University of Alberta

BAYESIAN SOLUTIONS TO MULTI-MODEL INFERENTIAL SENSING PROBLEMS

by

Shima Khatibisepehr

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Process Control

Department of Chemical and Materials Engineering

Fall 2013
Edmonton, Alberta

To *peace*, *knowledge*, and *dignity*,
as humankind journeys into an uncertain era of
dynamic changes and constrained resources

Abstract

In many industrial plants, development and implementation of advanced monitoring and control techniques require real-time measurement of process quality variables. However, on-line acquisition of such data may involve difficulties due to inadequacy of measurement techniques or low reliability of measuring devices. These concerns motivate the design of inferential sensors to infer process quality indicators from real-time measurable process variables. Development and implementation of inferential sensors entail many challenges that are often addressed in a rather *ad hoc* manner. Although many of the encountered challenging issues are interconnected, most of the existing solutions are disjoint. The main contribution of this dissertation is development of an integrative and holistic Bayesian inferencing paradigm to provide general and integrated solutions to certain outstanding inferential sensing problems.

The core component of an inferential sensor is the process model which is identified through first-principles and process data analysis. The problem of model identification from contaminated data is formulated under a hierarchical Bayesian framework to simultaneously consider different aspects of data analysis and inferential modeling.

A Bayesian approach is developed for identification of multi-modal systems switching among non-linear continuous-state dynamics. The proposed procedure provides a framework to accommodate the overlapping operating regions, facilitate the inclusion of prior knowledge about the operating conditions, and include a global adaptation mechanism within the envelope of previously identified operating conditions.

Real-time identification of inferential models can be viewed as a special modeling technique for design of multi-model inferential sensors with infinite number of local models. A Bayesian framework is developed to provide a systematic and computationally feasible method for real-time similarity function parametrization and model structure selection in just-in-time/space modeling methods.

One of the practical challenges faced in implementation of inferential sensors is to assess the accuracy of their real-time predictions. A data-driven Bayesian approach is proposed to capture conditional dependence of the reliability of inferential sensor predictions on characteristics of the input space and reliability of the empirical process model.

The practicality and validity of the proposed Bayesian frameworks are verified using data from various simulation configurations, experimental set-ups, and industrial processes.

Preface

”My grandfather once told me that there are two kinds of people: those who work and those who take the credit. He told me to try to be in the first group; there was less competition there.” *Indira Gandhi*

Acknowledgements

It is a great pleasure to acknowledge with sincere gratitude those people who have assisted me in my doctoral journey.

First and foremost, I would like to express my deepest appreciation to Prof. Biao Huang, my research supervisor and life mentor, whose influence on my intellectual and personal growth has been exceptionally valuable. His willingness to give his time so generously to help me to discover my true expectations and inspirations for my research as well as my career is very much appreciated. I am grateful to him not only for his scholarly guidance and useful critiques throughout my graduate studies, but also for his infinite patience and continuous encouragement during the period I was struggling with the mid-PhD crisis.

I would like to thank the members of my defense committee, Drs. Sirish L. Shah, Vinay Prasad, Hong Zhang, and Jay H. Lee for helpful suggestions provided. My special thanks are also extended to Dr. Manabu Kano and Mr. Sanghong Kim for numerous helpful suggestions and inspiring discussions while I was a visiting researcher at the Kyoto University in Japan.

I have greatly benefited from the opportunities provided by Alberta Health and Wellness, Syncrude Canada Ltd., and Suncor Energy Inc. to gain industrial research experience. I appreciate the feedback offered by Drs. Fangwei Xu, KwanHo Lee, Enbo Feng, Elom Domlan, Ramesh Kadali, and Errol Goberdhansingh. I would like to offer my special thanks to Aris Espejo from Syncrude Canada Ltd. for putting his confidence in me, encouraging me to assume challenging roles, and providing me with his invaluable

guidance.

I have had the support and enjoyed the friendship of many wonderful members of the Process Control Group at the University of Alberta. I am particularly grateful to Aditya Tulsyan, Mulang Chen, Lei Chen, Nima Sammaknejad, Elham Naghoosi, Ming Ma, Ruben Gonzalez, Mahdi Ale Mohammad, Drs. Fadi Ibrahim, Swanand Khare, Da Zheng, Yu Miao, Yuri Shardt, Yu Zhao, and Venkat Raghavan for our stimulating discussions as well as their constructive comments and warm encouragement during the best and worst moments of my graduate studies.

I would like to thank the National Science and Engineering Research Council (NSERC) of Canada for the financial support granted through the Alexander Graham Bell Doctoral Scholarship.

I owe my profoundest debt of gratitude to my parents along with my grandmother, who always taught me that I could achieve anything if I believed in myself. Thank you for supporting me in all kinds of ways, for enduring my absence, and for not allowing me to forget the important things in life. Without your unconditional love, support, and encouragement I would not be where I am today. I also would like to thank my sister and everlasting friend, Shiva, for being present and supporting me in every way possible. Together or apart, you are always the greatest source of laughter and joy in my life. I would like to extend my gratitude to my parents-in-law, whom I count as both friends and family. Their commitment to education and their independence of mind have been an inspiration to me.

There were many evenings and weekends spent with my face in a monitor. My task dedication during this lengthy process would not have been possible without my incredible Kasra. Thank you not only for being the most accommodating husband in the world, but also for being my best friend who has continuously bumped me forward. You have forced me to new levels and motivated me to do more than I ever thought possible.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

AIC	Akaike's information criterion
AIC_c	Second-order information criterion
ANN	Artificial neural network
ARX	Autoregressive with exogenous input
BIC	Bayesian information criterion
BLVR	Bayesian latent variable regression
BPCA	Bayesian principle component analysis
CDF	Cumulative distribution function
CFR	Continuous fermentation reactor
CI	Cell index
CSTH	Continuous stirred tank heater
CSTR	Continuous stirred tank reactor
CVA	Canonical variate analysis
D:B	Diluent to Bitumen ratio
DA	Data augmentation
DCS	Distributed control system
DTW	Dynamic time warping

EKF	Extended Kalman filter
EM	Expectation-maximization
EnKF	Ensemble Kalman filter
ERM	Empirical risk minimization
FIR	Finite impulse response
IPS	Inclined plate settler
JPDF	Joint probability density function
LOOCV	Leave-one-out cross-validation
LS-SVM	Least squares support vector machine
LW-PLS	Locally weighted partial least squares
LWR	Locally weighted regression
MAD	Median absolute deviation from median
MAE	Mean absolute error
MAP	Maximum a posteriori
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov chain Monte Carlo
MI	Multiple imputation
MIMO	Multiple-input multiple-output
MISO	Multiple-input single-output
ML	Maximum likelihood
MPC	Model predictive control
MRSE	Mean relative estimation error

MSE	Mean squared error
MVT	Multivariate trimming
NARX	Non-linear autoregressive with exogenous input
NIR	Near-infrared
NMAR	Not missing at random
OLS	Ordinary least squares
OPC	Object linking and embedding for process control
PCA	Principle component analysis
PCR	Principle component regression
PDF	Probability density function
PF	Particle filter
PLS	Partial least squares
PsBF	Pseudo-Bayes factor
PSV	Primary separation vessel
PWARX	Piecewise autoregressive with exogenous input
PWOE	Piecewise output error
QAIC	Quasi-likelihood information criterion
RBS	Random binary sequence
RI	Reliability index
RLS	Recursive least squares
RMSE	Root mean squared error
RMSECV	Root mean squared error of cross-validation
RPLS	Recursive partial least squares

RT-CES	Real-time cell electronic sensing
RVP	Reid vapor pressure
SIC	Schwarz information criterion
SRM	Structural risk minimization
SSVS	Stochastic search variable selection
StdE	Standard deviation of errors
SVC	Support vector classification
SVM	Support vector machine
SVR	Support vector regression
TIC	Takeuchi' information criterion
TMC	Total misclassification cost
UKF	Unscented Kalman filter

Mathematical Symbols

$\mathbb{E}$	Expectation operator
$\mathbb{F}$	Feature space
$\mathbb{R}$	Set of real number
$\mathcal{D}$	Identification (training) data-set
$\mathcal{H}$	Model structure
$\mathcal{J}$	Cost function
$\mathcal{K}$	Kernel function
$\mathcal{L}$	Likelihood
$\mathcal{N}$	Normal distribution
∇	Gradient

Ω	Mapping function
Φ	Set of hyperparameters
σ^2	Standard deviation
$\sim$	Probability distribution
$\mathbf{A}$	Matrix
$\mathbf{a}$	Vector
$\mathbf{I}$, $\mathbf{I}_N$	The $N \times N$ identity matrix
Θ	Set of model parameters
$\vec{\mathbf{0}}_N$	An *N*-long all-zeros vector
$\vec{\mathbf{1}}_N$	An *N*-long all-ones vector
a	Scalar
p	Probability density function
$tr(\mathbf{A})$	Trace of the matrix $\mathbf{A}$
z^{-1}	Back-shift operator
det	Determinant
diag($\mathbf{a}$)	Diagonal matrix whose entries are the elements of the vector $\mathbf{a}$
exp	Exponential function
log	Logarithm

Chapter 1

Introduction

1.1 Motivation

In many industrial applications, successful implementation of advanced monitoring and control techniques highly depends on representativeness of identified process models as well as accuracy and reliability of measurements (Qin and Badgwell, 2000). Specifically, real-time analysis of key performance indicators constitutes an essential prerequisite for advanced monitoring and control of industrial processes. However, on-line measurement of process quality variables is often restricted by inadequacy of measurement techniques or low reliability of measuring devices. Even if appropriate instrumentation exists, the key performance indicators are normally determined by off-line sample analysis in laboratory or on-line product quality analyzers which are often expensive and require frequent and high-cost maintenance. Furthermore, discontinuity and significant delays associated with laboratory analysis or slowly-processed quality measurements of on-line analyzers can reduce the efficiency of control policies. In industrial processing plants, such limitations can have a severe influence on the quality of products, production of waste, and safety of operations.

In the last two decades, there has been a growing interest in the development of **inferential models**, also called **soft sensors**, to provide frequent on-line estimates of quality variables on the basis of their correlation with real-time process measurements.

Such predictive models devoted to producing real-time estimates of desired plant variables can help to reduce the need for measuring devices, improve system reliability, and develop tight control policies (Fortuna *et al.*, 2007).

There are several advantages of inferential sensors in comparison with traditional instrumentation:

1. They give more insight into the process through capturing the information hidden in data.

2. They are an emerging technology that allows industrial users to improve productivity, become more energy efficient, reduce environmental impact, and improve business profitability by reducing the production cost associated with off-specification products.

3. They can be easily implemented on existing hardware. Further, various on-line identification algorithms can be used to maintain the model when system parameters change.

4. They involve little or no capital costs such as the cost of installation, management of the required infrastructure, and commissioning.

The range of tasks fulfilled by inferential sensors is broad. Inferential models may not only be used as a substitute or complement to physical sensors, but can also perform several other tasks efficiently. Good reviews of inferential sensor applications in a number of different fields of process engineering can be found in Stephanopoulos and Han (1996); Chiang *et al.* (2001); Qin and Badgwell (2003); Fortuna *et al.* (2007); Kadlec *et al.* (2009); Kano and Ogawa (2010). At a general level, these fields can be divided into three broad categories:

1. Process monitoring

- Substituting/complimenting on-line instrumentation
- Predicting process quality variables or key performance indicators
- Monitoring and analysis of process trends
- Fault detection

2. Process control

   - Development of advanced control strategies, such as model predictive control
   - Heuristics and logic in planning and scheduling of process operations

3. Off-line operation assistance

   - Diagnosis of process operations
   - Knowledge-based engineering design
   - Development of plant simulator

As indicated by existing research efforts, development and implementation of industrial inferential sensors entail many challenges (Hangos and Cameron, 2001; Paoletti *et al.*, 2007; Kadlec *et al.*, 2009; Pani and Mohanta, 2011; Kano and Fujiwara, 2013). Despite the increasing number of publications dealing with industrial applications, several issues remain open for future investigation. The main objective of this research is to develop novel Bayesian frameworks to reformulate and solve some of these outstanding problems. Specific problems investigated in this thesis are briefly described in Section 1.2. An equally important objective of this work is to implement the developed Bayesian frameworks in experimental and industrial case studies to demonstrate practicality and validity of the methods.

1.2 Problems of Interest

Incorporation of prior process knowledge. Depending on the level of *a priori* knowledge, two different philosophies may guide the choice of modeling strategies, namely **first principles analysis** and **statistical data analysis** (Ljung, 1999). First principles or knowledge-driven models are obtained based on formulating and solving a set of differential and algebraic equations representing physical phenomena. Development of such models requires a deep understanding of transport phenomena, possible reaction pathways, and thermodynamic behavior of the studied systems. The complexity of chemical, petroleum, and biological processes could make first principles modeling infeasible or prohibitively difficult. Therefore, decades of research have been devoted to developing system identification techniques for situations in which complete understanding of the internal mechanisms governing the system dynamics is absent. Traditionally, data-driven models are constructed only based on computational inference of historical relations among system components. However, it has been widely realized that good modeling practice requires exploitation of all available sources of information. The limited knowledge offered by first principles analysis of known mechanisms may form the core of inferential process models, while the impacts of the observed but not sufficiently understood phenomena can be accounted for through system identification and computational inference techniques. Therefore, synthesizing the information obtained from first principles analysis and statistical data analysis is one of the issues arising in some inferential sensing problems. Since available process knowledge cannot be easily expressed in many of the classical formalisms, it might become challenging to fully incorporate *a priori* knowledge about the process operation and underlying mechanisms into the identification procedure.

Controlling the complexity of inferential models. Proper identification of a representative process model is another essential prerequisite for development of an

efficacious inferential sensor. The choice of knowledge-driven models for industrial processes depends on the complexity of the underlying physical systems and thus the availability of phenomenological knowledge of the involved unit operations. In the absence of any process knowledge, the task is to find a suitable inferential model that is well supported by historical data. Therefore, a data-driven model is identified without investigating the internal mechanisms. In such cases, the main criteria to be considered in model selection are simplicity, generality, and flexibility (Hangos and Cameron, 2001). The more degrees of freedom are allowed in the inferential model structure, the closer the model can approximate the identification data-set. On the other hand, too much flexibility might reduce the generalization performance of the developed inferential sensor when the process is operated under a wide range. Determination of a proper model structure plays a key role in achieving a compromise between accuracy and complexity of the model. The problem of model complexity control is often dealt with in a rather *ad hoc* manner. Thus, there is a need to develop a reliable systematic method for model structure selection.

Identification of inferential models from contaminated data. Some of the challenging issues encountered in inferential sensing problems arise due to the varying quality of industrial data. In the context of process industry, measurement noise, missing measurements, outlying observations, multi-rate data, and drifting disturbances are the common factors affecting the quality of operational and laboratory data. Satisfactory performance of inferential sensors can be achieved only if such challenging issues are addressed. Therefore, process data analysis in general and data quality assessment/pre-processing in particular is of essential significance for design of inferential sensors. The problems of process data pre-processing and inferential model identification are often interconnected. However, most of the existing solutions are disconnected and each solution targets mainly one problem. Therefore, it is desired to seek for a unified framework that simultaneously considers different aspects of data analysis and inferential modeling.

Design and implementation of multi-model inferential sensors. Representation of multi-modal processes is another issue that may arise in the identification of inferential models. Some chemical processes experience discrete changes superimposed on their predominantly continuous dynamic behavior. The continuous-state dynamics is typically associated with physical phenomena involved, while the discrete-state dynamics may come from switching controllers, inherent non-linearities in the system, different operating conditions, or any other external discrete events influencing the process under investigation. In such cases, only **multi-model** inferential sensors can describe both the continuous dynamic behavior and the transitions between discrete modes. Multi-model inferential sensors can also be used to approximate complex processes by concatenating multiple local models with simple structures. Real-time model identification (Cleveland, 1979; Atkeson *et al.*, 1997), also known as **just-in-time/space modeling** (Zheng and Kimura, 2001), can be viewed as a special modeling technique for design of multi-model inferential sensors with infinite number of local models. The multi-model paradigm has attracted increasing attention in the process control community due to its many potential industrial applications. Commonly, the existing identification methods hinge on the assumption that any operating space can be partitioned into a finite number of linearly separable regions. Consequently, identification data points lying in the proximity of the intersection of multiple regions cannot be effectively handled. Besides, available process knowledge and relevant background information cannot be easily incorporated in partitioning the operating space and identifying the sub-systems. Therefore, development and implementation of multi-model inferential sensors require further investigation to meet the specific requirements of the process industries.

