splitting and addition mechanisms to improve the converge of regions. Uosaki and Hatanaka proposed a hybrid regime selection method to improve the quality of the global model based on three criteria - Kullback Discrimination Information, Akaike Information Criterion and Mean Square Error, in partition and integration of regimes to build up suitable local regimes (Uosaki and Hatanaka, 2008). Kohonen's self-organizing map (SOM) using vector quantization techniques was used to improve the operating regime partition for multiple model approaches in nonlinear systems (Nelles, 2001). This method overcomes the limitation of hyper-rectangle local model domains in

previous works. Moreover, instead of dividing individual input variables, the operation space is directly partitioned with Voronoi tessellation by defining a set of weight vector $\{\varepsilon_m, m = 1, \dots, M\}$ in the Euclidean sense (Martinetz et al., 1993):

$$R_m = \{\boldsymbol{x}: \|\boldsymbol{x} - \varepsilon_m\| \leq \|\boldsymbol{x} - \varepsilon_i\|, \forall i \neq m\} \tag{4.4}$$

where $\boldsymbol{x}$ is the input variables. As shown in Figure 4-1, the weight vector defines the central location of each operating regime. The Voronoi tessellation ends up forming the boundaries of each operating regime by aggregating "similar" input-output patterns into one sub-region through unsupervised clustering. Moreover, the growing mechanism is further explored and added to SOM, such as growing cell structure and growing neural gas, which enables SOM to grow to an appropriate size (Fritzke, 1994b, 1994a). Such autonomous addition and removal capability in SOM can facilitate operating regime partition in multiple model approach (i.e., the number of local models can autonomously grow and adapt to new data). Liu introduced the growing structure multiple model system (GSMMS) method, which utilized a growing SOM to decompose input-output space of a dynamic system into sub-regions and identify local model parameters to describe the system behavior in each sub-region with the least squares method (Liu et al., 2009).

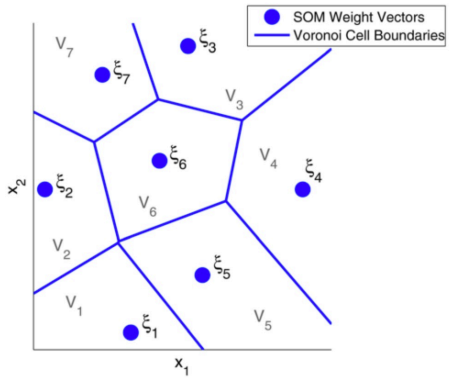


Figure 4-1 Voronoi tessellation with SOM weight vectors

4.2.2 Multiple model approach for fault diagnosis and prognosis

In fault diagnosis and prognosis problems, the multiple model and operating regime approaches have attracted significant attentions. Takagi-Sugeno fuzzy models (TSFMs) that represented local dynamics in different state space regions by local linear systems were employed to generalize linear parity relations for fault estimation in nonlinear systems (Nguang et al., 2007). Deshpande utilized a Bayesian approach to identify local models and generalized likelihood approach for fault identification in a nonlinear system (Deshpande and Patwardhan, 2008). In (Wang et al., 2008), degradation patterns were characterized under different operating regimes and a similarity-based prognostic approach was studied to estimate the remaining useful life. Operation-specific hidden Markov models were developed to characterize degradation process in a semiconductor manufacturing system (Bleakie and Djurdjanovic, 2016). However, those approaches lack adaptive capability to learn new operating regimes automatically from new data. GSMMS that integrated the growing SOM with efficient local model parameter estimation was proposed for anomaly detection and fault diagnosis in nonlinear dynamic systems (Liu et al., 2009). This method is capable of discovering new operating regime and refining local models with new data from the system so that it can generate operating regime dependent residuals for more reliable anomaly detection. Moreover, the authors have compared the model accuracy, training and testing time among the proposed GSMMS method and other methods such as the TSFM, NARX and ARX, and showed that the GSMMS outperformed those methods because of its relatively high model accuracy and fast speed during testing. Later, the GSMMS was further employed for precedent free fault isolation in diesel engine exhaust gas recirculation system (Cholette and Djurdjanovic, 2012) and quality estimation in a semiconductor manufacturing process (Bleakie and Djurdjanovic, 2016). Nevertheless, those studies are

restricted in the implementations that local model parameters can be identified by a closed-form solution with the linear least squares method.

To address this weakness, this research revises the GSMMS algorithm by formulating the local model identification problem into an optimization problem based on the loss minimization framework and solving it with a gradient descent method.

4.3 Methodology

Following the multiple model approach proposed by Johansen and Foss (Johansen and Foss, 1993, 1995, 1997), the global model is defined as:

$$\mathrm{F(t)} = \sum_{m=1}^{M} v_m(s(t)) f_m(s(t)) \tag{4.5}$$

$$v_m(s(t)) = \begin{cases} 1 & s(t) \in V_m \\ 0 & otherwise \end{cases} \tag{4.6}$$

where $V_m, m = 1, \dots, M$ is a disjoint partition of the operating space that particular model input vectors $s(t)$ reside. Model input vectors s(t) can consist of manipulated inputs, measured outputs and auxiliary variables at a given sampling time $\mathrm{t}, \mathrm{t} = 1, \dots, \mathrm{T}$. $\mathrm{F(t)}$ is the global output, which is a weighted summation of the output from each local model $f_m(s(t))$. The validity function v_m indicates how much each local model output contributes to the global output. Here, it is defined with a gating function so that only one local model will be used to describe the system dynamics at any given sampling time t. This provides tractability, which will benefit the model parameter estimation during learning, model stability and facilitate describable control capabilities.

From the Eqs. (4.5) and (4.6), two major elements need to be identified to describe a complex system with the multiple model approach – V_m that partitions the operating space and $f_m\big(s(t)\big)$ that represents the local dynamic behaviors in the system. Following the training

process of the GSMMS method, the multiple model system is established by identifying 1) structural parameters – number and locations of the weight vectors in the SOM, and 2) local model parameters. In this section, a revised version of the GSMMS is introduced, which determines the structural parameters with the growing SOM algorithm, and formulates local model identification into an optimization problem based on the loss minimization framework solving with a gradient descent method.

4.3.1 Identification of structural parameters

The growing SOM method involves three learning processes – competitive learning, cooperative learning and adaptive learning (Fritzke, 1994b). The competitive learning process means with a given SOM network as shown in the left-hand side plot of Figure 4-2 a), each node will try to compete with the others and win new sample data based on the Euclidean distance. The winning node is defined as the best matching unit (BMU) and the new sampling data *s* will be assigned to it as shown in the right-hand side plot of Figure 4-2 a). Second, with the new sample data *s*, the weight vectors of each node will be updated with the cooperative learning capability, i.e., adjusting the location of the BMU as well as its direct topological neighbors. Therefore, the distance among each node shown in the left-hand side plot of Figure 4-2 b) is changed and an updated SOM network is presented in the right-hand side plot of Figure 4-2 b). At last, the existing node that has the highest value such as visiting frequency or quantization errors will be marked shown in left-hand side plot of Figure 4-2 c). If it exceeds a predefined threshold, a new node will be inserted between this node and its furthest direct neighbor to reduce its value as shown in the right-hand side plot of Figure 4-2 c). This process is called adaptive learning.

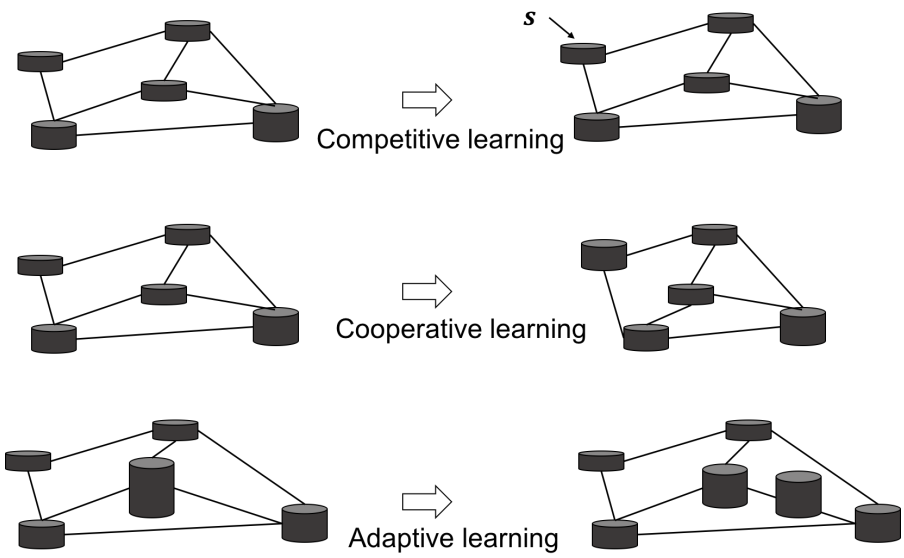


Figure 4-2: Learning mechanisms of a growing SOM (Fritzke, 1994b)

In this study, the growing SOM is employed to facilitate the operating regime partition in the GSMMS. However, instead of using the visiting frequency or quantization errors, the growth of the GSMMS is according to the fitness of the local model in each operating regime.

The SOM weight vectors $\{\varepsilon_m, m = 1, \dots, M\}$, which determine the center location of each operating regime are adjusted with the recursive updating format:

$$\varepsilon_m(k+1) = \varepsilon_m(k) + \zeta_m(k)h\big(k, dis(m,c)\big)[s - \varepsilon_m], \;\; m = 1, \dots, M \qquad (4.7)$$

where $dis(.,.)$ is the topological distance between two operating regimes, which is computed using the breadth-first procedure (Cormen et al., 2009). c is the BMU - the index of the weight vector that is closest to the training input s. k is the number of updates of weight vectors in a GSMMS network with a fixed number of operating regimes. The maximum number of k is predefined to stop the weight vectors updates. k will be reset to 0 when a new operating regime is added to the network.

For each input $s(t)$ at a given sampling time t, its BMU $c(t)$ will be determined by Eq.

(4.8). The matching process of the input s and the BMU c indicates the competitive learning capability of the growing SOM method. Operating regimes compete with each other and use the data they won to determine local models in the corresponding regimes.

$$c(t) = \arg\min_{m} \|s(t) - \varepsilon_m\| \tag{4.8}$$

The neighborhood function $h\big(k, dis(m,c)\big)$ is defined with Gaussian kernel shown in Eq. (4.9). The use of the neighborhood function is a cooperative learning process among each operating regime. The neighborhood function updates the weight vectors not only with the training data falling into the corresponding operating regime (Voronoi set), but also the training data in the neighboring regimes.

$$h\big(k, dis(m,c)\big) = exp\left(\frac{-dis(m,c)^2}{2\sigma^2(k)}\right) \tag{4.9}$$

where σ is a non-increasing function of time and defines the effective range of the neighborhood function.