Monitoring the real-time performance of inferential sensors. Real-time performance assessment of inferential sensors is another important topic to be further investigated. In order to maintain the reliability of an inferential sensor, it is required to track its on-line

performance. However, designing a performance index and specifying a threshold are not straightforward. The main body of research in this area has been focused on exploiting advanced strategies for development of inferential sensors; only a few publications have provided methodologies for on-line reliability analysis of inferential models. The proposed methodologies are rather *ad hoc* and have a number of practical and theoretical limitations. Hence, it is of paramount importance to search for general criteria and techniques for on-line performance assessment of inferential sensors.

In view of the aforementioned challenging issues, this thesis concerns formulating the stated problems of interest as rigorous conditional probabilistic problems within systematic Bayesian frameworks. In principles, Bayesian methods suggest a general solution for many types of systems including linear and non-linear systems, in the presence of Gaussian or non-Gaussian disturbances, with or without constraints, and in handling regular or irregular data samples. As a result of the demonstrated potential of Bayesian methods in dealing with certain outstanding issues associated with inferential modeling, interest in investigating these methods has grown in recent years. Combined with a suite of inference and learning algorithms, Bayesian methods have proven to be powerful in many applications (Korb and Nicholson, 2004; Khatibisepehr and Huang, 2008; Shao *et al.*, 2011; Qi and Huang, 2011). However, these methods are not yet widely applied to inferential sensing practices in the process industry.

1.3 An Overview of Bayesian Inference

Bayesian philosophy originates from an interpretation of Bayes' theorem (Bayes, 1763/1958), which updates the probability of a query variable, x, conditioned on observed data, $\mathcal{D}$, in the light of new information (Korb and Nicholson, 2004):

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p(x)}{p(\mathcal{D})} \tag{1.1}$$

where,

- $p(x)$ is the prior probability of the query variable, which represents the state of knowledge about x before incorporating any information about the observed data, $\mathcal{D}$.

- $p(x|\mathcal{D})$ is the posterior probability of the query variable, which is derived from or depends upon the observed data.

- $p(\mathcal{D}|x)$ is the conditional probability of the observed data given that the query variable takes on a certain value; as a function of x given $\mathcal{D}$, it is also called the likelihood function.

- $p(\mathcal{D})$ is the prior or marginal probability of $\mathcal{D}$, and acts as a normalizing constant, *i.e.* $p(\mathcal{D}) = \int p(\mathcal{D}|x)p(x)dx$.

Bayes' theorem is commonly applied to solving probabilistic inference problems. The steps considered necessary in performing Bayesian inference can be summarized as follows. First, the subjective and/or objective prior knowledge is taken into consideration in order to specify the prior distribution of query variables, $p(x)$. Next, the effect of observed data is investigated in order to incorporate the likelihood of various values of x by multiplying the prior distribution with the likelihood function, $p(\mathcal{D}|x)$. The posterior distribution of query variables, $p(x|\mathcal{D})$, is thus more concentrated than the prior distribution. Finally, the maximum *a posteriori* (MAP) estimates of query variables are obtained from the following expression:

$$x^{\text{MP}} = \underset{x}{\operatorname{argmax}}\, p(x|\mathcal{D}) \tag{1.2}$$

It is noteworthy that all conclusions drawn from evaluation of posterior distributions depend on the quality and extent of the prior information included in Bayesian inference processing. Nevertheless, the choice of prior distribution becomes less significant as more observations are collected. In the case of the non-informative priors (*i.e.*, uniform distribution), the MAP estimates are identical to the maximum likelihood (ML) estimators

which can be expressed as follows:

$$x^{\mathrm{ML}} = \underset{x}{\operatorname{argmax}}\, p(\mathcal{D}|x) \tag{1.3}$$

There are two other types of intractable problems inherently related to the Bayesian statistics that play an important role in Bayesian inference:

- **Marginalization**: Given the joint probability density function $p(x, y)$, the marginal probability density function of the random variable x can be obtained by integrating out y such that

$$p(x) = \int_Y p(x, y)dy \tag{1.4}$$

- **Expectation**: Given the conditional probability density function $p(x|y)$, the expected value of an arbitrary function of the random variable x, $g(x)$, is calculated as

$$\mathbb{E}_{p(x|y)}[g(x)] = \int_X g(x)p(x|y)dx \tag{1.5}$$

Adopting Bayesian methods to formulate and solve the inferential sensing problems bears several benefits.

1. Process knowledge can be easily incorporated in a Bayesian scheme by specifying proper prior distributions over model parameters, functional forms, and constraints.

2. Bayesian methods force one to make the tacit assumptions explicit in the prior distributions. In this way, the assumptions are easier to evaluate, criticize, and modify.

3. The model identification problem can be rigorously formulated under a principled framework, which features fewer heuristic design choices. For instance, a Bayesian approach to modeling can naturally deal with complexity control to avoid over-fitting by integrating out the uncertain model parameters and/or hyperparameters.

4. Within a full Bayesian framework, the uncertainty in model parameters is characterized through posterior probability density functions which give rise to a so called **predictive distribution**. Thus, probabilistic predictions are made by marginalizing over the parameters.

5. General Bayesian learning techniques convert the identification problem into an equivalent problem of computing expectation or evaluating an integral as opposed to solving a global optimization problem as in likelihood methods.

6. Incomplete data and non-Gaussian distributions can be handled naturally.

1.4 Thesis Outline

The main contributions of this research are presented in six chapters the contents of which have been published or to be published in peer-reviewed journals:

1. Khatibisepehr, S., B. Huang and S. Khare (2013). Design of inferential sensors in the process industry: A review of Bayesian methods. *Journal of Process Control*. in press.

2. Khatibisepehr S., B. Huang, E. Domlan, E. Naghoosi, Y. Zhao, Y. Miao, X. Shao, S. Khare, M. Keshavarz, E. Feng, F. Xu, A. Espejo and R. Kadali (2013). Soft sensor solutions for control of oil sands processes. *The Canadian Journal of Chemical Engineering* **91**(8), 1416-1426.

3. Khatibisepehr, S. and B. Huang (2013). A Bayesian approach to robust process identification with ARX models. *AIChE Journal* **59**(3), 845-859.

4. Khatibisepehr, S. and B. Huang (2012). A Bayesian approach to design of adaptive multi-model inferential sensors with application in oil sand industry. *Journal of Process Control* **22**(10), 1913-1929.

5. Khatibisepehr S., B. Huang, F. Ibrahim, J.Z. Xing and W. Rao (2011). Data-based modeling and prediction of cytotoxicity induced by contaminants in water resources. *Computational Biology and Chemistry* **35**(2), 69-80.

6. Khatibisepehr S., B. Huang, S. Khare, E. Domlan, F. Xu, A. Espejo and R. Kadali. A probabilistic framework for real-time performance assessment of inferential sensors. Submitted to *Control Engineering Practice*.

The organization of the thesis is as follows.

Chapter 2 provides a general introduction to the main steps involved in development and implementation of industrial inferential sensors, and presents an overview of the relevant Bayesian literature. The potential Bayesian solutions to some of the main issues associated with inferential sensor design are discussed. A review of the literature on the industrial applications of Bayesian inferential sensors is also presented.

Chapter 3 provides a classical non-Bayesian framework to capture the non-linearity in the local region around a query point in a real-time manner. The proposed real-time model identification approach, also known as just-in-time/space modeling, can cope with variations in process characteristics and handle non-linearity of underlying mechanisms. An ν-support vector regression (ν-SVR) model is adopted to form the core of the predictive framework. The formulation of the SVR embodies the structural risk minimization (SRM) principle that is used to minimize an upper bound on the expected risk. Given a query point, a search algorithm is applied to select spatial and temporal nearest neighbors within the identification data-set. The selected sub-set of identification data is then used to identify a local ν-SVR model. Since the SRM principle provides means of constructing regularized risk functions, it can be motivated from a Bayesian perspective. The regularization term included in the regularized risk function can be interpreted as the prior belief over the level of complexity of the SVR model structure. The developed framework is implemented to facilitate real-time modeling and prediction of cytotoxicity effects on living cells induced

by certain water contaminants. The structural risk minimization approach is used for model order selection, while the cross-validation is performed for hyperparameter tuning.

In the method proposed in Chapter 3, the search for optimal model structure and hyperparameters are not interconnected. In Chapter 4, a unifying Bayesian framework is developed to facilitate real-time model structure selection and similarity function parametrization in just-in-time/space modeling methods. The proposed framework would bridge the gap between the model structure selection and hyperparameter tuning. Since partial least squares (PLS) regression can effectively handle the collinear identification data, the locally weighted PLS algorithm is adopted as the main modeling technique. It is assumed that the operating space can be partitioned into a finite number of sub-spaces. A Bayesian procedure is outlined for partitioning and characterizing the operating space. For each sub-space, the problem of finding the locally optimal LW-PLS model structure and similarity function hyperparameters is formulated under an iterative hierarchical Bayesian optimization framework. Thus, the real-time identification problem amounts to detecting the underlying operating sub-space and estimating the LW-PLS model parameters. The proposed method has the following attractive features: 1. The Bayesian model comparison allows us to perform objective comparisons between alternative model structures. Therefore, the resulting optimization problem in each sub-space would automatically be subjected to model complexity control to avoid over-fitting. 2. Objective criteria for local tuning of the hyperparameters of the similarity function are provided. 3. Real-time model structure selection and similarity function parametrization would become computationally efficient.

In Chapter 5, the problem of inferential model identification in presence of outliers is formulated and solved under a robust Bayesian framework consisting of consecutive levels of optimization. The resulting optimization problem was hierarchically decomposed and a layered optimization strategy was implemented. An iterative hierarchical Bayesian

approach is adopted to coordinate the solutions obtained in subsequent layers of optimization. The proposed optimization strategy not only yields maximum *a posteriori* estimates of model parameters, but also provides an automated mechanism for determining the hyper-parameters and investigating the quality of each observation. Moreover, the developed framework allows us to incorporate the prior knowledge of the contaminating distributions. Thereby, the restrictive assumptions made in traditional robust identification methods about contaminating distributions (*e.g.* symmetric noise distribution) are relaxed.

Chapter 6 presents a Bayesian approach for identification of multi-modal systems switching among non-linear continuous-state dynamics to meet the specific requirements of the process industries. The proposed identification procedure provides a framework to accommodate the overlapping operating regions and facilitate the inclusion of prior knowledge about the operating conditions. A Bayesian decision-support scheme has also been developed for real-time implementation of the multi-model inferential sensors. The developed scheme includes a global adaptation mechanism, within the envelope of previously identified operating conditions. The efficacy of the method is demonstrated through a successful industrial application of an adaptive multi-model inferential sensor designed for real-time monitoring of a key quality variable in an oil sands processing unit.

Chapter 7 presents a data-driven Bayesian approach for real-time performance assessment of inferential sensors. A statistical inference framework is developed to capture conditional dependence of the reliability of inferential sensor predictions on characteristics of the input space and reliability of the empirical process model. The details of the proposed Bayesian method are presented for both discrete and continuous operating statuses. Real-time performance assessment of multi-model inferential sensors is also discussed. The proposed method has the following attractive features: 1. *A priori* knowledge of process operation and underlying mechanisms can be easily incorporated in identifying the criteria for real-time performance assessment of the designed inferential

sensors. 2. Since probability density functions would reflect the actual data distribution, empty regions within the identification data-set can be diagnosed. 3. Correlations between input variables are taken into account. 4. Contribution of each input variable to prediction uncertainty is automatically considered. 5. Application of the method does not depend on the identification techniques employed for inferential model development. 6. Real-time implementation of the method is computationally efficient.

In Chapter 8, the methods proposed throughout the thesis are incorporated to lay out a novel unified Bayesian framework for the design of multi-model inferential sensors. The chapter also includes a summary of the major contributions of the thesis as well as recommendations for future research.

1.5 Main Contributions

This thesis can be used as a guide to Bayesian inferential sensing practice in process industries. The main contributions of this work are summarized below.

1. Development of an integrative and holistic Bayesian framework for design of adaptive multi-model inferential sensors from contaminated industrial data with little or no need for subjective knowledge. The proposed approach is the first attempt to integrate the otherwise disjoint steps required for development of inferential sensors including data quality assessment and model identification.

2. Providing objective criteria for simultaneous model structure selection and similarity function parametrization in just-in-time/space modeling methods. The developed framework provides a systematic approach for model structure selection and similarity function parametrization.

3. Proposing a reliability analysis methodology for real-time performance assessment of inferential sensors. The proposed method is of paramount importance to the

implementation of industrial inferential sensors.

4. Development of a data-driven framework to facilitate real-time prediction of cytotoxicity effects on living cells induced by certain water contaminants. The developed framework allows us to analyze intrinsic cell behavior and predict the trajectory of its progress (growth or death) over a considerable time horizon.

5. Design of adaptive multi-model inferential sensors for real-time monitoring of key quality indicators of an oil sands processing unit. The developed inferential sensors have been running on-line reliably and successfully since July 2011.

Bibliography

Atkeson, C. G., A. W. Moore and S. Schaal (1997). Locally weighted learning. *Artificial Intelligence Review* **11**, 11–73.

Bayes, T. (1763/1958). An essay towards solving a problem in the doctrine of chances. *Biometrika* **45**(3-4), 296–315.

Chiang, L. H., E. L. Russell and R. D. Braatz (2001). *Fault Detection And Diagnosis In Industrial Systems*. first ed.. Springer-Verlag. London, UK.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**(368), 829–836.

Fortuna, L., S. Graziani, A. Rizzo and M. G. Xibilia (2007). *Soft Sensors for Monitoring and Control of Industrial Processes*. first ed.. Springer-Verlag. London, UK.

Hangos, K. M. and I.T. Cameron (2001). *Process Modelling and Model Analysis*. first ed.. Academic Press. San Diego, USA.

Kadlec, P., B. Gabrys and S. Strandt (2009). Data-driven soft sensors in the process industry. *Computers and Chemical Engineering* **33**(4), 795–814.

Kano, M. and K. Fujiwara (2013). Virtual sensing technology in process industries: Trends and challenges revealed by recent industrial applications. *Journal of Chemical Engineering of Japan* **46**(1), 1–17.

Kano, M. and M. Ogawa (2010). The state of the art in chemical process control in Japan: Good practice and questionnaire survey. *Journal of Process Control* **20**(9), 969–982.

Khatibisepehr, S. and B. Huang (2008). Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Industrial and Engineering Chemistry Research* **47**(22), 8713–8723.

Korb, K. B. and A. E. Nicholson (2004). *Bayesian Artificial Intelligence*. first ed.. Chapman & Hall/CRC. London, UK.

Ljung, L. (1999). *System Identification - Theory For the User*. second ed.. Prentice Hall. Upper Saddle River, USA.

Pani, A. K. and H. K. Mohanta (2011). A survey of data treatment techniques for soft sensor design. *Chemical Product and Process Modeling*.

Paoletti, S., A. L. Juloski, G. Ferrari-Trecate and R. Vidal (2007). Identification of hybrid systems: A tutorial. *European Journal of Control* **13**(2-3), 242–260.

Qi, F. and B. Huang (2011). Bayesian methods for control loop diagnosis in presence of temporal dependent evidences. *Automatica* **47**(7), 1349–356.

Qin, S. J. and T. A. Badgwell (2000). An overview of nonlinear model predictive control applications. In: *Nonlinear Model Predictive Control* (F. Allgöwer, A. Zheng and C. I. Byrenes, Eds.). Vol. 26 of *Progress in Systems and Control Theory*. pp. 369–392. Birkhäuser Basel. Basel, Switzerland.

Qin, S. J. and T. A. Badgwell (2003). A survey of industrial model predictive control technology. *Control Engineering Practice* **11**(7), 733–764.

Shao, X., B. Huang, J. M. Lee, F. Xu and A. Espejo (2011). Bayesian method for multirate data synthesis and model calibration. *AIChE Journal* **57**(6), 1514–1525.

Stephanopoulos, G. and C. Han (1996). Intelligent systems in process engineering: a review. *Computers and Chemical Engineering* **20**(6/7), 743–791.

Zheng, Q. and H. Kimura (2001). Just-in-time modeling for function prediction and its applications. *Asian Journal of Control* **3**(1), 35–44.

Chapter 2

Design of Inferential Sensors in the Process Industry: A Review of Bayesian Methods

2.1 Introduction

Depending on the level of *a priori* knowledge of the process, three different classes of inferential models can be developed: 1. knowledge-driven, 2. data-driven, and 3. gray-box models. Knowledge-driven models, also called **first-principles models**, are developed on the basis of first principles analysis and, thus, require full phenomenological knowledge about the underlying mechanisms (Grantham and Ungar, 1990; Prasad *et al.*, 2002; Friedman *et al.*, 2002; Cinar *et al.*, 2003). Although first-principles models have many advantages, they can often be expensive and time-consuming due to the complexity of industrial processes. In contrast, data-driven models are constructed only based on the historical relations among the existing measurements, and prevent one from the laborious study of complex chemical and physical phenomena involved (Kano and Nakagawa, 2008; Olanrewaju *et al.*, 2010; Wang *et al.*, 2010*a*; Jampanaa *et al.*, 2010). Data-driven models, also called **black-box models**, are proposed for situations in which physical understanding of the process under investigation is absent or not relevant. In between the two extremes,

A version of this chapter has been accepted for publication in Journal of Process Control (Khatibisepehr *et al.*, 2013).

there are many possible combinations of knowledge-driven and data-driven models. The prior knowledge offered by the simplified first principles analysis forms the core of a so called **gray-box model**, while data-driven methods can compensate for fractions that cannot be modeled easily in terms of phenomenological models. Satisfactory results of gray-box techniques have been widely reported in literature, because any available source of information is exploited to refine such models (Bohlin and Graebe, 1995; Dadhe *et al.*, 2001; Jiaa *et al.*, 2011; Liu *et al.*, 2012). In this paper, we focus our attention mainly on gray-box modeling due to its growing popularity in industrial applications.

Regardless of which modeling approach is taken, an inferential sensor design procedure is an iterative process consisting of the following steps: 1. Process data analysis, 2. Model identification, 3. Model validation, 4. Model implementation and calibration. Figure 2.1 presents a flowchart of the inferential sensor design procedure. As a general guideline, Appendix A outlines some of the main tasks to be performed at different stages. Since each industrial application has unique requirements and challenges, the recommended tasks does not necessarily include all steps required for development or implementation of a specific inferential sensor.

Development and implementation of industrial inferential sensors entail many challenges. As a result of the demonstrated potential of Bayesian methods in dealing with certain outstanding issues associated with inferential modeling, interest in investigating these methods has grown in recent years. As indicated by existing research efforts, Bayesian methods suggest a general solution for many types of systems including linear and non-linear systems, in the presence of Gaussian or non-Gaussian disturbances, with or without constraints, and in handling regular or irregular data samples. Combined with a suite of inference and learning algorithms, Bayesian methods have proven to be powerful in many applications (Korb and Nicholson, 2004; Khatibisepehr and Huang, 2008; Shao *et al.*, 2011; Qi and Huang, 2011). Despite the increasing number of publications dealing with

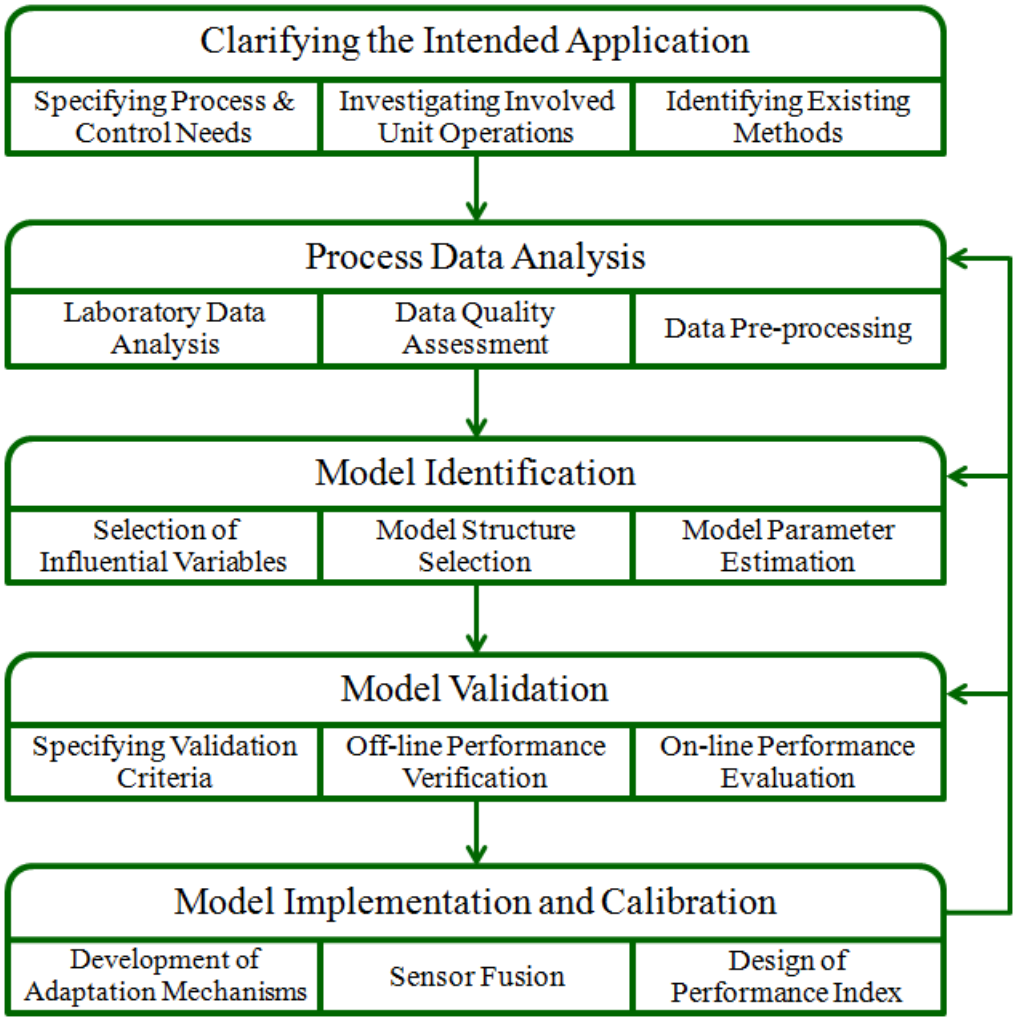


Figure 2.1: Flowchart of the inferential sensor design procedure

industrial applications, these methods are not yet widely applied to inferential modeling practices in the process industry.

The purpose of this Chapter is twofold. The first objective is to provide a general introduction to the main steps involved in development and implementation of industrial inferential sensors, and present an overview of the relevant Bayesian literature. The use of Bayesian techniques in industrial applications, in particular in design of inferential sensors for process industries, is relatively new. An equally important objective of this Chapter is thus to discuss the potential Bayesian solutions to some of the main issues associated with inferential sensor design. A review of the literature on the industrial applications of

Bayesian inferential sensors is also presented. This Chapter is not intended to provide a comprehensive review of the great variety of methods used in the design of inferential sensors, but is rather focused on the techniques that have their origin in Bayesian Statistics. Therefore, the main contribution of this work is complementing the existing reviews in the field. Gonzalez (1999); Fortuna *et al.* (2007); Kadlec *et al.* (2009); Pani and Mohanta (2011); Kano and Fujiwara (2013) are among the most cited publications providing reviews of the inferential sensor applications, the most popular inferential modeling techniques, and the challenging issues involved.

2.2 Process Data Analysis

Process data analysis is the initial step in the design of inferential sensors. Careful investigation of laboratory and operational data enables us to extract relevant information contained in historical data, select influential variables, and assess data quality (*e.g.* reliability, accuracy, completeness, and representativeness). In particular, the query variable measurements should be thoroughly assessed to ensure that reference data of sufficiently high quality and variability are used in the design of inferential sensors. In this phase, conducting interviews with plant experts and operators plays a key role in fully exploiting the wealth of historical data. The experiences and expertise of those involved in day to day operation provide valuable insight into underlying mechanisms, relevant process variables, performance of measuring devices, and operating modes, among others.

The collected process data is often divided into three subsets: the identification data-set, the validation data-set, and the test data-set. The identification data-set is used for inferential model identification purposes, while the validation and test data-sets are reserved for evaluating the performance of the developed inferential sensor. The difference between the latter two data-sets will be explained in Section 2.4. It is noteworthy that the distribution of identification data within the process operating region is crucial to ensuring

the quality of inferential sensors. Therefore, the identification data-set should adequately represent the possible operating modes of the process under investigation.

In Bayesian data analysis, marginal and joint probability distributions of observed and query variables are investigated in order to extract hidden patterns (*e.g.* dependencies and multiple operating ranges) from historical process data. These patterns can be considered as a summary of the input data, which can then be used to obtain more accurate results by a decision support system (Gelman *et al.*, 2003).

2.2.1 Characteristics of Laboratory Data

From an inferential modeling point of view, laboratory data are often considered as reference measurements. Therefore, it is important to acquire sufficient knowledge about the characteristics of laboratory data, such as sampling interval, sampling procedure, analysis techniques, and basics of measuring devices.

Although laboratory analysis often provides accurate and reliable measurements, the quality of laboratory data might be affected by the following factors:

1. The exact sampling instants are often not recorded; rather nominal time, as required by the assigned sampling intervals, is attributed to each collected sample.

2. There are potential human errors that may occur in collecting samples, conducting experiments, and recording the results.

3. The laboratory equipment is frequently calibrated within a specified accuracy. However, the tolerable range of inaccuracy for a certain instrument may introduce a considerable error when the operational range of variation of the measured property is relatively small.

Therefore, laboratory data quality assurance is of essential significance for design of inferential sensors.

2.2.2 Data Pre-processing

Development and implementation of inferential sensors entail many challenges that may arise due to the varying quality of industrial data. In the context of process industry, measurement noise, missing measurements, outlying observations, multi-rate data, measurement delay, and drifting disturbances are the common factors affecting the quality of process data. Satisfactory performance of inferential sensors can be achieved only if such challenging issues are addressed. As a preliminary step, data pre-processing is often required in order to obtain a data-set which adequately represents the characteristic properties of the process under investigation (Kadlec *et al.*, 2009; Pani and Mohanta, 2011).

2.2.2.1 Incomplete Data

In many industrial plants, missing measurements and irregularly sampled data are commonly experienced mainly due to hardware sensor failure or routine maintenance, data acquisition system malfunction, different acquisition rates from different sensors, or delays associated with laboratory analysis. Background information about the pattern and extent of data incompleteness is often not included in process data analysis. Rubin (1976) developed a probabilistic framework to describe different plausible assumptions that might be made about the incompleteness mechanisms. Suppose the identification data can be segregated into two parts: the complete and the incomplete attributes, *i.e.* $\mathcal{D} = \{\mathcal{D}_c, \mathcal{D}_{ic}\}$. A matrix of binary indicator variables, $\mathbf{M}$, can be constructed to denote whether or not a data point is observed. The incompleteness mechanism can be described as the posterior probability distribution of $\mathbf{M}$ given the identification data *i.e.* $p(\mathbf{M}|\mathcal{D}, \Phi)$, where Φ denotes hyperparameters characterizing this conditional probability distribution. Depending on the extent of conditional dependence between $\mathbf{M}$ and $\mathcal{D}$, the incompleteness mechanism can be categorized into three classes (Rubin, 1976; Imtiaz and Shah, 2008):

1. **Missing Completely At Random (MCAR)**: The incompleteness mechanism is

defined to be MCAR, if the posterior probability distribution of $\mathbf{M}$ does not depend on any part of $\mathcal{D}$, *i.e.* $p(\mathbf{M}|\mathcal{D}, \Phi) = p(\mathbf{M}|\Phi)$. For instance, incomplete data resulting from instrument failures or transmission problems may not follow a discernible pattern.

2. **Missing At Random (MAR)**: The incompleteness mechanism is defined to be MAR, if the posterior probability distribution of $\mathbf{M}$ depends on complete attributes $\mathcal{D}_c$, but not on the incomplete ones, *i.e.* $p(\mathbf{M}|\mathcal{D}, \Phi) = p(\mathbf{M}|\mathcal{D}_c, \Phi)$; this is a considerably weaker assumption. In some industrial plants, frequent measurement of key performance indicators is costly or time-consuming. In such cases, the process is monitored and controlled through regularly-measured variables. That is, the quality variables are measured only when process variables indicate the process is drifting away from the normal operating conditions. Thus, incompleteness of quality variables depends on the regular measurements of process variables.

3. **Not Missing At Random (NMAR)**: The incompleteness mechanism is defined to be NMAR, if the posterior probability distribution of $\mathbf{M}$ depends on both complete and incomplete attributes. Under this assumption, the cause of incompleteness has to be identified and included in the process data analysis, *i.e.* the cause of incompleteness is not ignorable. For instance, if measured variables violate the technological limitations of the measuring device, the instrument would fail to measure values falling outside its nominal range.

The techniques for handling incomplete data can be divided into two broad categories, namely *ad hoc* methods and statistically principled methods (Little and Rubin, 2002; Osborne, 2008; Mason *et al.*, 2012). Under the MCAR assumption, case-wise deletion and single imputation are the most primitive *ad hoc* methods. It is noteworthy that removal of incomplete samples may lead to a considerable loss of information and biased estimates, which could negatively impact the prediction and/or generalization performance of the

inferential models. Also, the statistical distribution of the data is distorted by substituting all missing values of a variable with a single statistical measure (*e.g.* mean or median of the corresponding observed values) (Schafer and Graham, 2002).

Unlike *ad hoc* methods, statistically principled methods include explicit assumptions about the incompleteness mechanisms to take into the consideration the statistical uncertainty introduced by the imputed values of missing measurements (Osborne, 2008). A wide variety of statistically principled methods has been developed; among them maximum likelihood (ML), multiple imputation (MI), and data augmentation (DA) are most widely used (Dempster *et al.*, 1977; Rubin, 1987; Tanner and Wong, 1987).

Let Θ denote a set of unknown parameters that govern the identification data likelihood, $\mathcal{L}(\Theta|\mathcal{D})$. Under the MAR assumption, all the relevant information about Θ is contained in the fully-observed data likelihood, $\mathcal{L}(\Theta|\mathcal{D}_c) = p(\mathcal{D}_c|\Theta)$. Through marginalization, this likelihood can be expressed as

$$\mathcal{L}(\Theta|\mathcal{D}_c) = \int_{\mathcal{D}_{ic}} p(\mathcal{D}_{ic}, \mathcal{D}_c|\Theta) d\mathcal{D}_{ic} \tag{2.1}$$

Direct maximization of $\mathcal{L}(\Theta|\mathcal{D}_c)$ is often intractable due to the presence of integral. To circumvent the difficulties associated with direct maximization of $\mathcal{L}(\Theta|\mathcal{D}_c)$, the expectation-maximization (EM) algorithm is often used to obtain maximum likelihood estimates of the query parameters. Comprehensive overviews of the formal procedures and key properties of EM algorithm have been given by Rubin (1987); Osborne (2008); Graham (2009). Briefly, the EM algorithm is an iterative procedure consisting of two consecutive steps:

1. **Expectation Step**: Given complete data and the current parameter estimates, calculate the expectation of the logarithm of the likelihood of the full identification

data with respect to the incomplete or missing data:

$$Q_{\Theta|\hat{\Theta}^k} = \mathbb{E}_{p(\mathcal{D}_{ic}|\mathcal{D}_c,\hat{\Theta}^k)}[\log \mathcal{L}(\Theta|\mathcal{D}_{ic},\mathcal{D}_c)] = \int_{\mathcal{D}_{ic}} \log p(\mathcal{D}_{ic},\mathcal{D}_c|\Theta) p(\mathcal{D}_{ic}|\mathcal{D}_c,\hat{\Theta}^k) d\mathcal{D}_{ic} \tag{2.2}$$

Note that rather than directly filling in the missing observations, the sufficient statistics of the full identification data likelihood is used.