The penalty term $\zeta_m(k)$ is determined by the value of the local loss function J_m that established for local model parameter identification. The increment of ζ_m will lead to the weight vector move toward regimes with higher loss functions. It aligns with the adaptive learning capability of the growing SOM but in this research, it employs different criteria for growth - the value of the loss function replaces the visiting frequency or quantization errors that are used in the traditional growing SOM method.

$$\zeta_m(k) = \frac{J_m(k)}{\sum_{m=1}^{M} J_m(k)} \tag{4.10}$$

If the largest ζ_m exceeds a predefined threshold, a new node can be inserted between the node with the largest ζ_m and its furthest neighbor according to the Euclidean distance. This leads

to a finer partition of the operating regime that cannot be sufficiently described by current local model and requires further decomposition.

At last, the growth of the SOM network is terminated based on two stopping criterion: 1) all $\zeta_m, \mathrm{m} = 1, \ldots, \mathrm{M}$ are below the predefined threshold, and 2) the number of SOM nodes (number of operating regimes) exceeds a predefined number.

4.3.2 Identification of local model parameters

After the operating regime is partitioned, the local model will be developed based on the training data falling into the corresponding regime. Previous work in the GSMMS has demonstrated the development of local models with linear least squares. In this section, the learning process of local model parameters is cast as an optimization problem based on the loss minimization framework and solved with the gradient descent method. This approach can handle those local models that do not have a closed-form solution (e.g., logistic regression) or are nonlinear (e.g., quadratic).

A loss function quantifies how undesirable it is to use the parameters β_m for prediction on x when the correct output is y. To obtain optimal parameters for the model that has the maximum similarity with the real behaviors, an optimization problem is formed to minimize the weighted sum of the loss function J_m as shown below:

$$J_m{}^* = \min_{\beta_{m1},\ldots,\beta_{mp}} \frac{1}{T}\sum_{t=1}^{T} w_m(s(t)) J_m(\beta_m; s_m(t)) \tag{4.11}$$

where weighting function w_m vector is defined as $w_m(s(t)) = exp\,(\frac{-dis(m,c(t))^2}{2\sigma^2})$. It can smooth the discontinuities along the boundaries of adjacent regions; the further the operating regime m away from the BMU c(t), the less impact of the current training sample on the local model

identification. $\beta_m = [\beta_{m1}, \dots, \beta_{mp}]$ are the parameters for the local model in the mth operating regime, and s_m are training samples that are assigned to the corresponding regime. The loss function J_m evaluates how much the current local model is deviated from the real system behaviors, which is used as the splitting criteria for operating regime partition in Section 4.3.1.

Typically, the training loss (empirical risk or training error) is minimized by determining the minimum average loss of all training samples, which requires making tradeoffs across all samples since it is difficult to find a set of parameters that makes every sample with a small loss in real applications. A general approach to achieve that is to use the iterative optimization – gradient descent, which starts at some point β_m (e.g., all zeros), and tries to tweak the parameters based on the gradient of the function. The gradient provides the direction to move in to decrease the objective the most. The gradient descent method is one of the most popular algorithms to perform optimization and has been widely used for machine learning algorithms. It has two hyperparameters – 1) step size η and 2) the number of iterations, which need to be customized based on different optimization problems (Ruder, 2016).

The standard gradient descent algorithm – batch gradient descent finds the optima with the entire training dataset:

$$\beta_m{}^{new} = \beta_m{}^{old} - \eta \nabla_{\beta_m} J_m(\beta_m; s_m) \tag{4.12}$$

The parameters β_m are updated in the opposite direction of the gradient of the loss function w.r.t. $\nabla_{\beta_m} J_m$. The step size η is an important hyperparameter that determines how fast to converge to a (local) minimum. Larger step sizes are likely to drive faster so that will have faster convergence but may get overshoot and end up unstable results. Smaller step sizes, on the other hand, lead to very slow convergence to the destination. A common strategy to define the step

size is to set it as a decreasing function of the number of updates. In this work, the initial step size is set to be one and decreases as the inverse of the root square of the number of updates that have been taken so far.

One downside of the batch gradient descent method is that it is slow since the training loss is based on a sum over all training data. For each iteration, the gradients for the whole dataset need to be calculated. It re-computes gradients for similar examples so that performs redundant computations for large datasets. Therefore, it is intractable for large datasets that don't fit in memory and usually does not allow online model updates. Stochastic gradient descent addresses this redundant computation problem by updating parameters based on each training sample instead of looping through all training samples shown in Eq. (4.13). It is a stochastic approximation of the gradient descent optimization, which usually is much faster - the cost of each update is reduced from $\mathcal{O}(\mathrm{Tp})$ to $\mathcal{O}(\mathrm{p})$ (Kasai, 2017), and can be used for online model updates. Even if the stochastic gradient descent may take more updates than batch gradient descent, with large data sets, it usually prefers to have many updates based on cheap estimates of the gradient rather than few updates based on good but expensive ones. Moreover, the noisiness of the stochastic gradient descent can help escape from saddle points or local minima in non-convex optimization problems.

$$\beta_m{}^{new} = \beta_m{}^{old} - \eta \nabla_{\beta_m} J_m(\beta_m; s_m(t)) \tag{4.13}$$

Nevertheless, since the stochastic gradient descent method updates is based on one training sample, it sometimes suffers from large variance that results in unstable convergence and objective function to fluctuate severely. To balance the tradeoff between the variance (robustness of the stochastic gradient descent) and the speed (efficiency of batch gradient descent), a compromise approach is to compute gradient against more than one training samples at each

update, which is called mini-batch stochastic gradient descent shown in Eq. (4.14). This approach results in smoother convergence since it computes gradient over more training samples, and remains its capability of reducing the computation cost and achieving better results as it can kick the objective function from local saddle point. It can also take advantage of vectorization and paralleling implementation to further speed up the training process.

$$\beta_m{}^{new} = \beta_m{}^{old} - \eta \nabla_{\beta_m} J_m(\beta_m; s_m(t{:}\, t+n)) \tag{4.14}$$

In this research work, the mini-batch stochastic gradient descent method is employed to solve the optimization problem. The step size is defined as

$$\eta = \frac{1}{idx^2}\ , idx = 1, \dots, I \tag{4.15}$$

where idx is the number of local model parameters updates that have been taken so far with the gradient descent method.

To sum up, the flowchart of the sequential training process of the revised growing structure multiple model system is shown in Figure 4-3.

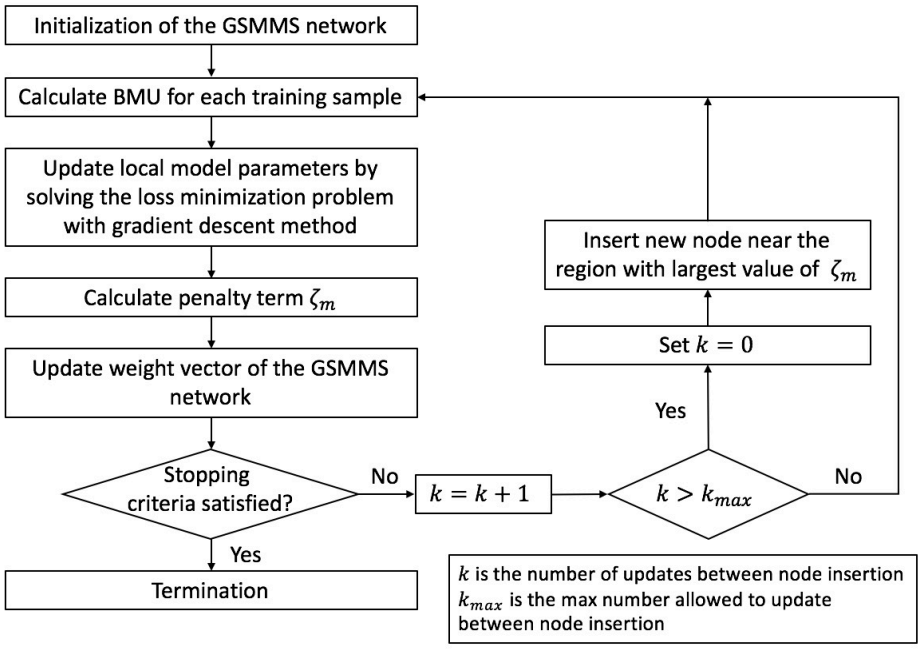


Figure 4-3: Flowchart of the sequential training process with the revised growing structure multiple model system

4.4 Case study

In this section, the proposed modeling method for multiple regime anomaly detection and fault diagnosis is demonstrated and validated through a case study of sensor fault diagnosis under multiple operating conditions. Data from a registration process in an R2R manufacturing system is employed for model training and testing. Data under both normal operating condition and with sensor failures are collected to demonstrate the effectiveness of the revised GSMMS algorithm: 1) normal data from operating conditions 1 and 2, respectively, and 2) abnormal data with one sensor degraded under operating conditions 1 and 2, respectively.

4.4.1 Multiple regime sensor fault diagnosis in R2R manufacturing systems

For a high-throughput R2R manufacturing system, sensors and inspection systems are installed to feedback information for system operation, supervise the system performance and

guarantee the product quality. However, those sensors often work under severe environment with high temperature or strong vibration induced by rollers, which results in sensor degradation and failures over time. To ensure sensor performance and avoid wrong decisions due to sensor failures, real-time sensor fault diagnosis should be realized in the production systems. In Chapter 3, a generalized nonlinear analytical redundancy method has been introduced for sensor fault diagnosis. However, it has been pointed out that the performance of analytical redundancy methods varies, depending on the inevitable uncertainty in the knowledge of system dynamics and measurement noise under different operating conditions (Chow and Willsky, 1984; Qin and Li, 1999). As shown in Section 3.5.1, when the parity coefficients determined by nominal data from operating condition 1 are used to generate parity residuals with nominal data from operating condition 2, it results in deviations that misjudge system disturbance/sensor noise as a sensor fault. Therefore, in this section, multiple regime sensor fault diagnosis is achieved by the proposed method to address this problem by integrating the growing SOM for operating regime partition and parity space approach for local model identification.

The overall framework is shown in Figure 4-4. Given a dynamic system with multiple sensors, the dynamic model is obtained based on physics and engineering knowledge, and input signals and output (sensor measurements) are collected for the GSMMS network training. In the GSMMS network, each node represents one operating regime and is updated according to the parity residuals that are generated from local parity relations in the corresponding operating regime. The determination of parameters for parity relations is converted into solving the loss minimization problem with the mini-batch stochastic gradient descent method.

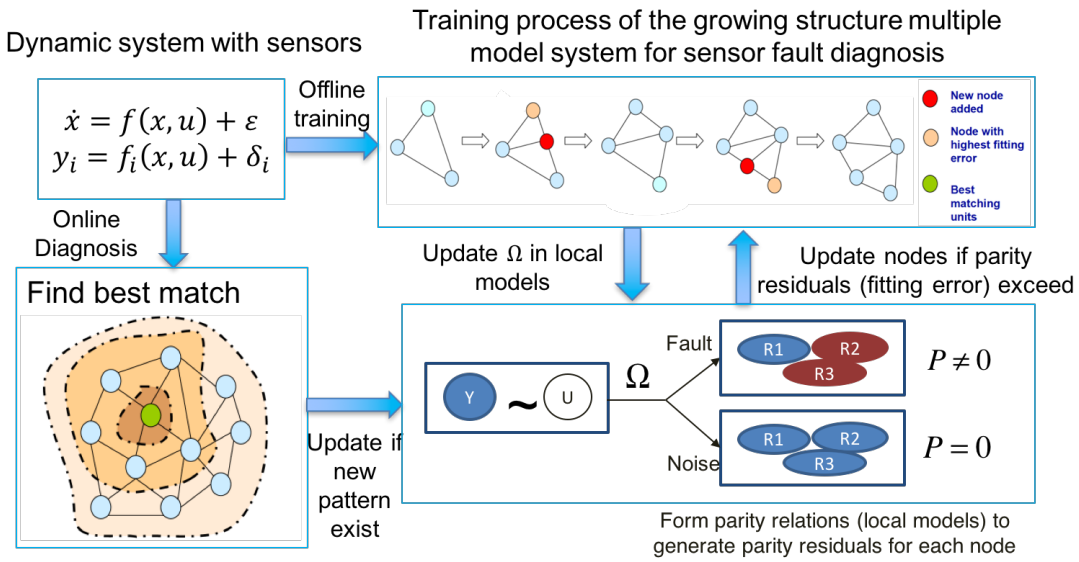


Figure 4-4: An operation dependent sensor fault detection scheme

4.4.2 GSMMS network training with parity space approach

As shown in Eq. (4.16), the sensor measurements consist of three parts – system model estimation, measurement noise and model uncertainty, and deviations that induced by sensor degradation/failures. Different operating conditions such as material changeover in R2R processes or various production speeds might result in different magnitude of the measurement noise and model uncertainty. To detect and differentiate the deviations induced by sensor degradation/failures from measurement noise/model uncertainty, the revised GSMMS network proposed in this chapter is employed.

$$y = \hat{y} + r_{noise+uncertainty} + r_y \tag{4.16}$$

To learn the GSMMS network with the parity space approach for sensor fault diagnosis, the R2R registration process is formed as a nonlinear system with one state, one actuator input, and two sensor measurements A and B. Eq. (4.17) shows the general formulation of the nonlinear system, of which the model details can be obtained based on empirical knowledge and

physical understanding of the system.

$$\begin{aligned} \dot{x} &= f(x,u)+\varepsilon \\ y_A &= h_1(x,u)+\delta_1 \\ y_B &= h_2(x,u)+\delta_2 \end{aligned} \tag{4.17}$$

and its parity relations are formulated as:

$$P = \Omega^{\perp}\left\{\begin{bmatrix} y_A \\ \dot{y}_A \\ y_B \\ \dot{y}_B \end{bmatrix} - \begin{bmatrix} 0 \\ \frac{\partial h_1}{\partial u} \\ 0 \\ \frac{\partial h_2}{\partial u} \end{bmatrix}\dot{u}\right\} \tag{4.18}$$

and

$$O_{NL}(x_0,u^*) = \begin{bmatrix} h_1 \\ \frac{\partial h_1}{\partial x} f \\ h_2 \\ \frac{\partial h_2}{\partial x} f \end{bmatrix} \tag{4.19}$$

Following the optimization design in Chapter 3, optimal model parameters are selected for parity relations so that under a no-fail situation, the parity residuals are close to zero in each operating regime as shown below:

$$\min_{\Omega_m} \quad P_m^2 = \left[\Omega_m{}^{\perp}(\mathrm{Y}-\mathrm{EU})\right]^2 = \left\{\Omega_m{}^{\perp}\left(\begin{bmatrix} y_A \\ \dot{y}_A \\ y_B \\ \dot{y}_B \end{bmatrix} - \begin{bmatrix} 0 \\ \frac{\partial h_1}{\partial u} \\ 0 \\ \frac{\partial h_2}{\partial u} \end{bmatrix}\dot{u}\right)\right\}^2 \tag{4.20a}$$

$$\text{s.t.} \qquad \Omega_m{}^{\perp}\Omega_m = 1 \tag{4.20b}$$

$$\Omega_m{}^{\perp}\nabla\begin{bmatrix} h_1 \\ \frac{\partial h_1}{x} f \\ h_2 \\ \frac{\partial h_2}{x} f \end{bmatrix} = 0 \tag{4.20c}$$

The operating regimes can represent different operating conditions such as during steady-state operation, material changeover or under different operation speeds. To obtain parameters for local parity relations in the corresponding operating regime, the optimization design in Eqs. (4.20) is converted to an optimization problem based on the loss minimization framework shown in Eqs. (4.21) and (4.22):

$$J_{mt}\big(\Omega_m; s_m(t)\big) = \big[\Omega_m{}^{\perp}\big(\mathrm{Y(t)} - \mathrm{EU(t)}\big)\big]^2 + \lambda_1\big(\Omega_m{}^{\perp}\Omega_m - 1\big)^2 + \lambda_2(\Omega_m{}^{\perp}\nabla O_{NL}(x(t), u(t)))^2 \tag{4.21}$$

$$J_m{}^* = \min_{\Omega_{m1},\ldots,\Omega_{m4}} \frac{1}{n}\sum_{t=1}^{n} w_m(s(t)) J_{mt}(\Omega_m; s_m(t)) \tag{4.22}$$

where λ_1 and λ_2 are hyperparameters that regulate the constraint functions in Eq. (4.20). Large values of λ_1 and λ_2 indicate that the loss function will be more sensitive to the violation of the constraints, vice versa. n is the batch size, which usually can be set as power of 2, ranging from 32 to 256 (Ruder, 2016).

The local model parameters Ω_m are determined by the mini-batch stochastic gradient descent in each operating regime, as shown in Eq. (4.23).

$$\Omega_m{}^{new} = \Omega_m{}^{old} - \eta\nabla_{\Omega_m} J_m(\Omega_m; s_m(t{:}\, t+n)) \tag{4.23}$$

where $s_m = [y_A, y_B, \dot{y}_A, \dot{y}_B, x, u, \dot{x}, \dot{u}]$ are the input-output signals from the system. Here the subset of s_m that only contains $[x, u, \dot{x}, \dot{u}]$ is used to partition the operating space.

4.4.3 Validation results

To demonstrate the advantage of the proposed multiple model approach for sensor fault diagnosis, the diagnostic performance under two different modeling scenarios is presented below. The diagnostic accuracy under different operating conditions with and without sensor faults is listed in Table 4-1 to evaluate the diagnostic performance of each modeling scenarios. Measurement deviations are added to sensor A to simulate different severity of sensor degradation - 5%, 10%, 15% and 20% (P5, P10, P15 and P20 respectively) in the case study. The MATLAB library – SGDLibrary is employed (Kasai, 2017) to solve the optimization problem. The batch size for the mini-batch stochastic gradient descent method is tuned and set as 128 during the training process. Figure 4-5 shows the training time of different batch size with the mini-batch stochastic gradient descent method, while Table 4-1 compares the training time and the RSME value of optimal objective function (parity residuals) among different gradient descent methods.

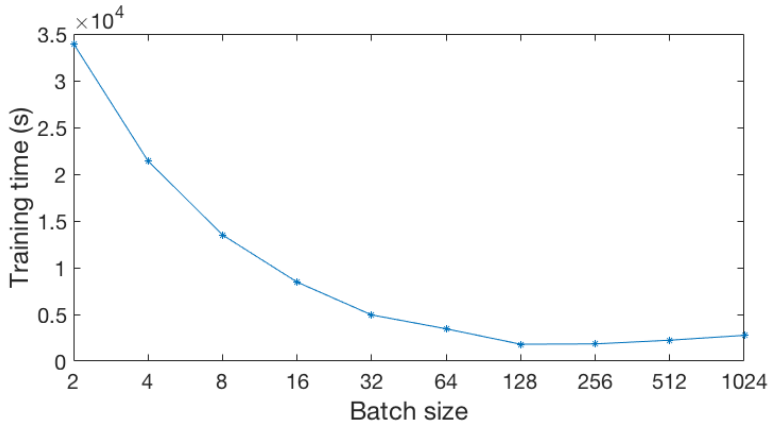


Figure 4-5: Training time with the mini-batch stochastic gradient descent method

Table 4-1: Comparison of three gradient descent algorithms

	Training time (s)	RMSE of parity residuals
Standard gradient descent	2377	0.642
Stochastic gradient descent	48506	0.632
Mini-batch stochastic gradient descent	1813	0.639

Scenario 1: Sensor fault diagnosis with a global model approach

The parity relations are trained with data under normal operating conditions that involve both operating conditions 1 and 2. As shown in Figure 4-6, the testing result shows that the parity residuals generated with the global model approach have large variance and are not consistent. By setting the threshold with 2 sigma, under normal condition, the false alarm is 6.9%, while the detection power with sensor A degraded by 20% is only 27.22%. The comparison result of parity residuals shown in Figure 4-7 fails to show a clear discrepancy between a normal sensor and a degraded sensor. Therefore, the global model approach is not effective to detect sensor faults/degradation in this case study.