2. **Maximization Step**: Maximize the expression above obtained with respect to Θ to find $\hat{\Theta}^{k+1}$:

$$\hat{\Theta}^{k+1} = \underset{\Theta}{\operatorname{argmax}}\, Q_{\Theta|\hat{\Theta}^k} \tag{2.3}$$

Multiple imputation (MI) bears a close resemblance to the EM algorithm. The MI procedure involves the following distinct steps (Graham, 2009; Lin, 2010):

1. Fill in the missing elements of the incomplete data-set with plausible values drawn randomly (with replacement) from a proper predictive distribution, $p(\mathcal{D}_{ic}|\mathcal{D}_c)$, in order to construct M complete data-sets.

2. Perform analysis on each of the complete data-sets applying standard complete-data techniques.

3. Combine the results obtained from the M complete data-sets into a single set of results. The confidence intervals are obtained by calculating the within and between imputation variance.

Suppose that $\mathcal{D} = \{\mathcal{D}_c, \mathcal{D}_{ic}\}$ follows a parametric model $p(\mathcal{D}|\Theta)$, where $\mathcal{D}_{ic}$ is caused by an ignorable incompleteness mechanism. Through marginalization, we obtain

$$p(\mathcal{D}_{ic}|\mathcal{D}_c) = \int_{\Theta} p(\mathcal{D}_{ic}|\mathcal{D}_c,\Theta) p(\Theta|\mathcal{D}_c) d\Theta \tag{2.4}$$

For proper multiple imputation the parameters Θ governing the predictive distribution of $\mathcal{D}_{ic}$ are first sampled from their complete data posterior distribution $\Theta^{(m)} \sim p(\Theta|\mathcal{D}_c)$. Next,

a plausible value of $\mathcal{D}_{ic}$ is randomly drawn from $p(\mathcal{D}_{ic}|\mathcal{D}_c, \Theta^{(m)})$. Therefore, it is natural to motivate multiple imputation from a Bayesian perspective in which the state of knowledge about the parameters is represented through a posterior distribution (Schafer and Graham, 2002).

Markov Chain Monte Carlo (MCMC) methods such as data augmentation (DA) are commonly applied to simulate random draws from Bayesian posterior distributions under complicated parametric models. DA may be viewed as Bayesian counterpart of the EM algorithm in which the deterministic expectation and maximization steps are replaced by their stochastic equivalents. As described by Allison (2002), DA is an iterative process involving two main steps as detailed below:

1. **Imputation Step**: Perform a random imputation of $\mathcal{D}_{ic}$ given the current parameter estimates, *i.e.* $\mathcal{D}_{ic}^{(k+1)} \sim p(\mathcal{D}_{ic}|\mathcal{D}_c, \Theta^{(k)})$.

2. **Sampling Step**: Draw plausible values of unknown parameters from a Bayesian posterior distribution reconstructed from the observed and imputed data, *i.e.* $\Theta^{(k+1)} \sim p(\Theta|\mathcal{D}_{ic}^{(k+1)}, \mathcal{D}_c)$.

This iterative procedure generates a Markov chain that eventually stabilizes to $p(\Theta|\mathcal{D}_c)$ and $p(\mathcal{D}_{ic}|\mathcal{D}_c, \Theta)$, the distributions from which MIs are generated.

Successful applications of EM and DA algorithms to handle incomplete data have been widely reported. Khatibisepehr and Huang (2008) conducted a comparative study on a variety of incomplete data handling techniques widely adopted for process data analysis. The authors concluded that under the assumption of MAR the EM-based Bayesian algorithm outperforms the other procedures in terms of accuracy of the parameter estimates. Raghavan *et al.* (2006) presented an EM-based strategy for data-driven identification of state-space models when output observations are missing at regular or irregular intervals. The proposed strategy is applied to developing an inferential sensor for a bleaching unit at Millar Western's Bleached-Chemi Thermo-Mechanical Pulp (BCTMP) mill located in

Whitecourt, Alberta. The developed inferential sensor is intended to provide optimal predictions of pulp brightness as one of the quality variables of the BCTMP process. Jin *et al.* (2012) presented a linear parameter-varying scheme for inferential sensor development in which the EM algorithm is employed for handling the irregular and incomplete output data. Imtiaz and Shah (2008) proposed a method which combines the principal component analysis imputation algorithm with the ideas of Bootstrap resampling and DA strategies. Moreover, multivariate missing data handling techniques are combined with dynamic time warping (DTW) to synchronize uneven length batch process data. The proposed method conserves the correlation between the variables and leads to a compact latent variable model. Qi *et al.* (2010) proposed a novel Bayesian method based on marginalization over underlying complete evidence matrix to handle incomplete data in data-driven control loop diagnosis. To enhance the MPC performance monitoring for an industrial diluted bitumen heating process, the proposed Bayesian approach is used to synthesize monitor outputs to distinguish different problems of similar phenomena. Ge and Song (2011) introduced a semi-supervised Bayesian method through which the information contained in the incomplete data-set $\mathcal{D}_{ic}$ can be incorporated into development of probabilistic principle component regression (PCR) models. The parameter estimation problem is formulated under the EM framework. The authors have followed their proposed approach to develop various industrial inferential sensors for advanced process monitoring of a sulphur recovery unit and a debutanizing distillation process.

2.2.2.2 Outlying Observations

Industrial data-sets are generally corrupted by the presence of outlying observations, also called **outliers**. Outliers are observations which appear to deviate markedly from the typical ranges of other observations (Grubbs, 1969). The outliers in operational data mostly represent a random error caused by issues such as process disturbances, instrument degradation, and transmission problems (Zeng and Gao, 2009; Lee *et al.*, 2011). In

some cases, however, outliers may arise due to infrequent yet important changes in system dynamics (Hodge and Austin, 2004). Statistical analysis of the process data contaminated with outliers may lead to biased parameter estimation and plant-model mismatch. Therefore, outlier identification constitutes an essential prerequisite for design of inferential sensors (Khatibisepehr and Huang, 2013). A comprehensive review of the outlier identification problem and several outlier identification methods is provided by Hodge and Austin (2004); Kadlec *et al.* (2009); Chandola *et al.* (2009).

Ben-Gal (2010) has distinguished two main categories for outlier detection methods, namely **parametric** and **non-parametric**. Since the focus of this review is on the Bayesian methods, we limit our literature review to parametric, also called **statistical**, approaches.

Statistical methods often indicate those observations that widely deviate from the center of the data distribution. For instance, the simplest statistical technique is the 3σ edit rule in which data are assumed to follow a Gaussian distribution. In this method, a data point x_i is labeled as an outlier if $|x_i - \mu_x| > 3\sigma_x$, where the distribution mean μ_x and standard deviation σ_x are calculated from all attribute values including the query value $x(i)$. To reduce the influence of outliers in estimating the distribution mean and standard deviation, Davies and Gather (1993) introduced the Hample identifier in which median and median absolute deviation from median (MAD) are used to represent the underlying distribution.

Several solutions have been proposed for solving the outlier detection problem by estimating a probability density of the normal data. For instance, in (Bishop, 1994) the density distribution of the input space is first estimated by a standard Parzen window approach with Gaussian kernel functions:

$$\hat{p}(\mathbf{x}) = \frac{1}{n(2\pi)^{d/2}\sigma^d} \sum_{n=1}^{N} \exp\left\{ - \frac{|\mathbf{x} - \mathbf{x}_n|^2}{2\sigma^2} \right\} \tag{2.5}$$

where $\mathbf{x}_n$ represents a data vector from the training set, d is the dimension of input space, and σ is the smoothing parameter. Next, a suitable threshold is specified based on the identification data-set which is known to be representative of normal data. The new

observation is then flagged as an outlier if the value of the density function $\hat{p}(\mathbf{x}_{new})$ is above the threshold. Roberts (1999) suggested the use of extreme value statistics when the identification data-set is contaminated by the presence of outliers. Yu (2012) proposed a Bayesian approach to estimate the posterior probabilities of all samples within the model input space and specify the appropriate confidence levels. A calibration procedure is then followed to correct the observations identified as outliers.

An alternative approach for probability density estimation is to model normal instances as a mixture of parametric distributions. Bishop (1994) and Agarwal (2006) used Gaussian mixture models for such techniques. In Ritter and Gallegos (1997), both normal instances and outliers are modeled as separate parametric distributions *i.e.* $\mathcal{D}_{Mix} = (1-\delta)\mathcal{D}_{Reg} + \delta\mathcal{D}_{Out}$, where δ is the prior probability of appearance of an outlier. First, the ellipsoidal multivariate trimming (MVT) technique (Rousseeuw and Leroy, 1996) is used to detect outliers and to estimate distribution parameters of both outliers and regular observations. Next, a Bayesian classifier is designed to compare certain linear combinations of posterior densities of each data vector with respect to the estimated distributions. In Khatibisepehr and Huang (2013), a contaminated distribution is adopted to describe the observed data and a set of indicator variables is introduced to denote the quality of each data point. The model identification problem in the presence of outliers is then formulated under a robust Bayesian framework consisting of consecutive levels of optimization. The proposed solution strategy not only yields maximum *a posteriori* (MAP) estimates of model parameters, but also provides hyperparameters that determine data quality as well as the prior distribution of model parameters. Assuming a uniform distribution of outliers, Eskin (2000) provided a measure to determine the likelihood of an observation being an outlier by comparing the change in the log likelihood of the mixture distribution (*i.e.* $\log \mathcal{L}_i(\mathcal{D}_{Mix}) - \log \mathcal{L}_{i-1}(\mathcal{D}_{Mix})$) if the observation is removed from the regular instance distribution, $\mathcal{D}_{Reg}$. Several variations of Bayesian classification technique have further

been proposed in Varbanov (1998); Ghosh-Dastidar and Schafer (2006); Das and Schneider (2007), and many others. The underlying principle of these methods is to evaluate the posterior probability of an observation acting as an outlier conditional upon the observed data and estimated values of the distribution parameters.

It is noteworthy that there exist a number of comparative studies on various outlier identification methods such as the work presented in Lalor and Zhang (2001); Penny and Jolliffe (2001); Ben-Gal (2010). These studies have shown the benefits of using a battery of methods to boost the performance of outlier identification procedures.

2.2.2.3 Collinearity

Process variables are often causally related, consequently, process measurements are strongly collinear. From the inferential modeling point of view, such collinear data provide little independent information. Some classical modeling techniques such as regression-based methods cannot deal with collinear identification data. There are several issues that might arise as a result of collinearity (Greene, 2007):

1. Since the identification data do not contain sufficient information to estimate all parameters simultaneously, precision of the estimated parameters would be degraded.

2. Parameters of the identified model may not be statistically significant.

3. Parameter estimates might have incorrect signs and/or implausible magnitudes.

4. Small changes in identification data may result in wide swings in parameter estimates.

Therefore, it is important to assess the degree of collinear relationships among process variables (Belsley and Welsch, 2004). Although a definite criterion for evaluating multi-collinearity does not exist, various techniques have been developed in an attempt to detect and assess collinearity. Draper and Smith (2003) proposed construction of

a correlation matrix of process variables to examine pairwise correlation coefficients. However, relatively large correlation coefficients do not necessarily imply collinearity. Marquardt (1970) suggested to evaluate the magnitude of diagonal elements of the variance-covariance matrix, also known as variance inflation factors. The major drawback of this method is that significant multi-collinearity between three or more variables cannot be indicated. To overcome this shortcoming, Belsley and Welsch (2004) introduced a method called **condition index analysis** to detect multi-collinearity based on singular value decomposition of the data matrix and decomposition of regression variance estimates.

Principal component analysis (PCA), partial least squares (PLS), and canonical variate analysis (CVA) are among the methods commonly employed to deal with the collinearity problem in the process industry (Marjanovic *et al.*, 2006; Mobaraki and Hemmateenejad, 2011; Lin and Jørgensen, 2011; Shao *et al.*, 2012). The basic idea behind such methods is to project process variables into a lower number of orthogonal latent variables (Lin *et al.*, 2007). However, the over-fitting phenomenon may occur if the number of identification data points is small relative to the number of variables (Huopaniemi *et al.*, 2009). As pointed out by Yu (2012), conventional PCA and PLS models also become ill-suited for non-linear processes with non-Gaussian disturbances.

The Bayesian solution to collinearity problem is to incorporate subjective and/or objective prior information in order to alleviate the weak identification data problem (Leamer, 1973; Western and Jackman, 1994). In general, there are two different levels at which expert knowledge may be included in handling issues caused by collinear process measurements (MacKay, 2002; Gelman and Hil, 2007):

1. **Parameter estimation**: Since Bayesian methods treat parameters as random variables, subjective and/or objective process information can be used to describe informative prior distributions for model parameters. For instance, Nounou and Bakshi (2004) proposed Bayesian latent variable regression (BLVR) as a

new approach for linear process modeling that can handle collinear variables. Temperature measurements at different trays of a distillation column are used to develop an inferential sensor providing real-time estimates of ethanol content of the distillate stream. Prior knowledge about regression parameters and measured variables is incorporated through BLVR method to handle highly collinear input data. It is noteworthy that conventional regularization techniques such as ridge regression and mixed-estimation can be viewed as special cases of Bayesian parameter estimation dealing with collinear data (Belsley and Welsch, 2004; Marco and Gutierrez-Galvez, 2012).

2. **Model selection**: Expert knowledge about influential process variables and functional relationships of causally related variables can be represented by prior distributions to eliminate redundant collinear variables or determine a set of plausible model structures (Lambers *et al.*, 2006; de Vocht *et al.*, 2012). For instance, Prívara *et al.* (2010) presented a Bayesian algorithm to incorporate prior information about the model structure, such as static gain and input-output feed-through, into subspace identification of multiple-input multiple-output (MIMO) systems. The proposed algorithm was applied to handle multi-collinear measurements collected for identification of an industrial HVAC system.

Interested readers are referred to Dormann *et al.* (2013) for a comprehensive overview of the major aspects and topics related to the collinearity problem.

2.3 Model Identification

The design of inferential sensors finds its roots in process modeling. Therefore, proper identification of a representative process model is an essential prerequisite for development of an efficacious inferential sensor. Generally, the model identification procedure comprises two steps, model structure selection and model parameter estimation.

Having established the objective of the inferential sensor, one of the key problems in process identification is to find a suitable model structure that best describes the underlying system dynamics. Depending on the level of *a priori* knowledge of the process, two different philosophies may guide the choice of model structure (Ljung, 1999): 1. First principles analysis and 2. Process data analysis. Balance and constitutive equations often form the basis of quantitative and/or qualitative first principles analysis performed for inferential sensing purposes. In principle, a knowledge-driven model structure can be obtained based on formulating and solving a set of differential and algebraic balance equations at microscopic and macroscopic levels. Selection of appropriate system boundaries is essential for derivation of mass, momentum, and energy balance equations. Depending on the objective of first principles analysis, the system boundaries might enclose an entire plant, a few unit operations, or an equipment. For a defined system, the general balance equation can be stated

$$\begin{bmatrix} Accumulation \\ within\ System \end{bmatrix} = \begin{bmatrix} Input\ through \\ Boundaries\ of\ System \end{bmatrix} - \begin{bmatrix} Output\ through \\ Boundaries\ of\ System \end{bmatrix} + \begin{bmatrix} Generation \\ within\ System \end{bmatrix} - \begin{bmatrix} Consumption \\ within\ System \end{bmatrix} \tag{2.6}$$

Himmelblau and Riggs (2004) provide an introduction to the principles and techniques used in formulating and solving balance equations. Comprehensive coverage of principal chemical engineering unit operations including fluid mechanics, heat transfer, mass transfer, and equilibrium stages can also be found in McCabe *et al.* (2005); Bird *et al.* (2007); Perry and Green (2008).