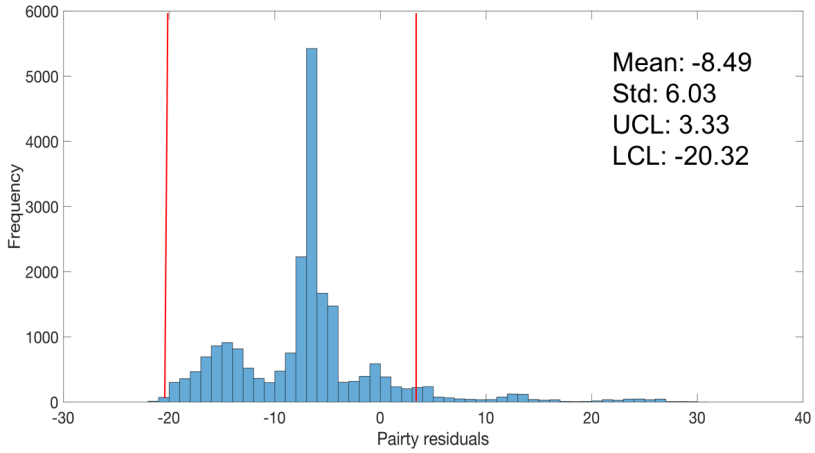


Figure 4-6: Histogram for parity residuals that generated with normal sensors - global model approach

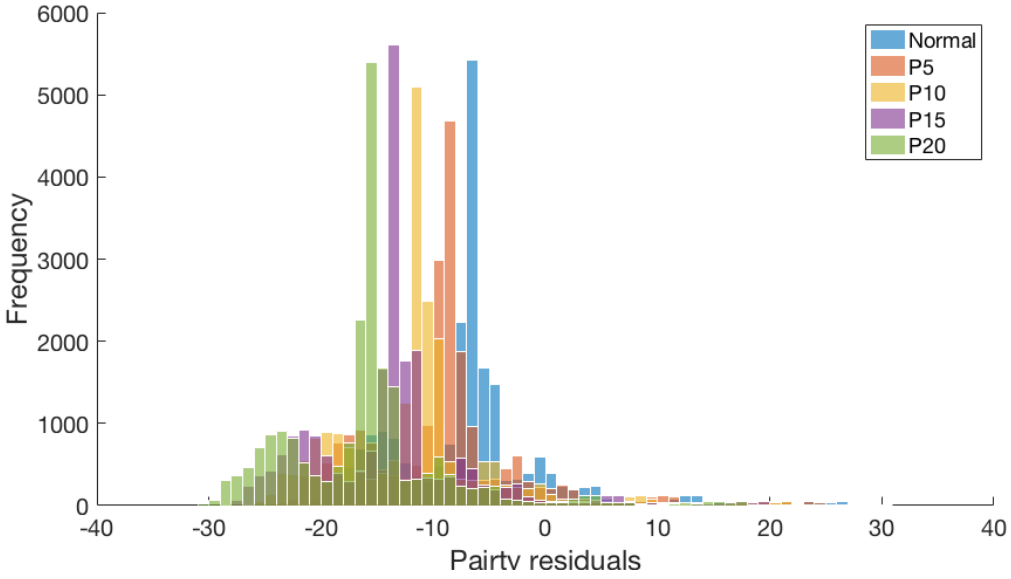
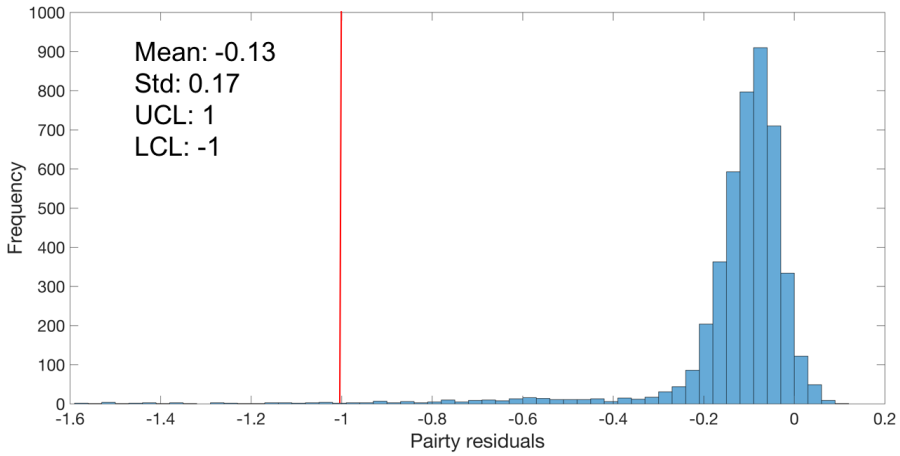


Figure 4-7: Histogram for parity residuals that generated with the sensor A degraded - global model approach

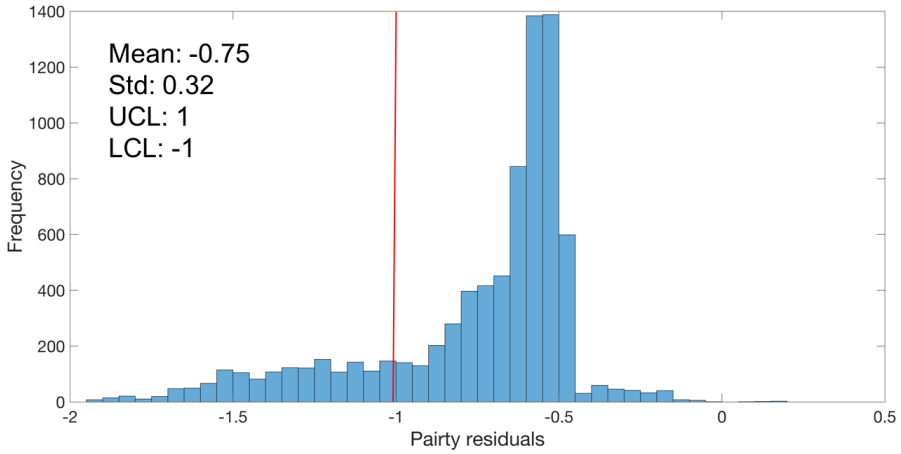
Scenario 2: Sensor fault diagnosis with the multi-regime approach (GSMMS)

In the last scenario, the GSMMS is constructed with the proposed method and data from normal operating conditions 1 and 2. The diagnostic performance of the generated local parity

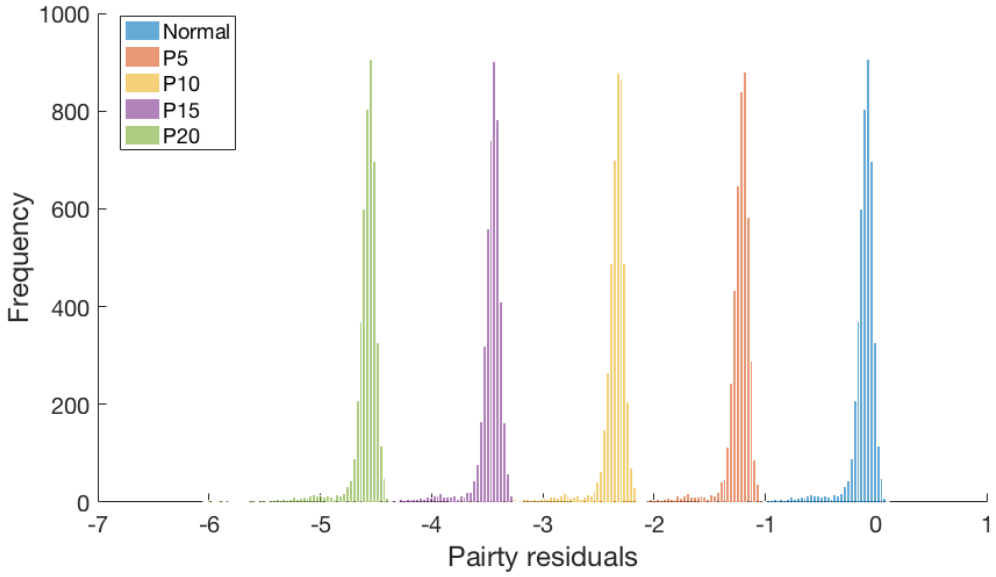
residuals is tested with normal and abnormal (sensor A degraded) data that collected from both operating conditions 1 and 2. Figures 4-8 a) and b) demonstrate that the GSMMS network trained with normal data from operating condition 1 and 2 is able to recognize that there are two different operating conditions, partition the entire operating space into two operating regimes, and develop two local models (M1 – local parity relations in operating regime 1, and M2 – local parity relations in operating regime 2) accordingly. The testing results in Table 4-1 shows that the false alarm rate under normal operating conditions 1 and 2 is 1.84% and 6.43%, respectively. The developed GSMMS network tested with sensor A degraded by 20% demonstrates its capability of detecting the anomaly with detection power of 100% and 98.83% under operating conditions 1 and 2, respectively. Also, as shown in Figures 4-8 c) and d), the parity residuals generated with the GSMMS network can be used as a heath indicator to reveal the severity of the sensor degradation.



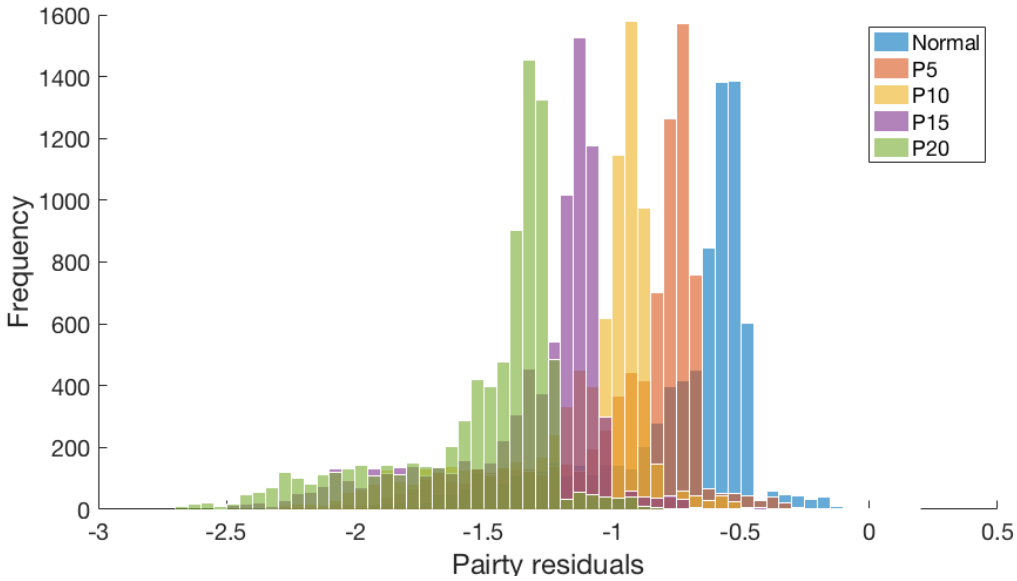
a) Parity residuals with normal sensors under operating condition 1 (N1)



b) Parity residuals with normal sensors under operating condition 2 (N2)



c) Parity residuals with sensor A degraded under operating condition 1 (N1_f1)



d) Parity residuals with sensor A degraded under operating condition 2 (N2_f1)

Figure 4-8: Parity residuals under different operating regimes with the multiple model approach

Table 4-2: Summary of the diagnostic accuracy under different scenarios

Scenarios	N1	N2	N1_f1	N2_f2
1. global model approach	93.10%	100.00%	27.22%	100.00%
2. multiple model approach	98.16%	93.57%	100.00%	98.83%

4.5 Discussion

Table 4-2 summarizes the diagnostic accuracy of the models developed under different scenarios and shows that the overall diagnostic performance of the sensor fault diagnosis is significantly improved with the multiple model approach. Moreover, it is observed that with the multiple model approach, the diagnostic performance with M1 is better than the one with M2 in their corresponding operating regimes. This is because the operating condition 2 involves larger

variance and noise than operating condition 1. As shown in Figure 4-9, the original system model accuracy under those two operating conditions is different, which is mainly affected by sensor measurement noise and system model uncertainty. Such different variations of the uncertainty and noise under the two operating conditions affect the local model development in the GSMMS framework, and therefore lead to different diagnostic performance.

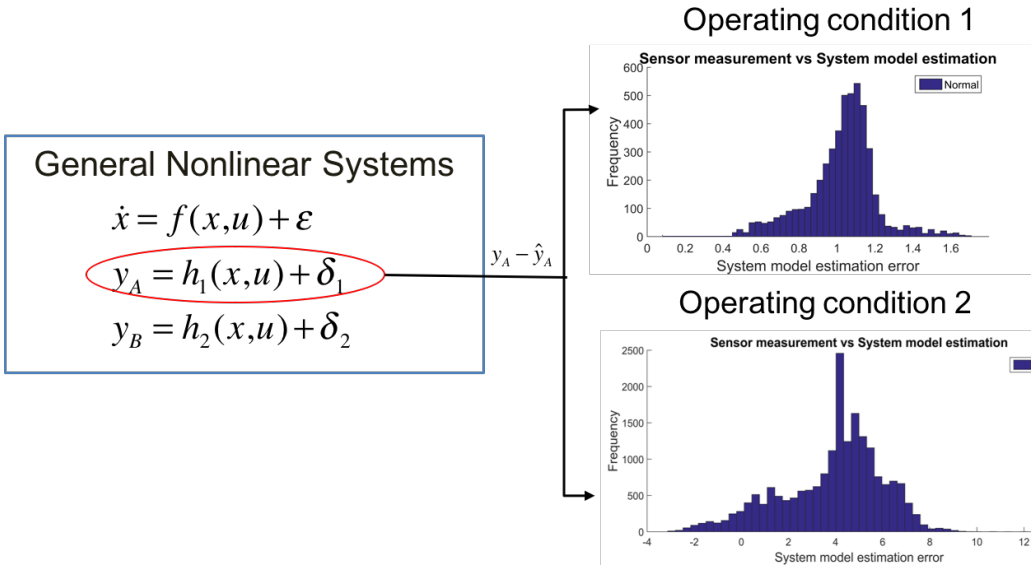


Figure 4-9: System model estimation results under different operating conditions

In Figure 4-10, the parity residuals are generated with M1 (x-axis) and M2 (y-axis) with data from operating conditions 1 (opt) and 2 (op2) with and without sensor A degraded. It is observed that the fault signature from the same failure source varies under different operating regimes. Those signatures can provide additional information for identifying the failure types and might provide hints for the root cause of the failures.

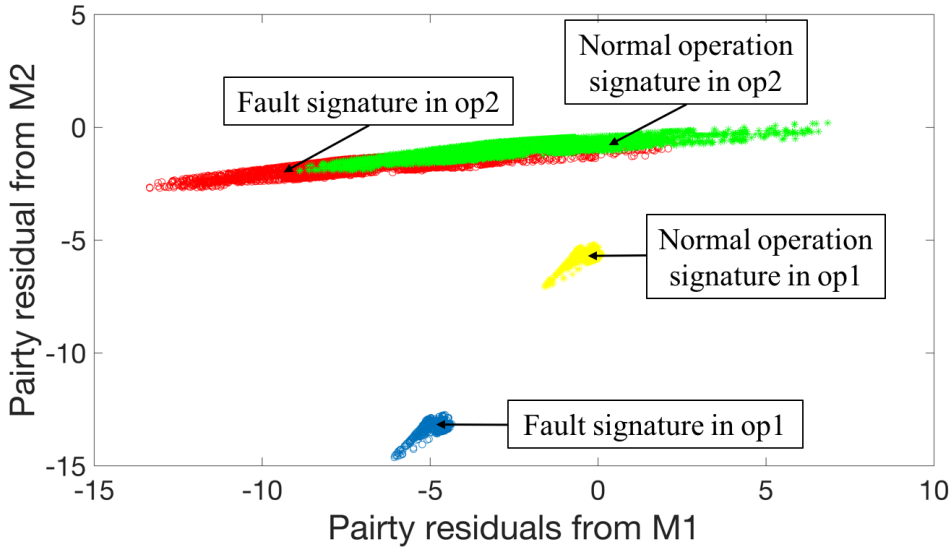


Figure 4-10: Parity residuals signatures under different operating regimes with and without sensor fault

4.6 Conclusion

In this chapter, a multi-regime modeling framework for anomaly detection and fault diagnosis is introduced. Following the strategy of "divide and conquer", the multiple model approach is adopted, which divides the entire system operating range into small operating regimes and then models the local dynamics individually. This research employed and revised the GSMMS to develop local models for anomaly detection and fault diagnosis in complex systems by integrating the growing SOM for partitioning operating regime and the gradient method for identifying local model parameters. The Voronoi sets in the SOM naturally partition the entire operating space into smaller operating regimes, in which local models are developed to describe the system behaviors individually. The number of the operating regimes is automatically determined by the growing SOM, which is capable of detecting and learning new data patterns that are generated from new operating regimes.

For the local model parameter identification problem, instead of using a linear least squares algorithm, this research formulates it as an optimization problem based on the loss minimization framework and finding the optimal solution with the mini-batch stochastic gradient descent method. This modification can help to handle the parameter identification problem of local models that do not have a closed-form solution or are nonlinear.

The case study of sensor fault diagnosis in an R2R registration process demonstrates the performance of the proposed method by comparing the accuracy of the diagnosis results from a global model approach and from the proposed multiple model approach. Both normal data and abnormal data collected from two different operating regimes are used to validate the detection capability of the proposed method.

The results show that the proposed multiple model approach outperforms the traditional global model approach in terms of diagnostic accuracy. The revised GSMMS can handle the local model identification for the parity relations, which does not have a closed-form solution. Moreover, with its continuous learning capability, the proposed method can enhance its detection capability over time with data from new operating regimes and is capable of identifying different system behaviors induced by various normal operating conditions or sensor degradation and failures.

To sum up, the revised GSMMS method for multi-regime anomaly detection and fault diagnosis can effectively detect anomalies and identify faults under different operating regimes. Such multiple model approach eases the modeling tasks for system dynamics by developing local models for each smaller sub-region instead of for the entire system. It also enables input-dependent fault diagnosis in complex systems, which provides additional insight to interpret the residual errors between the model output and the actual systems. The proposed method has a

wide range of industrial applications in anomaly detection and fault diagnosis with its benefit of an efficient online training process and effective diagnostic capability.

CHAPTER 5 CONCLUSION AND FUTURE WORK

5.1 Conclusions

This doctoral research has presented the research attempts in developing practical approaches to increase system visibility without additional sensors, in order to enhance real-time monitoring and fault diagnosis capability in R2R manufacturing systems. In particular, the research focused on 1) a multistage modeling method that characterizes the twofold error propagation and its associated quality measurements in R2R manufacturing systems, 2) a nonlinear analytical redundancy method for sensor fault detection in general nonlinear systems, and 3) a multiple regime anomaly detection and fault diagnosis scheme with a revised version of the GSMMS method that integrates both the growing SOM and gradient descent method. The effectiveness of those proposed methods is demonstrated with data from an R2R web handling process.

Chapter 2 characterizes the twofold error propagation include process-centric and product-centric variation propagation in R2R processes. A hybrid modeling method that integrates both physical models and regression models is proposed to describe the multistage process-centric variation propagation, which uniquely exists in R2R processes, and its associated product quality measurements from one stage to another. This modeling method can help release the requirement of sophisticated knowledge of a system and sensor/inspection systems in every stage. Moreover, the case study of an unwinding process indicates that the hybrid modeling method outperforms

pure data-driven methods with the improved accuracy of tension estimation by 15% and the pitch length by 70%. Finally, the estimation results from the multistage model can serve as virtual sensing and virtual metrology to monitor operation performance and product quality in each stage, thereby increase the visibility of the R2R manufacturing system without additional physical sensors.

Chapter 3 presents a model-based analytical redundancy approach for sensor fault diagnosis by employing nonlinear observation matrix to formulate parity relations for residuals generation. The proposed method extends the analytical approach from the linear system to the general nonlinear system in which, both input and output equations are nonlinear functions of states and inputs. Moreover, it is able to generate residuals that are robust to noise and model uncertainties while sensitive to sensor faults. Finally, the case study of the R2R registration process indicates that the different types of sensor faults/degradations can be detected and isolated by the designed analytical redundancies without additional physical sensors.

Chapter 4 introduces a multiple regime modeling approach for anomaly detection and fault diagnosis framework by integrating the growing self-organizing map for partitioning operating regime and the gradient method for identifying local model parameters. The proposed method enables input-dependent fault diagnosis in complex systems, which facilitates the interpretation of the residual errors between the model output and the actual system. Moreover, the case study of the R2R registration process indicates that the proposed multiple model approach outperforms the traditional global model approach in terms of diagnostic accuracy. Finally, the continuous learning capability of the proposed method can enhance its detection capability over time with data from new operating regimes and is capable of identifying different system behaviors under different operating conditions.