The choice of knowledge-driven model structures for industrial processes depends on the complexity of the underlying physical systems and thus the availability of phenomenological knowledge of the involved unit operations. In the absence of any process knowledge, the task is to find a suitable model structure that is well supported by historical data. Therefore, a data-driven model structure is selected without investigating the internal mechanisms. In such cases, the main criteria to be considered in model selection are

simplicity, generality, and flexibility (Hangos and Cameron, 2001). A comprehensive overview of the wide variety of black-box structures (*e.g.* autoregressive models with exogenous inputs and state-space models) as well as an in-depth discussion of the general aspects of the choice of model structure (*e.g.* model order selection) can be found in Ljung (1999). Also, a general introduction to non-linear black-box structures including neural networks, radial basis networks, wavelet networks, hinging hyperplanes, and fuzzy models is provided by Sjöberg *et al.* (1995).

Having selected a representative structure, various classical or Bayesian estimation techniques can be applied in order to estimate the relevant model parameters. We follow with a comprehensive overview of the Bayesian methods that have been widely applied in identification of inferential sensors. The methods listed below are those most commonly suggested in the literature, though others can also be found.

2.3.1 Classical Bayesian Model Identification Methods

In this section, we introduce several classical Bayesian methods to build models for inferential sensors. In the case of data-driven methods, the emphasis is given on building simple models with a minimum number of influential variables. These compact models are easy to implement and maintain on-line.

One of the most important issues in the design of inferential sensors is the concept of model complexity. The more degrees of freedom are allowed in the model structure, the closer the model can approximate the identification data-set. On the other hand, too much flexibility might reduce the generalization performance of the developed inferential sensor when the process is operated under a wide range. Determination of a proper model structure (*e.g.* model order within a specified class) plays a key role in achieving a compromise between accuracy and complexity of the model.

2.3.1.1 Selection of Influential Variables

The problem of variable selection arises when complete process knowledge is not available. In such cases, influential variables are usually identified based on the limited process knowledge as well as the sensitivity analysis of operational and laboratory data. The main purpose of sensitivity analysis is two-fold. First, it is conducted to investigate how changes in the candidate input variables influence the query variable and, consequently, identify the most informative variables (Warne *et al.*, 2004). Second, it is performed to determine the degree of interaction between potential influential variables thereby preventing the undesired effects of collinearity in the process measurements (Chan *et al.*, 1997). General introduction to different aspects of variable selection as well as good reviews of the non-Bayesian methods of sensitivity analysis can be found in Saltelli *et al.* (2000); Guyon and Elisseeff (2003); Yuan and Lin (2006); Bhuyan (2011); Fujiwara *et al.* (2012).

In the context of inferential modeling for industrial applications, correlation analysis is the simplest and most widely used statistical approach to detect linear dependencies between input and query variables (Warne *et al.*, 2004; Komulainena *et al.*, 2004; Fortuna *et al.*, 2007). One can plot the color coded graph as illustrated in Figure 2.2 to identify variables with maximum correlation. A threshold value for the correlation coefficient can be chosen to decide on the number of variables to be selected.

The other commonly used variable screening techniques are step-wise methods such as forward selection and backward elimination (Wang *et al.*, 2006; Smits *et al.*, 2006; Fuchs and Maria, 2007; Wang *et al.*, 2010*a*). As a general implementation procedure of such methods, the following steps are iteratively performed to evaluate the significance of all candidate variables. In each stage, one variable is first added to (or removed from) the existing inferential model. The prediction performance of the revised model is then assessed on a suitable validation data-set to check whether or not the model's predictive capability has been improved. The classical Bayesian step-wise method involves evaluating

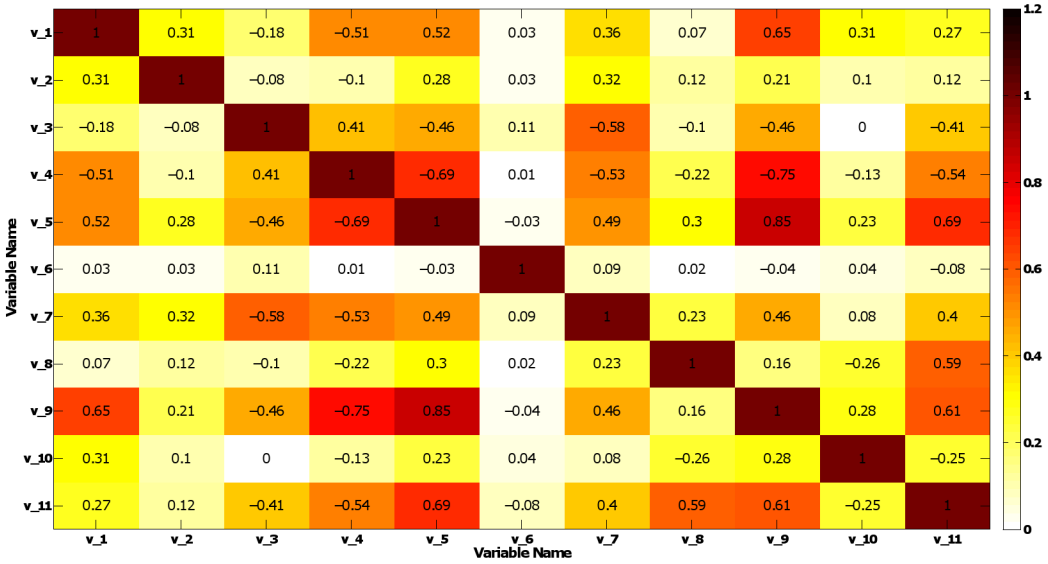


Figure 2.2: Color coded graph for correlation analysis

the Bayesian information criterion (BIC), also known as Schwarz information criterion (SIC), over a set of candidate models (Schwarz, 1978). The best approximating model is the one with minimum value of BIC or equivalently the one with highest posterior probability. In general, the BIC is defined as

$$BIC = -2\log\left(\mathcal{L}(\hat{\Theta}|\mathcal{D})\right) + K\log(T) \tag{2.7}$$

where K is the number of parameters, T is the number of observations, and $\mathcal{L}(\hat{\Theta}|\mathcal{D})$ is the likelihood of the estimated model parameters given the identification data. It is noteworthy that the likelihood term tends to decrease as more parameters are added to the model (*i.e.* K increases). A number of alternative information criteria also exist. These include: Akaike's information criterion (AIC), Takeuchi's information criterion (TIC), second-order information criterion (AIC_c), and quasi-likelihood information criterion(QAIC), among others. All these criteria are intended to minimize the prediction error of the model while penalizing the number of freely estimated parameters in order to identify a model that is both parsimonious and accurate (Burnham and Anderson, 2002; Lütkepohl, 2006; Shittu and Asemota, 2009). For neural network modeling with small identification data-sets,

Ingrassia and Morlini (2005) proposed to modify the number of degrees of freedom to be used in BIC. The authors used the modified BIC to select influential variables from an identification data-set concerning a vibration severity chart for centrifugal pumps in an ethylene system. Kaneko and Funatsu (2012) used different information criteria, including AIC and BIC, to identify important process variables affecting the operation of a distillation column at Mizushima plant of Mitsubishi Chemical Corporation. Also, the efficiency of different variable selection methods was compared by evaluating the accuracy and complexity of the resulting models.

The variance-based methods form an important class of probabilistic sensitivity analysis approaches in which the relative importance of each candidate input variable is quantified in terms of the resulting reduction in the output variance (Saltelli *et al.*, 2000; Lind and Ljung, 2005, 2008). Oakley and O'Hagan (2004) presented a unifying Bayesian framework for estimating various sensitivity measures. Their proposed framework provides a link between the sample-based regression measures and variance-based sensitivity analyses. Dufour *et al.* (2005) presented a neural network-based strategy for detection of feedstock variations in a continuous pulp digester. The authors studied sensitivity of the network outputs to the typical manipulated variables using the variance-based sensitivity analysis. Gonzagaa *et al.* (2009) developed an inferential sensor to provide on-line estimates of the viscosity of Polyethylene Terephthalate (PET). The authors used sensitivity analysis to select a proper set of process variables considering their degree of correlation with the polymer viscosity.

In many Bayesian variable selection approaches the problem is transformed into separating non-zero regression coefficients $\theta_j \neq 0$ from zero regression coefficients $\theta_j = 0$ (O'Hara and Sillanpää, 2009; Frühwirth-Schnatter and Wagner, 2011). First, a binary indicator variable, $I_j \in \{0, 1\}$, is associated with each coefficient. Next, a mixture of Gaussian prior distribution is specified for θ_j such that $p(\theta_j|I_j) = (1 - I_j)\mathcal{N}(0, \sigma_{0j}^2) + I_j\mathcal{N}(0, \rho\sigma_{1j}^2)$, where $\sigma_{0j}^2 \ll \sigma_{1j}^2$. Finally, the posterior inclusion probability for each

candidate variable is evaluated based on the estimated values of I_j and θ_j. If $I_j = 0$ and σ_{0j}^2 is close to zero, it can be concluded that θ_j is likely to be close to zero, *i.e.*, the corresponding process variable is practically not significant (Fahrmeir *et al.*, 2010). In the indicator model selection approach proposed by Kuo and Mallick (1998), an additional auxiliary variable is introduced such that $\theta_j = I_j\beta_j$. Each coefficient has a spike-and-slab prior distribution characterized by a spike at zero (*i.e.* $\sigma_{0j}^2 = 0$) and a flat slab equal to β_j elsewhere (*i.e.* $\sigma_{1j}^2 \to \infty$). Also, it is assumed that $p(\beta_j|I_j) = p(\beta_j)$. To identify a set of influential process variables, the posterior distribution of indicator variables is approximated by the means of the Markov chain Monte Carlo (Robert and Casella, 2004). George and McCulloch (1993) developed a stochastic search variable selection (SSVS) procedure utilizing a hierarchical Gaussian mixture model such that $\sigma_{1j}^2 = \rho\sigma_{0j}^2$, with $\rho \gg 1$. To obtain a computationally efficient sampling scheme, it is further assumed that $\sigma_1^2 = \sigma_{1j}^2$ and $\sigma_{0j}^2 = \sigma_0^2$. The posterior distribution is then evaluated through Gibbs sampling. Finally, influential process variables are selected according to their frequency of appearance in the sequences of Gibbs sample. The major drawback of SSVS is that ρ and σ_0^2 are assumed to be known and fixed. To address the aforementioned issue, Meuwissen and Goddard (2004) treated σ^2 as an uncertain hyperparameter to be estimated in an intermediate step.

Another Bayesian approach to select influential variables is to specify a continuous prior distribution on $\theta_j = 0$ that approximates the spike-and-slab shape without the inclusion of indicator variables, *i.e.* $\theta_j = \beta_j$ with $\beta_j \sim \mathcal{N}(0, \sigma_j^2)$. The main task is to define a prior distribution over hyperparameter σ_j^2 such that the values of β_j are shrunk towards zero if the corresponding process variable is practically not significant. Griffin and Brown (2010) discussed the interpretations of different prior distributions over σ_j and concluded that a wide range of shrinkage behavior can be specified through a gamma distribution. A special case of a gamma prior distribution would result in a Bayesian formulation of

Lasso (Tibshirani, 1996) which is a very popular classical method of variable selection. As discussed by O'Hara and Sillanpää (2009), the methodologies of Bayesian variable selection generally possess the following properties:

1. Subjective prior probabilities of variable inclusion can be incorporated to set the required degree of sparseness.

2. Model tuning parameters can be included by specifying data-based prior distributions. Hyperparameters of prior distributions can also be estimated in an intermediate step of hierarchical variable selection approaches.

4. Posterior variable inclusion probability can be evaluated through marginalization over different models.

A general introduction to various Bayesian variable selection methods as well as a comprehensive review of proper prior distributions are given by George and McCulloch (1997); Oakley and O'Hagan (2004); O'Hara and Sillanpää (2009); Frühwirth-Schnatter and Wagner (2011).

Despite many advantages of classical Bayesian variable selection methods, they have not been widely explored for inferential modeling practices in the process industry. One of the successful applications of Bayesian variable selection methods is reported by Ge and Song (2010). This paper introduced a Bayesian regularization method to effectively determine dimensionality of latent variables within a probabilistic PCA framework for multi-mode process monitoring. The authors treated the latent variable dimensionality as a model complexity problem, which can be handled using a Bayesian variable selection method.

2.3.1.2 Delayed Measurements

In many industrial plants, some process variables affect quality variables only after some time-delays. Therefore, another important aspect to be taken into consideration is that of

estimation of time-delays in the process variables. In order to select optimal time-delay of each variable, one can treat lagged process variables as independent variables and apply the same techniques introduced in Section 2.3.1.1. For instance, Gonzagaa *et al.* (2009) used sensitivity analysis to simultaneously select influential process variables and their time-delays for on-line prediction of the viscosity of Polyethylene Terephthalate (PET). Knowledge of plant operation (*e.g.* residence time inside a separation vessel, reaction time in a batch reactor, and etc.) can be included to determine possible upper bounds on time-delays.

2.3.1.3 Parameter Estimation

Once an appropriate model structure has been selected, Bayesian estimation techniques can be employed in order to estimate unknown model parameters, Θ. Since Bayesian methods treat parameters as random variables, subjective and/or objective process information can be used to specify informative prior distributions for model parameters. The importance of this point should be emphasized when estimating the parameters of complex model structures chosen on the basis of first-principles analysis. Such complex models contain a large number of highly correlated parameters that need to be estimated from operational or experimental identification data (Chu *et al.*, 2009). For inferential sensor applications the focus is on identifying a model providing accurate prediction rather than estimating each physical parameter. Although the true values of the model parameters are not known, the available process knowledge can often be translated into proper prior distributions of the parameters. This would prevent the over-fitting phenomenon commonly encountered in estimating the parameters of complex knowledge-driven model structures.

In classical Bayesian parameter estimation techniques ML estimates are penalized by prior knowledge. That is, the posterior distribution of model parameters, $p(\Theta|\mathcal{D})$, is

maximized to obtain a vector of single-point MAP estimates:

$$\begin{aligned}\Theta^{\text{MP}} &= \underset{\Theta}{\operatorname{argmax}}\, p(\Theta|\mathcal{D}) \\ &= \underset{\Theta}{\operatorname{argmin}} \Big[-\log \mathcal{L}(\Theta|\mathcal{D}) - \log p(\Theta|\mathcal{D}) \Big] \end{aligned} \tag{2.8}$$

Equation 2.8 suggests that classical Bayesian parameter estimation bears a close resemblance to regularized parameter estimation (Fahrmeir *et al.*, 2010). It is important to note that the above solution is not fully Bayesian because the entire *a posteriori* information obtained for parameters is represented by point estimates. Yet, classical Bayesian estimation techniques are still of interest in some industrial applications. For instance, in on-line implementation of inferential sensors, it might not be computationally feasible to integrate over the distribution of model parameters. Gunawan *et al.* (2003) proposed a systematic MAP estimation approach combining *a priori* information with after-anneal boron secondary ion mass spectroscopy (SIMS) profiles to obtain estimates of transient enhanced diffusion parameters. In Nounou and Bakshi (2004), temperature measurements at different trays of a distillation column were used to develop an inferential sensor providing real-time estimates of ethanol content of the distillate stream. The authors incorporated prior knowledge about model parameters in a Bayesian framework in order to obtain their MAP estimates. Fujiwara *et al.* (2005) used Bayesian estimation to compute the MAP parameter estimates of multiple inferential models developed for advanced control of pharmaceutical crystallization processes.