5.2 Contributions of this thesis

This thesis facilitates a full realization of the potential advantages of R2R manufacturing systems and addresses existing barriers in real-time monitoring and fault diagnosis in complex systems. The detailed contributions are summarized as followings:

First, this thesis investigated the variation propagation mechanism and the associated printing quality issues in the continuous R2R manufacturing processes. A twofold variation model is developed to describe how variations are introduced and transformed as the substrate goes from one operation stage to another (**product-centric variation model**), and how variation propagate instantaneously to the downstream substrate (**process-centric variation model).** A multistage model based on the formulation of SoV is developed via a novel modeling approach - a *hybrid modeling* that integrates physics-based models (torque equilibrium, and Hooke's law) with data-driven methods (e.g., censored regression, and linear/logistic regression) to address the complex variation propagation phenomenon. Specifically, sensor data analytics complement the lack of full physical knowledge of an R2R process dynamics, while the physical knowledge minimizes the requirement for sensing and inspection at each stage. The estimates of the state variables (e.g., web tension) serve as a virtual sensor, while the outputs of the observation equations serve as a virtual metrology tool for intermediate product quality measurements based on system inputs (e.g., material properties and operational variables). As a result, the model serves as a foundation for process diagnosis/prognosis, quality control and improvement in R2R manufacturing systems. To the best of our knowledge, this is the first work that uses the SoV model to characterize the process-centric variation propagation and product variability in a continuous manufacturing system, whose mechanism is vastly different from the one in discrete manufacturing systems.

Second, this thesis proposed a nonlinear analytical redundancy method for sensor fault diagnosis problem in general nonlinear systems in which, both input and output equations are nonlinear functions of states and inputs. Following the idea of the linear analytical redundancy method – parity space that utilizes observation matrix to construct input-output relations to describe the relationships between system behaviors and sensor measurements, this study extends the parity space from linear to nonlinear systems by decomposing nonlinear observation matrix to build the input-output relations. The number of available analytical redundancies that can be added for sensor fault diagnosis is determined by the rank of nonlinear observation matrix and the generated residuals from each redundancy are used to detect and identify sensor faults in general nonlinear systems. Moreover, a robust optimization is designed to identify the model coefficients so that the generated residuals are sensitive to sensor failures but robust to noise/uncertainties. Post-process sensitivity analysis is also conducted to evaluate the effect of the changing operating conditions on the parity residuals generation, and provides the valid operating range given optimal model coefficients.

Finally, this thesis introduced a multi-regime modeling framework for anomaly detection and fault diagnosis. Following the idea of "divide and conquer", the multiple model approach is adopted, which divides the entire system operation space into small sub-regions and then models the local dynamics individually. The GSMMS is employed and revised to develop multiple models for anomaly detection and fault diagnosis in complex systems by integrating the growing self-organizing map for partitioning operating regime and the gradient method for identifying local model parameters. The novelty of this study is that instead of using the linear least squares algorithm, this research formulates the local model parameter identification problem as an optimization problem based on the loss minimization framework and finding the optimal solution

with the mini-batch stochastic gradient descent method. This modification can handle the parameter identification problem of local models that does not have a closed-form solution or has a nonlinear structure. The residuals generated from each operating regime enables input-dependent fault diagnosis in complex systems, which can provide extra information for fault identification. Moreover, the continuous learning capability of the proposed method can enhance its detection capability over time with data from new operating regimes and is capable of identifying different system behaviors induced under various normal operating regimes.

5.3 Future work

Some future research work can be conducted in the area addressed by this thesis. A short summary of possible directions is listed below.

1) The development of the multistage model is applied to R2R processes with considering the process-centric variation as tension propagation in Chapter 2. However, chemical reactions and heat treatment sometimes are also integrated in R2R processes, which might bring additional types of process-centric variation and will be an interesting direction for further investigation.
2) Sensors and inspection systems are valuable for quality assurance in R2R processes, which will detect defects and trace product quality at each individual operation step. Therefore, insufficient sensors might delay the realization of defects and increase the time of finding the root cause. However, it is not applicable nor cost-effective to install sensors and inspections system for every operation step due to constraints of budget, space and process speed. Therefore, an optimal sensor placement in an R2R process that balances the tradeoff between the cost of sensors (installation and maintenance) and the cost of defects (capital and time) can be further studied.

3) The nonlinear analytical redundancy method proposed in Chapter 3 for sensor fault diagnosis assumes the actuator works properly. However, it is not a rare situation that both sensors and actuators fail in a nonlinear system. A robust fault diagnosis scheme that is capable of detecting and isolating both sensor and actuator failures for a general nonlinear system is worthy of further explorations.
4) Due to the increasing complexity of manufacturing systems, and large volume of data under different operating conditions, the optimization problem that is designed for the local model identification in Chapter 4 often requires high computation power during the training process and online adaptive learning. Larger batch size for the mini-batch stochastic gradient descent method can reduce communication time but will slow down the convergence rate, which will diminish the benefits of the reduced communication cost. Guidelines for selecting a proper batch size for different applications can be explored. Moreover, parallelized and distributed algorithms for learning the local model can be further investigated to improve the efficiency of the online learning process.

APPENDICES

A.I: Dynamic equations in R2R processes

According to the *torque equilibrium analysis*, the dynamic equations of the material rolls, driven rollers, and idle rollers can be written in the following general forms (Branca et al., 2012).

$$\frac{d}{dt}\left(J_k(t)\omega_k(t)\right) = -\tau_{fk} + R_k(t)\left(T_{k+1}(t) - T_k(t)\right) + \tau_k(t) \tag{A.1}$$

where $J_k(t)$ is the inertia of the roll, $\omega_k(t)$ is the angular velocity of the roller, $\tau_{fk}(t)$ denotes the friction torque, $R_k(t)$ denotes the radius of the roller, $T_{k+1}(t)$ are the tension of the substrate coming out of the roller, $\tau_k(t)$ is the applied torque transmitted from motor to the roller.

Material roll: The inertia $J_k(t)$ and the radius $R_k(t)$ are both a function of time (decreasing as material unwinds while increasing as material rewinds) that can be expressed as:

$$J_k(t) = J_0 + \frac{\pi}{2}\rho_w w_w (R_k^4(t) - R_c^4) \tag{A.2}$$

$$\dot{R}_k(t) = \lim_{\Delta t \to 0} \frac{\Delta R_k}{\Delta t} = \frac{t_w \omega_k(t)}{2\pi} \tag{A.3}$$

where J_0 is a fixed inertia of the core shaft and the motor. ρ_w is the density of the substrate, w_w is the width of the substrate, R_c is radius the of the core shaft, t_w is the material thickness.

Driven roller: $J_k(t)$ is time-invariant and the dynamics is represented by Eq. (A.4).

$$\frac{d}{dt}\left(J_k\omega_k(t)\right) = -\tau_{fk} + R_k\left(T_{k+1}(t) - T_k(t)\right) + \tau_k(t) \tag{A.4}$$

Idle roller: There is no $\tau_k(t)$ and $\mathrm{J_k(t)}$ is time-invariant so that the dynamic equation is shown

as Eq. (A.5)

$$\frac{d}{dt}(J_k\omega_k(t)) = -\tau_{fk} + R_k(T_{k+1}(t) - T_k(t)) \quad \text{(A.5)}$$

Undesired slippage: The occurrence of slippage between a roller and the web is caused by excessive air-entrainment between them. During slippage, the velocity of the roller falls below the web velocity and the roller cannot guide the web anymore. Therefore, the web will tend to shift in transverse direction by external disturbances, and wrinkles or scratching may be generated on the web. Avoiding such slippage events can not only reduce defective products but also improve productivity by maintaining the web transportation at the designed speed. With the intermediate tension estimated from the proposed multistage model, the belt friction equation can be employed to monitor the occurrence of the slippage in an R2R manufacturing system (Whitworth and Harrison, 1983).

$$\textbf{Belt friction equation}: T_k \leq T_{k-1}exp(\mu_k\beta_k) \quad \text{(A.6)}$$

where μ_k is the frictional coefficient between a roller and web, β_k is the wrap angle of a roller.

A.II: Derivatives of parameter estimation in the censored regression model

$$T_k^{id} = \begin{cases} T_k^{id^*} & if\ T_k^{id^*} > L \\ 0 & if\ T_k^{id^*} \le L \end{cases} \tag{A.7}$$

where T_k^{id} is a real measurement (longitudinal tension) and $T_k^{id^*}(t) = E_k T_{k-1}^{id-1}(t) + \boldsymbol{F}_k \boldsymbol{R}_k(t) + Z_k$ is a latent variable, $L = 0$. Parameters E_k and F_k can be obtained by maximum likelihood estimation.

Case 1: $T_k^{id} > 0$

When $T_k^{id} > 0$, $T_k^{id} = T_k^{id^*}$ so that the conditional probability of T_k^{id} is same as that of $T_k^{id^*}$. Therefore, the probability of the tension on the current pitch conditioning on the upstream pitch tension and the roll radius is the following:

$$f\left(T_k^{id} \middle| T_k^{id-1}, \boldsymbol{R}_k\right) = f\left(T_k^{id^*} \middle| T_k^{id-1}, \boldsymbol{R}_k\right) = \frac{1}{\sigma_z\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{T_k^{id} - \left(E_k T_{k-1}^{id-1} + \boldsymbol{F}_k \boldsymbol{R}_k\right)}{\sigma_z}\right)^2\right\} =$$

$$\frac{1}{\sigma_z}\phi\left(\frac{T_k^{id} - \left(E_k T_{k-1}^{id-1} + \boldsymbol{F}_k \boldsymbol{R}_k\right)}{\sigma_z}\right) \tag{A.8}$$

where $T_k^{id^*} \sim N(E_k T_{k-1}^{id-1} + \boldsymbol{F}_k \boldsymbol{R}_k, {\sigma_z}^2)$ and ϕ is the probability density function (pdf) of a standard normal distribution. (σ_z is the standard deviation of the normal distribution)

Case 2: $T_k^{id} = 0$

When $T_k^{id} = 0$, the mass conditional probability can be obtained by:

$$Pr\left(T_k^{id} = 0 \middle| T_k^{id-1}, \boldsymbol{R}_k\right) = Pr\left(T_k^{id^*} < 0 \middle| T_k^{id-1}, \boldsymbol{R}_k\right)$$

$$= Pr\left(E_k T_{k-1}^{id-1} + \boldsymbol{F}_k \boldsymbol{R}_k + Z_k < 0 \middle| T_k^{id-1}, \boldsymbol{R}_k\right) = Pr\left(Z_k \le -(E_k T_{k-1}^{id-1} + \boldsymbol{F}_k \boldsymbol{R}_k) \middle| T_k^{id-1}, \boldsymbol{R}_k\right)$$

Since $Z_k \sim N(0, {\sigma_z}^2)$, we have

$$Pr\left(T_k^{id}=0\middle|T_k^{id-1},\boldsymbol{R}_k\right)=\Phi\left(-\frac{\left(E_kT_{k-1}^{id-1}(t)+\boldsymbol{F}_k\boldsymbol{R}_k(t)\right)}{\sigma_z}\right)=1-\Phi\left(\frac{\left(E_kT_{k-1}^{id-1}(t)+\boldsymbol{F}_k\boldsymbol{R}_k(t)\right)}{\sigma_z}\right) \tag{A.9}$$

where Φ is cumulative distribution function (cdf) of standard normal distribution.

By adding a dummy variable d_j, where

$$d_j=\begin{cases}1 & if\ T_k^{id}>0\\ 0 & if\ T_k^{id}\le 0\end{cases} \tag{A.10}$$

And the conditional pdf of T_k^{id} given T_k^{id-1} and $\boldsymbol{R}_k$ can be expressed as:

$$f\left(T_k^{id}\middle|T_k^{id-1},\boldsymbol{R}_k\right)=\left\{\frac{1}{\sigma_z}\phi\left(\frac{T_k^{id}-\left(E_kT_{k-1}^{id-1}+\boldsymbol{F}_k\boldsymbol{R}_k\right)}{\sigma_z}\right)\right\}^{d_j}\cdot\left\{1-\Phi\left(\frac{\left(E_kT_{k-1}^{id-1}(t)+\boldsymbol{F}_k\boldsymbol{R}_k(t)\right)}{\sigma_z}\right)\right\}^{1-d_j} \tag{A.11}$$

Then, the likelihood function L_M can be defined as:

$$L_M=\prod_{m=1}^{M}\left\{\frac{1}{\sigma_z}\phi\left(\frac{T_k^{id}-\left(E_kT_{k-1}^{id-1}+\boldsymbol{F}_k\boldsymbol{R}_k\right)}{\sigma_z}\right)\right\}^{d_j}\cdot\left\{1-\Phi\left(\frac{\left(E_kT_{k-1}^{id-1}(t)+\boldsymbol{F}_k\boldsymbol{R}_k(t)\right)}{\sigma_z}\right)\right\}^{1-d_j} \tag{A.12}$$

Therefore, the maximum likelihood estimator for the parameters E_k and $\boldsymbol{F}_k$ can be defined as:

$$\left(\hat{E}_k,\hat{\boldsymbol{F}}_k;\widehat{\sigma_z}\right)=\underset{E_k,\boldsymbol{F}_k;\sigma_z}{arg\ max}\{ln\,L_M(E_k,\boldsymbol{F}_k;\sigma_z)\}=$$

$$\underset{E_k,\boldsymbol{F}_k;\sigma_z}{arg\ max}\left\{\sum_{m=1}^{M}d_j\,ln\left[\frac{1}{\sigma_z}\phi\left(\frac{T_k^{id}-\left(E_kT_{k-1}^{id-1}+\boldsymbol{F}_k\boldsymbol{R}_k\right)}{\sigma_z}\right)\right]+\left(1-d_j\right)ln\left[1-\right.\right.$$

$$\left.\left.\Phi\left(\frac{\left(E_kT_{k-1}^{id-1}(t)+\boldsymbol{F}_k\boldsymbol{R}_k(t)\right)}{\sigma_z}\right)\right]\right\} \tag{A.13}$$

where M denotes the total number of training data sets ($m = 1,\ldots, M$)

REFERENCE

Abellán-Nebot, JV, Romero Subirón, F, Serrano Mira, J. 2013. Manufacturing variation models in multi-station machining systems. Int. J. Adv. Manuf. Technol. 64: 63–83.

Armacost, RL, Fiacco, A V. 1974. Computational experience in sensitivity analysis for nonlinear programming. Math. Program. 6: 301–326.

Babuška, R. 1998. Fuzzy Modeling for Control. Dordrecht: Springer Netherlands.

Barton, PI, Pantelides, CC. 1994. Modeling of combined discrete/continuous processes. AIChE J. 40: 966–979.

Bencze, WJ, Franklin, GF. 1994. A separation principle for hybrid control system design. In: Proceedings of IEEE Symposium on Computer-Aided Control Systems Design (CACSD). IEEE, p 327–332.

Bengio, Y. 2012. Practical Recommendations for Gradient-Based Training of Deep Architectures. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)., p 437–478.

Betta, G, D'Apuzzo, M, Pietrosanta, A. 1995. A knowledge-based approach to instrument fault detection and isolation. IEEE Trans. Instrum. Meas. 44: 1009–1016.

Bishop, CM. 2006. Summary for Policymakers. In: Intergovernmental Panel on Climate Change, editor. Climate Change 2013 - The Physical Science Basis. Cambridge: Cambridge University Press, p 1–30.

Bleakie, A, Djurdjanovic, D. 2016. Growing Structure Multiple Model System for Quality Estimation in Manufacturing Processes. IEEE Trans. Semicond. Manuf. 29: 79–97.

Blom, HAP, Bar-Shalom, Y. 1988. The interacting multiple model algorithm for systems with Markovian switching coefficients. IEEE Trans. Automat. Contr. 33: 780–783.

Branca, C, Pagilla, PR, Reid, KN. 2012. Governing Equations for Web Tension and Web Velocity in the Presence of Nonideal Rollers. J. Dyn. Syst. Meas. Control 135: 11018.

Broen, R. 1974. A nonlinear voter-estimator for redundant systems. In: 1974 IEEE Conference on Decision and Control including the 13th Symposium on Adaptive Processes. IEEE, p 743–748.

Bureau d'Enquêtes et d'Analyses pour la sécurité d. 2012. Final Report. Golden, CO.

Burnham, KP, Anderson, DR. 2002. Model selection and multiple inference: a practical information-theoretic approach, second. New York, USA: Springer. 60-64 p.

Camelio, J, Hu, SJ, Ceglarek, D. 2003. Modeling Variation Propagation of Multi-Station Assembly Systems With Compliant Parts. J. Mech. Des. 125: 673.

Chen, S, Hong, X, Harris, CJ, Wang, XX. 2005. Identification of nonlinear systems using generalized kernel models. IEEE Trans. Control Syst. Technol. 13: 401–411.

Cholette, ME, Djurdjanovic, D. 2012. Precedent-Free Fault Isolation in a Diesel Engine Exhaust Gas Recirculation System. J. Dyn. Syst. Meas. Control 134: 31007.

Chow, E, Willsky, A. 1984. Analytical redundancy and the design of robust failure detection systems. IEEE Trans. Automat. Contr. 29: 603–614.

Clark, RN. 1978. Instrument Fault Detection. IEEE Trans. Aerosp. Electron. Syst. AES-14: 456–465.

Cormen, TH, Leiserson, CE, Rivest, RL, Stein, C. 2009. Introduction to algorithms, Third edit. London, England: The MIT press. 594-601 p.

Das, R, Harrop, P. 2011. Printed, Organic & Flexible Electronics Forecasts, Players & Opportunities 2011-2021.

Department of Energy. 2015. Roll to Roll Processing Technology assessments.

Deshpande, AP, Patwardhan, SC. 2008. Online Fault Diagnosis in Nonlinear Systems Using the Multiple Operating Regime Approach. Ind. Eng. Chem. Res. 47: 6711–6726.

Djurdjanovic, D, Ni, J. 2003. Dimensional Errors of Fixtures, Locating and Measurement Datum Features in the Stream of Variation Modeling in Machining. J. Manuf. Sci. Eng. 125: 716–730.

Djurdjanovic, D, Ni, J. 2004. Measurement Scheme Synthesis in Multi-Station Machining Systems. J. Manuf. Sci. Eng. 126: 178.

Dong, J, Wang, M, Zhang, X, Ma, L, Peng, K. 2017. Joint Data-Driven Fault Diagnosis Integrating Causality Graph With Statistical Process Monitoring for Complex Industrial Processes. IEEE Access 5: 25217–25225.

Du, M, Mhaskar, P. 2014. Isolation and handling of sensor faults in nonlinear systems. Automatica 50: 1066–1074.

Dunia, R, Qin, SJ, Edgar, TF, McAvoy, TJ. 1996. Identification of faulty sensors using principal component analysis. AIChE J. 42: 2797–2812.

Elnokity, O, Mahmoud, II, Refai, MK, Farahat, HM. 2012. ANN based Sensor Faults Detection, Isolation, and Reading Estimates – SFDIRE: Applied in a nuclear process. Ann. Nucl. Energy 49: 131–142.

Fiacco, A V. 1976. Sensitivity analysis for nonlinear programming using penalty methods. Math. Program. 10: 287–311.

Frechard, J, Knittel, D, Dessagne, P, Pellé, JS, Gaudiot, G, Caspar, JC, Heitz, G. 2013. Modelling and fast position control of a new unwinding–winding mechanism design. Math. Comput. Simul. 90: 116–131.

Fritzke, B. 1994a. A Growing Neural Gas Learns Topologies. In: Tesauro, G, Touretzky, DS, Leen, TK, editors. NIPS'94 Proceedings of the 7th International Conference on Neural Information Processing Systems. Denver, Colorado: MIT Press Cambridge, MA, USA ©1994, p 625–632.

Fritzke, B. 1994b. Growing cell structures—A self-organizing network for unsupervised and supervised learning. Neural Networks 7: 1441–1460.

Gao, Z, Cecati, C, Ding, S. 2015a. A Survey of Fault Diagnosis and Fault-Tolerant Techniques Part II: Fault Diagnosis with Knowledge-Based and Hybrid/Active Approaches. IEEE Trans. Ind. Electron. 62: 1–1.

Gao, Z, Cecati, C, Ding, SX. 2015b. A Survey of Fault Diagnosis and Fault-Tolerant Techniques—Part I: Fault Diagnosis With Model-Based and Signal-Based Approaches. IEEE Trans. Ind. Electron. 62: 3757–3767.

Gill, BS. 2011. Development of Virtual Metrology in Semiconductor Manufacturing.

Hofleitner, A. 2013. A hybrid approach of physical laws and data-driven modeling for estimation : the example of queuing networks.

Huang, Q, Shi, J. 2004. Variation transmission analysis and diagnosis of multi-operational machining processes. IIE Trans. 36: 807–815.

Huang, Q, Zhou, S, Shi, J. 2002. Diagnosis of multi-operational machining processes through variation propagation analysis. Robot. Comput. Integr. Manuf. 18: 233–239.

Huang, W, Lin, J, Bezdecny, M, Kong, Z, Ceglarek, D. 2007. Stream-of-Variation Modeling—Part I: A Generic Three-Dimensional Variation Model for Rigid-Body Assembly in Single Station Assembly Processes. J. Manuf. Sci. Eng. 129: 821–831.