2.3.2 Full Bayesian Model Identification

The full Bayesian approach to model identification consists of two main steps. The first step concerns learning the model structure, $\mathcal{H}$, while the second step focuses on estimating the relevant model parameters, Θ.

Bayesian model comparison provides a probabilistic approach to rank alternative models without the introduction of *ad hoc* penalty terms (MacKay, 1992). In the light of

identification data-set, $\mathcal{D}$, the posterior probability of each model, $\mathcal{H}_i$, is evaluated as follows:

$$p(\mathcal{H}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{H}_i)p(\mathcal{H}_i)}{\sum_i p(\mathcal{D}|\mathcal{H}_i)p(\mathcal{H}_i)} \tag{2.9}$$

where $p(\mathcal{H}_i)$ represents the prior over $\mathcal{H}_i$ and $p(\mathcal{D}|\mathcal{H}_i)$ is the likelihood obtained by integrating over the $\mathcal{H}_i$'s parameter space:

$$p(\mathcal{D}|\mathcal{H}_i) = \int_{\Theta} p(\mathcal{D}|\Theta, \mathcal{H}_i)p(\Theta|\mathcal{H}_i)d\Theta \tag{2.10}$$

MacKay (2002) showed that under certain conditions the likelihood can be approximated by the height of the peak of the integrand $p(\mathcal{D}|\mathcal{H}_i, \Theta)p(\mathcal{H}_i|\Theta)$ times its width, $\sigma_{\Theta|\mathcal{D}}$:

$$p(\mathcal{D}|\mathcal{H}_i) \simeq p(\mathcal{D}|\Theta^{\text{MP}}, \mathcal{H}_i)p(\Theta^{\text{MP}}|\mathcal{H}_i)\sigma_{\Theta|\mathcal{D}} \tag{2.11}$$

where $p(\mathcal{D}|\Theta^{\text{MP}}, \mathcal{H}_i)$ is the best-fit likelihood that $\mathcal{H}_i$ can achieve and $p(\Theta^{\text{MP}}|\mathcal{H}_i)\sigma_{\Theta|\mathcal{D}}$ is known as **Occam factor**. Complexity of the model is then automatically penalized by the magnitude of the Occam factor. As pointed out by MacKay (2002), ideal Bayesian predictions do not involve model selection; rather, predictions are made by summing over all the alternative models, weighted by their probabilities.

Having selected a model structure, the posterior distribution of the parameters can thus be computed as

$$p(\Theta|\mathcal{D}, \mathcal{H}_i) = \frac{p(\mathcal{D}|\Theta, \mathcal{H}_i)p(\Theta|\mathcal{H}_i)}{p(\mathcal{D}|\mathcal{H}_i)} \tag{2.12}$$

As mentioned previously, the full Bayesian parameter estimation results in posterior distributions over parameters to reveal the level of uncertainty of the estimated values (Khatibisepehr and Huang, 2008).

Recently, there has been a growing interest in the application of full Bayesian model identification for the development of inferential sensors. Yan *et al.* (2004) developed an inferential sensor for real-time estimation of the freezing point of light diesel oil produced in a distillation column. The underlying inferential model was identified within a Bayesian

evidence framework first proposed by (MacKay, 1992). Beck and Yuen (2004) presented a Bayesian framework for selecting the most plausible class of models for a structural or mechanical system within some specified set of model classes. Bermak and Belhouari (2006) developed a gas classification framework based on Bayesian model identification as well as principal components analysis. For real-time monitoring of dynamic non-linear processes, Khawaja (2010) proposed a Bayesian failure prognosis scheme. The author applied the method to develop a Bayesian framework for detection and identification of cracks in the blades of a turbine high-power compressor disk.

2.3.3 Bayesian Interpretation of Classical Identification Methods

Many classical identification methods can be formulated within a Bayesian framework. Bishop and Tipping (2003); Tipping (2004) provided an overview of Bayesian formulation of the classical regression and classification problems. MacKay (1995) presented a Bayesian interpretation of neural network modeling. Kwok (2000); Suykens *et al.* (2002) derived a probabilistic formulation of the least squares support vector machine (LS-SVM) within a hierarchical Bayesian evidence framework. Nounou *et al.* (2002) developed a Bayesian PCA (BPCA) algorithm to integrate modeling and feature extraction by simultaneously solving parameter estimation and data reconciliation optimization problems.

2.3.4 Multi-model Inferential Sensors

Inferential model structures can be characterized as static and dynamic models; to develop a dynamic model the temporal dimension is added to the otherwise static model. Since temporal data arises in various areas of engineering, many industrial processes need to be naturally modeled as dynamic systems in order to express their behavior over time. In such cases, an inferential model structure should reflect both the static and dynamic characteristics of the process. Multi-model structures form an important class of model

structures that have been extensively adopted to represent time-varying dynamic behavior of industrial processes (Kim *et al.*, 2005; Li *et al.*, 2009; Domlan *et al.*, 2011). Multi-model inferential sensors typically describe both the continuous dynamic behavior and the transitions between discrete modes. The continuous dynamics is typically associated with the physical phenomena involved, while the discrete dynamics may come from switching controllers, inherent non-linearities in the system, different operating conditions, or any other external discrete events influencing the process under investigation. A general introduction to the identification of multi-modal processes, a discussion of the main issues connected with multi-modal system modeling, and an overview of the related literature are given in Paoletti *et al.* (2007); Lauer (2008). Readers are further directed to Murray-Smith and Johansen (1997); an early edition of the progress of work in the area of multi-model approach.

The multi-model paradigm has attracted increasing attention in the process control community due to its many potential industrial applications. The problem of multi-modal system modeling has been considered widely and to date several approaches have been proposed, such as the algebraic procedure (Vidal *et al.*, 2003), the clustering-based procedure (Ferrari-Trecate *et al.*, 2003), the EM-based procedure (Jin and Huang, 2010), and the bounded-error procedure (Bemporad *et al.*, 2005). In the recursive identification procedure implemented in these approaches, operating space is first partitioned into a finite number of non-overlapping regions. The regions are either defined a priori or estimated along with different sub-models. The operating regions and sub-models can be identified simultaneously by minimizing a suitable objective function. If the performance of the identified model is not satisfactory, the identification procedure is repeated with new sub-models and/or regions. The operating regions and sub-models can also be identified recursively. First, identification data is attributed to relevant regions based on descriptive classification criteria; the identification data-set is divided into multiple exclusive sub-

sets. Next, standard identification techniques are applied to develop sub-models that best describe the associated regions; the identified sub-models would be well supported by the corresponding identification data sub-sets.

A Bayesian identification procedure was proposed by Juloski *et al.* (2005) for piecewise autoregressive exogenous (PWARX) models and was extended by Juloski and Weiland (2006) for piecewise output error (PWOE) models. First, each attribute is classified to the mode with the highest probability by sequential processing of the identification data points. Next, Bayesian parameter estimation is performed to identify each sub-model from the corresponding data. A limitation of the described procedure is that the operating space is partitioned into a finite number of linearly separable regions, *i.e.*, at each time instant only one mode is active. If the identification data is not linearly separable or if the relevant residuals are comparable the violating attributes are excluded from analysis. In industrial applications, however, the operating modes are often overlapped or have non-linear boundaries in continuous unit operations. Moreover, the classification rule only relies on evaluating the residuals obtained from each sub-model. Thus, the available information about the process operation cannot be fully incorporated in the identification procedure. Khatibisepehr and Huang (2012) proposed a Bayesian procedure in order to accommodate the overlapping regions and facilitate the inclusion of prior knowledge about the operating conditions. The authors applied their proposed method to developing an adaptive multi-model inferential sensor for real-time monitoring of a key quality variable in an industrial oil sands processing unit. Li and Huang (2006) introduced a Bayesian-based model-set management method for selecting a statistically superior model-set for implementation of multi-model inferential sensors. The authors applied their method to design of a multi-model inferential sensors for automotive paint spray process where the thickness of the thin film on the vehicle surface should be precisely predicted. Suzdaleva and Nagy (2012) argued that combination of fault detection methods in the form of a hybrid

system allows us to exploit different types of knowledge and, consequently, leads to a more comprehensive intelligent supervisory control system. They further discussed that different types of information appearing in an on-line diagnostic system can be processed via combination of algorithms subject to probability distributions. Based on this argument, the authors proposed a decomposed version of Bayesian filtering specialized for hybrid dynamic systems with normal and discrete multinomial states and observations.

2.4 Model Validation

Model validation is the phase required to evaluate the performance of the identified inferential models. The validation criteria are chosen based on the intended applications of an inferential sensor. If the required criteria are not satisfied the inferential sensor design procedure should be reconsidered through close examination of each development step.

Off-line model validation usually comprises two steps, namely, **self-validation** and **cross-validation**. Self-validation determines the adequacy of fit by evaluating the prediction performance of the inferential model on the identification data. However, adequacy of model fit does not reliably ascertain the performance of the developed inferential sensor, *i.e.*, satisfactory prediction capability on the identification data does not guarantee generalization to other data-sets. Cross-validation assesses the generalization capability by evaluating the prediction performance of the identified inferential model on an independent data-set that has not been used for the model identification. Therefore, cross-validation plays an important role in preventing over-fitting the identification data.

Depending on the amount of available data, efficacy of the developed inferential sensor can be assessed on different data subsets. When the available data-set is sufficiently large, two independent subsets are constructed for cross-validation:

1. **Validation data-set**: The validation data-set is used to tune the identified inferential model and thus is indirectly involved in the identification procedure. Since such

data-sets are intended to guide the development of inferential sensors, satisfactory performance on the validation data might still be biased.

2. **Test data-set**: The prediction performance of the fully developed inferential sensor is evaluated on a test data-set consisting of completely independent data; test data is used neither in identification nor in tuning steps.

Often, accuracy of identification and reliability of validation procedures are sensitive to size of the corresponding data-sets. The required size of the historical data varies with application. In fact, whether the available data size should be small or large can be determined by several factors such as complexity of inferential model structure and extent of prior process knowledge. When the original data-set is relatively small, a single division of the available data into identification, validation, and test subsets is not feasible. Repeated partitioning and resampling of the available data are the main strategies adopted to overcome the limitations imposed by small data-sets (Ye, 2003). The techniques commonly applied in the inferential sensor applications include leave-one-out cross-validation (Wang *et al.*, 2010*a*), k-fold cross-validation (Kadlec, 2009; Chitralekha, 2011), and bootstrap resampling (Braga-Neto and Dougherty, 2004; Bolf *et al.*, 2009). In the k-fold cross-validation, the data are randomly divided into k equal partitions, k-1 of which are used for identification and the remaining one used for testing. This process is repeated until all the partitions are eventually used for both model identification and validation. The leave-one-out cross-validation procedure corresponds to a special case of k-fold cross-validation, in which k equals the number of data points. Applying the bootstrap resampling, an identification data-set is constructed by randomly sampling with replacement from the equally-likely available data. Given an original data-set of size n, the probability that a data point will not appear in the bootstrap identification data-set is $(1 - 1/n)^n \approx e^{-1} \approx 0.368$ (Efron and Tibshirani, 1993). Consequently, a fraction of the original data points will likely not be used in the identification phase and, thus, can be

reserved for validation purposes.

2.4.1 Performance Evaluation Criteria

Generally, the major purpose of model validation is to evaluate the accuracy and reliability of the developed inferential sensor. Accuracy is the level of agreement between the predicted and reference values, while reliability is the degree to which the prediction errors vary. Evaluating the performance of an inferential sensor amounts to analyzing the characteristics of prediction errors, which are also referred to as residuals. Some of the graphical and numerical methods are briefly described next.

2.4.1.1 Graphical Techniques

Common graphical techniques used in analysis of residuals include, but are not limited to, the following:

- **Scatter plot of predicted values versus target values**: The ideal case would be for all the data points to lie on the identity line ($y = x$), indicating perfect agreement between the predicted and target values.

- **Run-sequence plot of predicted and target values**: The time trend of the predicted and target values are plotted together to visually assess the accuracy and reliability of the inferential model.

- **Histogram of residuals**: The probability distribution of residuals is used to verify the assumptions made in the identification process about the error distribution.

- **Residual lag plot**: The lag plot of the residuals indicates whether or not the prediction errors are independent from their past values. Ideally, the auto-covariance function of residuals is a pulse function *i.e.* the auto-covariance is zero for all lags τ except for $\tau = 0$ (Shumway and Stoffer, 2000).

- **Residuals versus input variables**: The residuals are plotted versus the input variables (or independent variables) to verify that the prediction errors contain no information about system dynamics. Mathematically, the cross-covariance between residuals and each input variable should be zero (Shumway and Stoffer, 2000).

- **Residuals run-sequence plot**: The run-sequence plot of residuals is analyzed to search for any identifiable anomaly or pattern (*e.g.* drift in the process) in the prediction errors.

Comprehensive reviews of graphical techniques can be found in Chambers *et al.* (1983) and NIST/SEMATECH (2011).

2.4.1.2 Performance Measures

To provide a numerical basis for model assessment, a wide variety of performance measures have been proposed in the literature. Mean absolute error (MAE), standard deviation of errors (StdE), and mean squared error (MSE) are the most common statistical measures used for evaluating the performance of instruments and inferential sensors.

The MAE is a measure of accuracy defined as the average of absolute prediction errors:

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|\varepsilon_n| \tag{2.13}$$

where N is the number of observations and ε_i is the prediction error for the i_{th} observation. Since the MAE is an indication of the magnitude of prediction errors, small values correspond to accurate predictions with low bias and high precision (Pillai and Nair, 1997).

The StdE is a measure of reliability expressed through the variation of prediction errors:

$$\text{StdE} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\varepsilon_n - \bar{\varepsilon})^2} \tag{2.14}$$

where $\bar{\varepsilon}$ is the mean of error distribution. The StdE can be interpreted as the probability that prediction errors exceed the acceptable level of tolerance. Small values are obtained

from reliable inferential sensors, which exhibit consistent prediction performance, *i.e.*, the prediction errors are clustered closely around the mean (Pillai and Nair, 1997). It is noteworthy that if an inferential sensor provides accurate predictions, the mean of error distribution is around zero.

Finally, the MSE is used to indicate the overall prediction performance in terms of both accuracy and reliability:

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n^2} \tag{2.15}$$

Small values imply that the prediction errors are normally distributed around zero with a relatively small variance. This indicates that the inferential model produces accurate and reliable predictions (Pillai and Nair, 1997).

Other statistical performance measures are reviewed in Zhang (2004); Dawsona *et al.* (2007).

2.5 Dynamic Bayesian State Estimation

Temporal data arises in many areas of science and engineering. As a result, many real-world processes need to be naturally modeled as dynamic systems in order to describe their time-dependent behavior. State-space models are among the formulations extensively used to represent, and hence model, dynamic systems (Franklin *et al.*, 1998). Described as a generic state-space formulation, a dynamic Bayesian model (Murphy, 2002) can be derived to represent sequences of variables as they evolve over time. Let $\mathbf{x}_{1:t} \triangleq \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$, $\mathbf{u}_{1:t} \triangleq \{\mathbf{u}_1, \cdots, \mathbf{u}_t\}$, and $\mathbf{y}_{1:t} \triangleq \{\mathbf{y}_1, \cdots, \mathbf{y}_t\}$ denote the sequence of hidden state variables, input variables, and output variables, respectively. Suppose that at each time instant, t, the output variables, $\mathbf{y}_t \in \mathbb{R}^{n_y}$, have been generated from the hidden state variables, $\mathbf{x}_t \in \mathbb{R}^{n_x}$,

and the input variables, $\mathbf{u}_t \in \mathbb{R}^{n_u}$, such that

$$\mathbf{x}_{t+1} = f_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) \tag{2.16}$$

$$\mathbf{y}_t = h_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t) \tag{2.17}$$

where the uncorrelated random variables $\mathbf{v}_t$ and $\mathbf{w}_t$ denote the process noise and the measurement noise, respectively. The state transition function f_t describes the evolution of the state with time, while the measurement function h_t relates the noisy measurements to the state. Within a Bayesian framework, Equation 2.16 characterizes the state transition density function, $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$, and Equation 2.17 characterizes the likelihood of the measurements, $p(\mathbf{y}_t|\mathbf{x}_t)$. Note that it is assumed that the hidden state variables satisfy the first-order Markov condition, *i.e.* $p(\mathbf{x}_{t+1}|\mathbf{x}_1, \cdots, \mathbf{x}_t) = p(\mathbf{x}_{t+1}|\mathbf{x}_t)$. Since the inputs are always considered as known, it is convenient that all the PDFs of the form $p(.|., ., \mathbf{u}_{1:t})$ are denoted by $p(.|., .)$ without explicitly showing the dependence on the input.