Huang, Y, Gertler, J, McAvoy, TJ. 2000. Sensor and actuator fault isolation by structured partial PCA with nonlinear extensions. J. Process Control 10: 459–469.

Hwang, ES, Kwon, S, Kim, D, Cho, YT, Jung, YG. 2015. Positional accuracy of micropatterns in the roll-to-roll imprinting process using a wrapped roll mold. J. Mech. Sci. Technol. 29: 1697–1702.

Isermann, R. 1984. Process fault detection based on modeling and estimation methods—A survey. Automatica 20: 387–404.

Jiang, L. 2011. Sensor Fault Detection and Isolation Using System Dynamics Identification Techniques.

Jin, J, Shi, J. 1999. State Space Modeling of Sheet Metal Assembly for Dimensional Control. J. Manuf. Sci. Eng. 121: 756.

JOHANSEN, TA, FOSS, B. 1993. Constructing NARMAX models using ARMAX models. Int. J. Control 58: 1125–1153.

Johansen, TA, Foss, BA. 1995. Identification of non-linear system structure and parameters using regime decomposition. Automatica 31: 321–326.

Johansen, TA, Foss, BA. 1997. Operating regime based process modeling and identification. Comput. Chem. Eng. 21: 159–176.

Jordan, MI, Jacobs, RA. 1994. Hierarchical Mixtures of Experts and the EM Algorithm. Neural Comput. 6: 181–214.

Kano, M, Fujiwara, K. 2013. Virtual Sensing Technology in Process Industries: Trends and Challenges Revealed by Recent Industrial Applications. J. Chem. Eng. JAPAN 46: 1–17.

Kasai, H. 2017. SGDLibrary: A MATLAB library for stochastic gradient descent algorithms. 1–25.

Kim, Y. 1997. A framework for an on-line diagnostic expert system with intelligent sensor validation. KSME Int. J. 11: 10–19.

Lee, CW, Kang, HK, Park, CJ, Shin, KH. 2005. Fault diagnosis of roll shape under the speed change in hot rolling mill. IFAC Proc. Vol. 38: 45–50.

Leuschen, ML, Walker, ID, Cavallaro, JR. 2005. Fault residual generation via nonlinear analytical redundancy. IEEE Trans. Control Syst. Technol. 13: 452–458.

Li, H, Yu, D, Braun, JE. 2011. A review of virtual sensing technology and application in building systems. HVAC&R Res. 17: 619–645.

Li, J, Freiheit, T, Jack Hu, S, Koren, Y. 2007. A Quality Prediction Framework for Multistage Machining Processes Driven by an Engineering Model and Variation Propagation Model. J. Manuf. Sci. Eng. 129: 1088–1100.

Li, J, Shi, J. 2007. Knowledge discovery from observational data for process control using causal Bayesian networks. IIE Trans. 39: 681–690.

Li, W, Shah, S. 2002. Structured residual vector-based approach to sensor fault detection and isolation. J. Process Control 12: 429–443.

Liu, J, Djurdjanovic, D, Marko, K, Ni, J. 2009. Growing Structure Multiple Model Systems for Anomaly Detection and Fault Diagnosis. J. Dyn. Syst. Meas. Control 131: 051001–4:13.

Liu, J, Shi, J, Hu, SJ. 2008. Engineering-Driven Factor Analysis for Variation Source Identification in Multistage Manufacturing Processes. J. Manuf. Sci. Eng. 130: 041009–1:10.

Livesley, RK. 1971. Optimization methods for engineering design. Addison Wesley.

Martinetz, TM, Berkovich, SG, Schulten, KJ. 1993. "Neural-gas" network for vector quantization and its application to time-series prediction. IEEE Trans. Neural Networks 4: 558–569.

Mathioudakis, K, Romessis, C. 2004. Probabilistic neural networks for validation of on-board jet engine data. Proc. Inst. Mech. Eng. Part G J. Aerosp. Eng. 218: 59–72.

Negiz, A, Cinar, A. 1992. On the detection of multiple sensor abnormalities in multivariate processes. In: American Control Conference, 1992. Chicago, IL, USA: IEEE, p 2364–2368.

Nelles, O. 2001. Nonlinear System Identification. Berlin, Heidelberg: Springer Berlin Heidelberg.

Nguang, SK, Zhang, P, Ding, S. 2007. Parity Relation Based Fault Estimation for Nonlinear Systems: An LMI Approach1. In: Fault Detection, Supervision and Safety of Technical Processes 2006. Elsevier, p 366–371.

Orjuela, R, Marx, B, Ragot, J, Maquin, D. 2013. Nonlinear system identification using heterogeneous multiple models. Int. J. Appl. Math. Comput. Sci. 23.

Patton, R, Clark, R, Frank, PM. 1989. Fault diagnosis in dynamic systems: theory and application. New York, USA: Prentice Hall.

Petridis, V, Kehagias, A. 1996. Modular neural networks for MAP classification of time series

and the partition algorithm. IEEE Trans. Neural Networks 7: 73–86.

Qiang Liu, Tianyou Chai, Hong Wang, Si-Zhao Joe Qin. 2011. Data-Based Hybrid Tension Estimation and Fault Diagnosis of Cold Rolling Continuous Annealing Processes. IEEE Trans. Neural Networks 22: 2284–2295.

Qin, SJ, Li, W. 2001. Detection and identification of faulty sensors in dynamic processes. AIChE J. 47: 1581–1593.

Qin, SJ, Li, W. 1999. Detection and identification of faulty sensors with maximized sensitivity. In: Proceedings of the 1999 American Control Conference. IEEE, p 613–617.

Ruder, S. 2016. An overview of gradient descent optimization algorithms. Biometrics 55: 591–596.

Sanger, TD. 1991. A tree-structured adaptive network for function approximation in high-dimensional spaces. IEEE Trans. Neural Networks 2: 285–293.

Shelton, JJ. 1986. Dynamics of Web Tension Control with Velocity or Torque Control. In: Proceedings of American Control Conference. Seattle, WA, USA: IEEE, p 1423–1427.

Sherry, L, Mauro, R. 2014. Controlled Flight into Stall (CFIS): Functional complexity failures and automation surprises. In: 2014 Integrated Communications, Navigation and Surveillance Conference (ICNS) Conference Proceedings. Herndon, VA, USA: IEEE, p D1-1-D1-11.

Shi, J, Zhou, S. 2009. Quality control and improvement for multistage systems: A survey. IIE Trans. 41: 744–753.

Shui, H, Jin, X, Ni, J. 2018. Twofold Variation Propagation Modeling and Analysis for Roll- to-Roll Manufacturing Systems. IEEE Trans. Autom. Sci. Eng.

Shumsky, A. 2008. Robust Analytical Redundancy Relations for Fault Diagnosis In Nonlinear Systems. Asian J. Control 4: 159–170.

Sobieszczanski-sobieski, J, Rileyj, KM. 1982. Sensitivity of Optimum Solutions of Problem Parameters. AIAA J. 20: 1291–1299.

Subbaraman, H, Lin, X, Xu, X, Dodabalapur, A, Guo, LJ, Chen, RT. 2012. Metrology and instrumentation challenges with high-rate, roll-to-roll manufacturing of flexible electronic systems. In: Postek, MT, Coleman, VA, Orji, NG, editors. Instrumentation, Metrology, and Standards for Nanomanufacturing, Optics, and Semiconductors VI., p 846603.

Suykens, JAK, Vandewalle, JPL, De Moor, BLR. 1996. Artificial Neural Networks for Modelling and Control of Non-Linear Systems. Boston, MA: Springer US.

Takagi, T, Sugeno, M. 1985. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man. Cybern. SMC-15: 116–132.

Tesheng Hsiao, Tomizuka, M. 2005. Sensor fault detection in vehicle lateral control systems via switching kalman filtering. In: Proceedings of the 2005, American Control Conference, 2005. IEEE, p 5009–5014.

Tobin, J. 1958. Estimation of Relationships for Limited Dependent Variables. Econometrica 26: 24.

Tzafestas, SG, Zikidis, KC. 2001. NeuroFAST: on-line neuro-fuzzy ART-based structure and parameter learning TSK model. IEEE Trans. Syst. Man Cybern. Part B 31: 797–802.

Ulsh, M. 2014. EERE Quality Control Workshop Final Report EERE Quality Control Workshop Final Report.

Uosaki, K, Hatanaka, T. 2008. Hybrid Operating Regime Selection Algorithm in Local Modeling. IFAC Proc. Vol. 41: 13504–13508.

Upadhyaya, BR, Kerlin, TW. 1987. Estimation of response time characteristics of platinum resistance thermometers by the noise analysis technique. ISA Trans. 17: 21–38.

US Dept of Energy. 2013. Roll to Roll (R2R) Processing Technology Assessment. 1–36.

VANDERPLAATS, GN, YOSHIDA, N. 1985. Efficient calculation of optimum design sensitivity. AIAA J. 23: 1798–1803.

Wang, T, Jianbo Yu, Siegel, D, Lee, J. 2008. A similarity-based prognostics approach for Remaining Useful Life estimation of engineered systems. In: 2008 International Conference on Prognostics and Health Management. IEEE, p 1–6.

Whitworth, DPD, Harrison, MC. 1983. Tension variations in pliable material in production machinery. Appl. Math. Model. 7: 189–196.

Xiao, D, Jiang, J, Mao, Y, Liu, X. 2016. Process Monitoring and Fault Diagnosis for Piercing Production of Seamless Tube. Math. Probl. Eng. 2016: 1–13.

Xie, K, Wells, L, Camelio, J a., Youn, BD. 2007. Variation Propagation Analysis on Compliant Assemblies Considering Contact Interaction. J. Manuf. Sci. Eng. 129: 934–942.

Young, GE, Reid, KN. 1993. Lateral and Longitudinal Dynamic Behavior and Control of Moving Webs. J. Dyn. Syst. Meas. Control 115: 309–317.

Yu, DL, Shields, DN. 2001. Extension of the parity-space method to fault diagnosis of bilinear systems. Int. J. Syst. Sci. 32: 953–962.

Zeng, L, Zhou, S. 2007. Inferring the Interactions in Complex Manufacturing Processes Using Graphical Models. Technometrics 49: 373–381.

Zhang, L, Ni, J, Lai, X. 2008. Dimensional errors of rollers in the stream of variation modeling in cold roll forming process of quadrate steel tube. Int. J. Adv. Manuf. Technol. 37: 1082–1092.

Zhou, S, Huang, Q, Shi, J. 2003. State space modeling of dimensional variation propagation in multistage machining process using differential motion vectors. IEEE Trans. Robot. Autom. 19: 296–309.