Given all measurements up to and including time t, the main inference tasks performed in dynamic Bayesian models can be usually categorized as one of four possible types of query (Chen, 2003):

1. **Filtering**: The most common inference problem is to estimate the state at time instant t, *i.e.* $\hat{\mathbf{x}}(t|\mathbf{y}_{1:t}) \triangleq \hat{\mathbf{x}}_{t|t}$, which is amount to evaluating the posterior PDF of the state at time t, $p(\mathbf{x}_t|\mathbf{y}_{1:t})$.

2. **Smoothing**: It might be desired to estimate the states at previous time instants, *i.e.* $\hat{\mathbf{x}}(t-l|\mathbf{y}_{1:t}) \triangleq \hat{\mathbf{x}}_{t-l|t}$, by evaluating the posterior PDF of the state at time $t-l$, $p(\mathbf{x}_{t-l}|\mathbf{y}_{1:t})$, for $l \in [1, t-1]$.

3. **Prediction**: It is often required to predict the future states or outputs, *i.e.* $\hat{\mathbf{x}}(t+h|\mathbf{y}_{1:t}) \triangleq \hat{\mathbf{x}}_{t+h|t}$ and $\hat{y}(t+h|\mathbf{y}_{1:t}) \triangleq \hat{\mathbf{y}}_{t+h|t}$, which is amount to evaluating the prior PDF of the state at time $t+h$, $p(\mathbf{x}_{t+h}|\mathbf{y}_{1:t})$, or the prior PDF of the output at time $t+h$, $p(\mathbf{y}_{t+h}|\mathbf{y}_{1:t})$, for $h \geqslant 1$.

4. **Viterbi decoding**: Another interesting inference problem is to estimate the most likely sequence of states, *i.e.* $\hat{\mathbf{x}}(1:t|\mathbf{y}_{1:t}) \triangleq \hat{\mathbf{x}}_{1:t|t}$, by evaluating the probability density function $p(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})$.

In many industrial applications, the model identification problem should be formulated in an on-line inference scheme, for example, by augmenting states and parameters, in order to track the parameter and state trajectories using a sequence of noisy measurements. In such cases, a sequential Bayesian inference approach provides a rigorous framework for dynamic parameter and state estimation problems*. The basic idea behind the sequential Bayesian inference is to evaluate and propagate the probability density functions through an iterative process consisting of two steps:

1. **Prediction step**: The prior $p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})$ is evaluated to obtain $\hat{\mathbf{x}}_{t+1|t}$:

$$p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t}) = \int p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t})d\mathbf{x}_t \tag{2.18}$$

   where $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$ is defined by the state transition function $f_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t)$.

2. **Update step**: Once $\mathbf{y}_{t+1}$ becomes available, the posterior $p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t+1})$ is evaluated to obtain $\hat{\mathbf{x}}_{t+1|t+1}$:

$$p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t+1}) = \frac{p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1})p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})}{p(\mathbf{y}_{t+1}|\mathbf{y}_{1:t})} \tag{2.19}$$

   where $p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1})$ is defined by the measurement function $h_{t+1}(\mathbf{x}_{t+1}, \mathbf{u}_{t+1}, \mathbf{v}_{t+1})$, which essentially determines the measurement noise model.

Figure 2.3 illustrates the described sequential procedure of Bayesian inference.

In Chen (2003); Simon (2006), Bayesian filtering theory has been thoroughly discussed and different Bayesian filtering techniques have been comprehensively reviewed with emphasis on non-linear and non-Gaussian scenarios. Also, a review of recent developments

*It is well-known that the parameter estimation problem can be formulated as a state estimation problem.

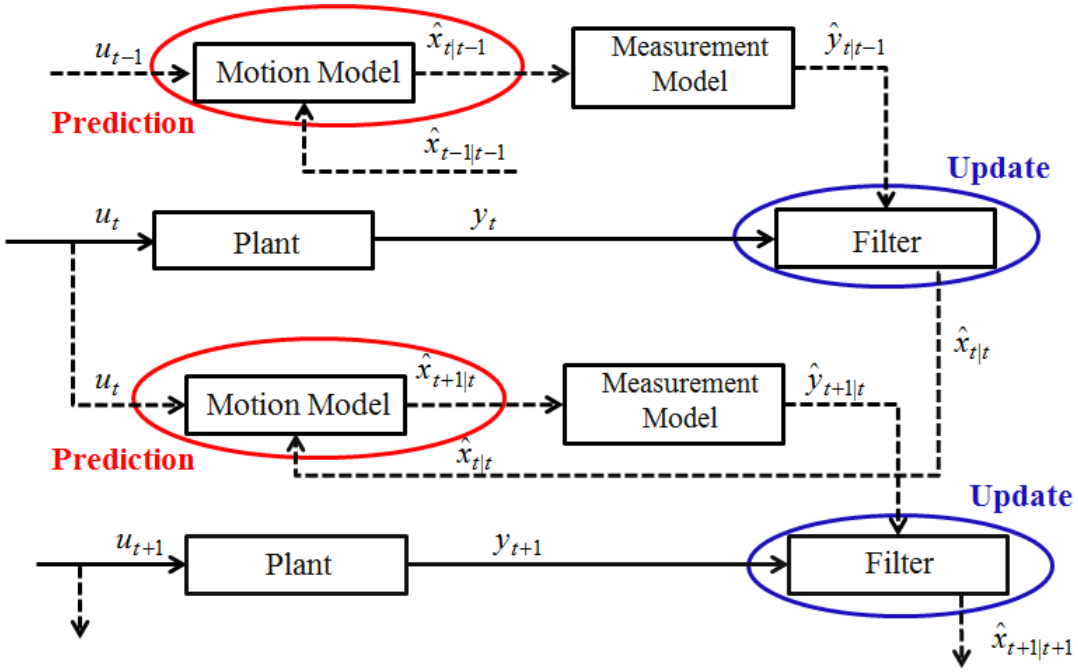


Figure 2.3: Sequential Bayesian inference

in the area of non-linear state estimators from a Bayesian perspective is given by Patwardhan *et al.* (2012). We follow with a brief overview of the theory and application of the most commonly used dynamic Bayesian state estimation techniques. Rather than being exhaustive, this section provides a general description of on-line state estimation in the context of inferential modeling.

2.5.1 Kalman-based Filters

Kalman filter (Kalman, 1960) can be viewed as a particular case of sequential Bayesian inference under the linear Gaussian assumptions. The computationally efficient nature of implementation of this filter makes it a very popular sequential Bayesian inference algorithm. Because of the efficiency of the Kalman filter in dynamic data processing, various Kalman-based fault detection and identification strategies have been proposed to develop a fault-tolerant control scheme (Prakash *et al.*, 2002; Villez *et al.*, 2011). For monitoring and controlling the operation of polymerization reactors, Freire and Giudici

(2004) derived an inferential model for joint estimation of the rate of heat generation and the overall heat transfer coefficient. The authors used a Kalman-based observer to estimate these two time-varying parameters from temperature measurements. For advanced monitoring of a biomass pre-treatment process, Prunescu *et al.* (2012) developed an inferential model capturing the environmental temperature differences inside a pressurized thermal reactor. The authors pointed out that it was hard to properly model the energy loss due to the open end of the reactor. Therefore, a Kalman filter was added to account for any missing dynamics.

Extended Kalman filter (EKF) (Sorenson, 1985) is an analytical approximation method used when the underlying process and/or measurement equations are non-linear. First, the non-linear state and measurement equations are linearized using first-order Taylor's series expansion. Next, the Kalman filter is applied to the linearized model of the non-linear equations. Therefore, the EKF enables us to apply Kalman filter structure to non-linear Gaussian systems. However, analytical computation of the state and output matrices involves evaluating the Jacobian of the non-linear models around the filtered and predicted values of the states at the previous time instant. Moreover, application of EKF would result in biased estimates for systems with significant non-linearity. As mentioned previously, the EKF estimates are accurate up to first order. Yet, this filtering technique has been widely used for the design of inferential sensors. Hagenmeyer and Nohr (2008) designed a flatness-based two-degree-of-freedom control scheme for temperature control in semi-batch reactors. Thereby, an EKF approach was chosen to estimate the reaction heat and the overall heat transfer coefficient. Bosca and Fissore (2011) applied an EKF-based approach to developing an inferential scheme for monitoring of the primary drying phase of a lyophilization process. The authors first derived first-principles models based on energy and mass balance analysis. The EKF algorithm was then used to estimate model parameters describing heat transfer to the product and mass transfer from the sublimation

interface to the bulk. Moreover, linearized models were developed to provide real-time predictions of product temperature as well as duration of primary drying within the EKF framework. Combining first-principles analysis and the EKF technique, Nair *et al.* (2011) developed an inferential sensor to obtain real-time estimates of furnace gas temperature for a tangentially-fired furnace used in steam generators of thermal power plants.

Unscented Kalman filter (UKF) (Norgaard *et al.*, 2000; Julier and Uhlmann, 2004) provides a method to approximate the probability density functions rather than approximating the non-linear functions. The main advantage of UKF over EKF is that it does not require explicit calculation of Jacobian and Hessian matrices (Chen, 2003). Moreover, UKF can better handle heavy-tailed distributions and hence is more tailored for non-Gaussian scenarios. Owing to the deterministic sampling approach followed, UKF is capable of estimating the posterior means and covariances accurately up to third-order for Gaussian data and at least second-order for non-Gaussian data (Julier and Uhlmann, 2004; Patwardhan *et al.*, 2012). Since the number of required samples is of the same order as the system, however, implementation of UKF would become computationally expensive for high dimensional model. Qu and Hahn (2009) introduced a moving horizon estimation formulation for non-linear constrained processes in which the arrival cost was determined by UKF. In Wang *et al.* (2010*b*), reliability of on-line tracking of a penicillin-fed batch fermentation process was improved by combining simplified mechanistic dynamic models and support vector regression (SVR)-based measurement equations. The authors developed an unscented Kalman filter for on-line estimation of key state variables. Salahshoor *et al.* (2012) proposed a new method for implementation of carbon dioxide (CO_2) sequestration process in saline aquifers through which reservoir pressure would follow a desired profile. The authors formulated their pressure control methodology within a non-linear model predictive control (MPC) to determine a sequence of optimum CO_2 injection rates. The reservoir pressure was predicted using a neural network model that was recursively trained

using EKF and UKF algorithms. To identify potential faults in controlling liquid levels in a three-tank hydraulic system, Mirzaee and Salahshoor (2012) developed a UKF-based inferential framework by integrating multiple stochastic models into an interpretive fuzzy decision-making scheme. Miyabayashi *et al.* (2012) developed a knowledge-driven state estimation system in order to detect catalyst deterioration and predict product concentration for monitoring of catalyst-packed tubular micro-reactors. The authors adopted UKF and EKF as non-linear filters and showed that UKF gave better estimation performance than EKF due to high non-linearity of underlying processes.

2.5.2 Particle Filters

Particle filter (PF) (Gordon *et al.*, 1993; Doucet *et al.*, 2001; Ristic *et al.*, 2004) provides a comprehensive approach to estimate the probability density functions of non-linear and non-Gaussian systems without making any explicit assumptions. The basic idea is to approximate the posterior PDFs through a set of weighted random samples, also called *particles*. Ensemble Kalman filter (EnKF) (Evenson, 2003) is a combination of Kalman and particle filtering techniques. The major advantage of EnKF over UKF is a reduced computational cost for high dimensional model, as samples are generated randomly and not deterministically. The general particle filtering approach to perform sequential Bayesian inference is outlined below.

First, N samples of the initial state, $\{\mathbf{x}_0^{(n)}\}_{n=1}^N$, are randomly drawn from the initial prior PDF of the state, *i.e.* $\mathbf{x}_0^{(n)} \sim p(\mathbf{x}_0)$. At each time instant t, the particles are propagated through the state transition and updated by measurement functions. N samples of $\mathbf{x}_{t+1}$ at time $t+1$ are generated from the posterior particles of the state and state noise at time t, *i.e.* $\mathbf{x}_{t+1}^{(n)} \sim p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(n)})$.

Once a new measurement becomes available, the posterior PDF can be approximated as

$$p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t+1}) = \frac{1}{N}\sum_{n=1}^{N} w_{t+1|t+1}^{(n)} \delta(\mathbf{x}_{t+1} - \mathbf{x}_{t+1}^{(n)}) \tag{2.20}$$

where $w_{t+1|t+1}^{(n)}$ denotes the importance weight assigned to the i^{th} particles and is given by

$$\begin{aligned} w_{t+1|t+1}^{(n)} &\triangleq p(\mathbf{x}_{t+1} = \mathbf{x}_{t+1}^{(n)} | \mathbf{y}_{1:t+1}) \\ &= \frac{p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}^{(n)})}{\sum_{j=1}^{N} p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}^{(j)})} \end{aligned} \tag{2.21}$$

A common problem with the above procedure is that after a few iterations, all but a few particles will have negligible importance weights. One way to avoid this degeneracy phenomenon is to remove particles that have small weights and to concentrate on particles with large weights (Arulampalam *et al.*, 2002). The idea is to resample N particles from the set $\{\mathbf{x}_{t+1}^{(n)}\}_{n=1}^{N}$ according to the importance weights $\{w_{t+1|t+1}^{(n)}\}_{n=1}^{N}$.

Chen *et al.* (2005) developed a particle filtering framework for on-line state and parameter estimation of first-principles inferential models of a highly non-linear batch process. Gopaluni (2008) proposed an identification algorithm formulated within the framework of expectation-maximization (EM) algorithm for identification of non-linear state-space models from incomplete identification data-sets. The complete log-likelihood functions in the expectation step of EM algorithm are approximated using the particle filtering technique. Similarly, Deng and Huang (2012) presented an EM-based framework for identification of non-linear parameter varying systems, in which the density functions of the expectation step are approximated using particle filters and smoothers. Zhao *et al.* (2011) presented a particle filtering strategy for on-line estimation of glucose and biomass concentration in a penicillin fermentation process. The prediction performance of the identified model was further improved by imposing state constraints on prior particles through projection of the violated particles onto a valid region. Jampanaa *et al.* (2010) proposed an inferential framework for detection of Bitumen-froth and Middlings interface level in separation cells of an oil sands primary extraction plant. The authors combined an image processing method, known as edge detection, with state-space model-based particle filtering in order to develop a PF-based vision sensor. Shenoy *et al.* (2010) applied various filtering techniques such as EKF, UKF, and PF onto the data from a Methyl Methacrylate

(MMA) continuous stirred tank reactor (CSTR) for various scenarios of Gaussian and non-Gaussian state and measurement noise sequences as well as plant-model mismatch. The authors concluded that for highly non-linear chemical processes, the UKF and PF would exhibit superior performance over the EKF. Moreover, it has been pointed out that the estimation performance of UKF and PF would vary depending on the degree of non-linearity of system dynamics, state and measurement noise levels, and the degree of plant-model mismatch.

2.6 Model Implementation and Calibration

On-line performance verification is the final step of the inferential sensor design procedure. If the off-line performance of the developed inferential sensor is satisfactory, the inferential sensor should be further tested on-line to address possible implementation issues.

The accuracy of an inferential sensor is guaranteed only for a particular region in which the model has been identified. However, most of the industrial processes exhibit a certain form of time-variant behavior due to fouling and/or abrasion in the process equipment, variation in the quality of feed, changes in the weather, and so on. In order to detect abrupt changes and gradual drifts in the process operations, process monitoring and on-line adaptation is often integrated in the implementation procedure. Once a significant variation is detected, the inferential model should be adjusted to compensate for deviations from the off-line design conditions. Several on-line adaptation methods have been proposed in the literature on the basis of moving windows techniques, recursive adaptation techniques, and ensemble-based methods (Kadlec *et al.*, 2011). Notwithstanding such precautions, periodic maintenance procedures should be considered.

2.6.1 Recursive and Real-time Identification Methods

In the cases where the prior process knowledge is not available, an inferential model for the quality variable can be built using only the historical data. In this case, the interactions between the process variables and their effects on the quality variable are not known. Therefore, data-driven inferential models are developed only based on the data mining techniques. Hence, these models are subject to recursive update as and when a new reference measurement becomes available. The advantage of the recursive methods is twofold. First, recursive identification of data-driven models ensures high prediction performance of inferential sensors. Second, recursive updates of model parameter estimates ensures that the model describes true current behavior of the underlying process. The most commonly used recursive methods are recursive least squares (RLS) and recursive partial least squares (RPLS) algorithms. In these algorithms, the influence of the past data is discounted for by using a forgetting factor. RLS algorithm can also be interpreted as a form of Kalman filter. This equivalence sometimes is helpful in implementation of RLS algorithm. For further details about the recursive methods and procedures of tuning the forgetting factor the readers are referred to Ljung (1999); Dayal and MacGregor (1997), and the references therein.

Locally weighted regression (LWR) (Cleveland, 1979; Atkeson *et al.*, 1997), also known as **just-in-time modeling** (Zheng and Kimura, 2001), is another on-line adaptation technique with a long history of development. The general idea behind the LWR algorithm is to identify a local model in real-time by prioritizing the identification data-points. The search for the nearest neighbors is carried out from the historical data-set using a predefined notion of similarity. This approach can cope with abrupt and gradual changes in the process characteristics and operating conditions. Kano and Fujiwara (2013) have provided a good review of the recent theoretical developments and successful industrial applications of the LWR algorithm.

2.6.2 Local Adaptation Mechanisms

In some industrial applications, recursive update of all parameters may not be feasible or desirable. In such cases, local adaptation mechanisms are integrated in the implementation procedure to detect and handle potential unknown drifts of process operating conditions.

Consider an input-output representation of an adaptive inferential sensor expressed as

$$y_t = \alpha_t f(\mathbf{r}_t, \Theta) + \beta_t + \nu_t \tag{2.22}$$

where α_t and β_t respectively denote the scale factor and discrepancy term of the model at time instant t and ν_t is the noise term. There are a variety of update rules that can be specified to guide the adjustment of the scale factor and discrepancy term. In many industrial applications, the general form of an exponentially weighted moving average filter is employed to develop local adaptation mechanisms, such that

$$\alpha_{t+1} = \lambda\left(\frac{y_t^{Ref} - y_{t-1}^{Ref}}{f(r_t, \Theta) - f(\mathbf{r}_{t-1}, \Theta)}\right) + (1-\lambda)\alpha_t \tag{2.23}$$

$$\beta_{t+1} = \kappa\Big(y_t^{Ref} - \alpha_{t+1} f(\mathbf{r}_t, \Theta)\Big) + (1-\kappa)\beta_t \tag{2.24}$$

where λ and κ are the smoothing parameters, also known as forgetting factors. As illustrated in Figure 2.4, the bias is updated to reduce the prediction offset. On the other hand, the scale factor is updated to adjust the slope of the imaginary line passing through the predictions. The formulation described above allows for straightforward implementation of on-line adaptation mechanisms for industrial inferential sensors, though others may also be used (Khatibisepehr and Huang, 2012).

State-space representation of inferential sensors with local adaptation mechanism can be considered as a special case of information synthesis as will be discussed in the next section.

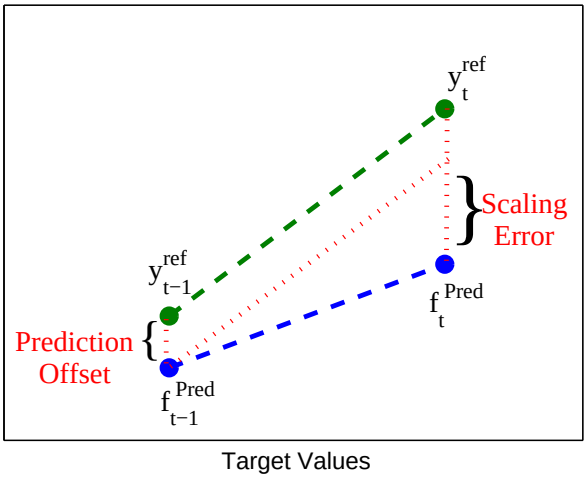


Figure 2.4: Inferential sensor calibration philosophy (Khatibisepehr and Huang, 2012)

2.6.3 Information Synthesis

State-space models are among the formulations extensively used to fuse the information gathered from several sources, such as physical and inferential sensors. An information synthesis problem is often posed as a filtering problem. In this way, the well-developed filtering techniques introduced in Section 2.5 can be adopted to solve this problem. In general, two different cases of information synthesis can be considered:

Case I. There is more than one instrumentation sensor and/or on-line analyzer along with the inferential model. However, each available sensor by itself may not be accurate. Therefore, it is required to incorporate all the information together in order to predict the query variable as accurately as possible.

Case II. The physical sensor is not available for some period due to limitations in the instrumentation. In such cases, the inferential sensor should provide real-time predictions of the query variable. Similarly, infrequent laboratory measurements are also used as another source of information. The inferential sensors are intended to provide real-time predictions of the query variable when the lab measurements are

not available. This case is the classic example of a multi-rate filtering problem.

2.6.3.1 Case I

Let $\mathbf{x}_t$ denote the true value of the query variable of interest, while $y_t^{\mathtt{i}}$ and $y_t^{\mathtt{m}}$ represent the measurements available from the physical sensor and inferential model, respectively. The state-space model can be formulated as follows:

$$\text{State equations:} \quad \mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{v}_t^{\mathtt{p}} \tag{2.25}$$

$$\alpha_{t+1} = \alpha_t + v_t^{\alpha} \tag{2.26}$$

$$\beta_{t+1} = \beta_t + v_t^{\beta} \tag{2.27}$$

$$\text{Measurement equations:} \quad y_t^{\mathtt{i}} = \mathbf{x}_t + w_t^{\mathtt{i}} \tag{2.28}$$

$$y_t^{\mathtt{m}} = \alpha_t \mathbf{x}_t + \beta_t + w_t^{\mathtt{m}} \tag{2.29}$$

Several terms in the proposed state-space model are discussed below:

1. **The process model (f):** Whenever the physical sensor is available and reliable, the sensor measurements can be considered as the true values of the query variable and then corresponding data can be used to build a dynamic model which would represent the true process behavior. Note that the data should be collected so that the sensor measurements are highly reliable and the data covers all possible operating conditions. If such data is not available or none of the physical sensors are reliable for building a model, one can use a simple random walk model in place of the identified model f as in the following equation:

   $$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{v}_t^{\mathtt{p}} \tag{2.30}$$

   Equation 2.30 simply reflects the fact that the query variable is varying

2. **Measurement noise ($w_t^{\mathtt{m}}$):** Tuning of measurement noise is a key issue in this information synthesis formulation. In many applications reliability of the

physical sensor measurements is specified with introduction of a performance index associated with the physical sensor. For instance, the higher the performance index, the better the accuracy of the sensor measurement. This performance index value can be exploited while tuning the measurement noise. In fact, measurement noise can be considered as a function of the performance index. This function is set such that the variance of noise is less when the performance index is higher and vice versa. A graphical representation of a typical function relating the performance index to the noise variance is shown in Figure 2.5.

3. **Process noise terms ($\mathbf{v}_t^{\mathrm{p}}$, v_t^{α}, v_t^{β}):** These terms represent process uncertainty and adapting parameter variation. As discussed in Section 2.6.2, the inferential sensor may drift away from the variable of interest over time and hence there is a need to develop adaptation mechanisms. In state-space representation of adaptive inferential sensors, Equations 2.26 and 2.27 are used to update the scaling factor, α_t, and the bias term, β_t, respectively. In this formulation, the noise variances are tuning parameters that affect adaptation speed and magnitude.

4. **Error in the inferential sensor prediction (w_t^i):** This noise or more precisely error term can be estimated using the historical data of the query variable as well as the predictions of the inferential model.

Note that with this set-up, we can estimate the query variable ($\mathbf{x}_t$) using standard filtering techniques like KF if f is linear and no scaling factor is considered, or otherwise EKF, UKF, or PF.

Shao *et al.* (2011) proposed a Bayesian framework, which facilitated inclusion of additional information in the form of prior knowledge and synthesis of multiple-source quality variable observations to derive a more accurate posterior distribution for the unknown state and parameters. To enhance the robustness of the proposed framework in the presence of abnormal data, the authors also developed a robust Bayesian fusion

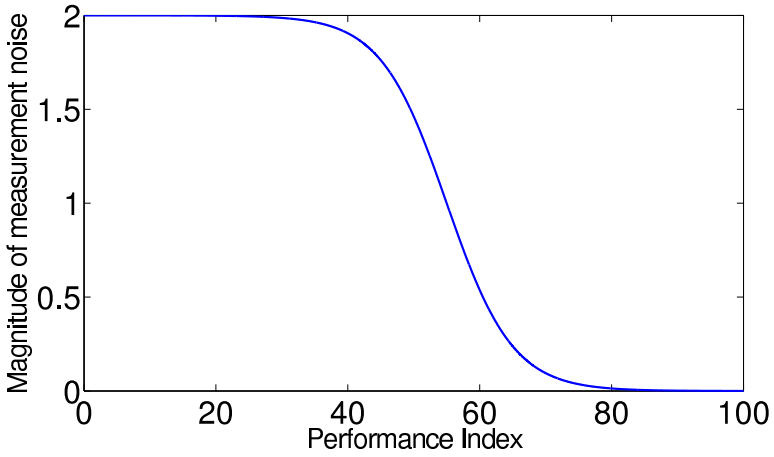


Figure 2.5: Performance index function

formulation with a time-varying measurement noise variance. In Shao *et al.* (2012), the Bayesian information synthesis approach of their previous was followed to fuse all the available information in order to obtain accurate and reliable predictions of bitumen froth quality in an oil sands natural froth lubricated transportation process.

2.6.3.2 Case II

A multi-rate state-space model would be developed to handle infrequently measured quality variables. This model can be described as follows:

$$\text{State equations:} \quad \mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{v}_t^{\mathtt{p}} \tag{2.31}$$

$$\alpha_{T+1} = \alpha_T + v_T^{\mathtt{a}} \tag{2.32}$$

$$\beta_{T+1} = \beta_T + v_T^{\mathtt{b}} \tag{2.33}$$

$$\text{Measurement equations:} \quad y_T^{\mathtt{i}} = \mathbf{x}_T + w_T^{\mathtt{i}} \tag{2.34}$$

$$y_t^{\mathtt{m}} = \alpha_t \mathbf{x}_t + \beta_t + w_t^{\mathtt{m}} \tag{2.35}$$

In this formulation, the slower sampling rate denoted by T corresponds to the slow-rate measurements of the query variable (*e.g.* laboratory data), while the faster sampling rate denoted by t corresponds to fast-rate measurements/predictions of the query variable. The

scaling factor and bias term are updated only when a reliable slow-rate measurement is available. In order to identify an inferential model, one needs to keep in mind the slow-rate sampling of the variable of interest. The presented filtering problem is then solved using multi-rate filtering techniques. Under multi-rate sampling conditions, Wu and Luo (2010) introduced a data fusion framework based on Kalman filter for on-line calibration of inferential sensors using infrequent laboratory measurements. The authors applied their method to the maintenance of an inferential sensor providing real-time estimates of the mixture quality in a blending system (*e.g.* pH neutralization system).

The concept of Bayesian information synthesis is not only used in process industry, but is also a popular tool for inferential modeling in other industries. For example, the Bayesian information synthesis helps in analyzing unstructured digital data in several forms. Autonomy (available at `http://www.autonomy.com/`) provides solution to such problems using innovative tools in Bayesian information synthesis.

2.6.4 Data Reconciliation and Gross Error Detection

Accuracy of measured process variables is a key requirement for successful implementation of inferential sensors. However, real-time measurements are subject to two types of errors: 1. random errors caused by imprecision of instruments and 2. systematic gross errors caused by instrument biases, malfunctioning of measuring devices, and significant heat or material loss (Narasimhan and Jordache, 2000; Romagnoli and Sánchez, 2000). The techniques used to improve the accuracy of measurements by reducing the effect of random errors are termed as data reconciliation methods. In order for data reconciliation to be effective, gross error detection methods are applied to identify and eliminate systematic errors. Generally, data reconciliation and gross error detection methods are intended to deal with instrument measurements that do not satisfy mass and energy balance constraints associated with the steady state or dynamic process operation (Tamhane and Mah, 1985).

As pointed out by Tamhane (1988), Bayesian methods provide a natural framework

for data reconciliation and gross error detection. Under certain restrictive assumptions, Tamhane (1988) presented a Bayesian scheme for detecting gross errors in chemical process data. For pseudo steady-state processes, Devanathan *et al.* (2005) proposed a Bayesian decision rule to detect mean shifts in process variables leading to an inference regarding the presence of multiple measurement biases. The authors discussed that the performance of their method for measurement bias identification would not be affected by the presence of leaks and cancelation of the effects of multiple biases in material balance. Gui *et al.* (2007) introduced a Bayesian framework for gross error detection that utilizes the available prior information on the unknown parameters of the mean shift model as well as the variance inflation model. Gonzalez *et al.* (2011) proposed a Bayesian approach for gross error detection allowing for the separation and estimation of measurement noise variance as well as process disturbance variance to gain more-informative estimates of gross errors. The proposed method focuses on estimating a model that is simultaneously consistent with mass balance equations and measurement noise covariance. Gonzalez *et al.* (2012) have developed a dynamic Bayesian methodology for real-time detection and quantification of instrument gross errors. This method can be considered as a type of switching Kalman filter through which future measurements are reconciled. In Gonzalez *et al.* (2011, 2012), the authors successfully applied their methods for on-site performance monitoring of weightometers in an oil sands slurry preparation plant, which could reduce costs of maintenance and aid in dealing with the unavoidable presence of systematic errors.

2.6.5 Monitoring of inferential Sensor Performance

In order to maintain the reliability of an inferential sensor, it is often necessary to track its on-line performance. However, designing a performance index and specifying a threshold are not straightforward. For each particular application, the historical process data and the prior physical knowledge should be exploited to identify the criteria that might affect on-line performance of the designed inferential sensor. Although much effort has been

devoted to dealing with several challenges associated with industrial applications, only a few publications provide a methodology for on-line performance assessment of inferential models. In Nomikos and MacGregor (1995), approximate confidence intervals have been developed to assess the accuracy of PLS predictions based on the traditional statistical properties. Kaneko *et al.* (2010) proposed a method to quantify the relationship between the applicability domains and accuracy of inferential sensor predictions. The authors discussed that a larger distance to the average as well as nearest neighbor of identification data would indicate a lower accuracy of prediction. Kaneko and Funatsu (2011) proposed an ensemble prediction method with time difference for inferential sensor development. In this work, the accuracy of the predictions is estimated using empirical models describing the relationship between the standard deviation of the multiple predicted values and the standard deviation of the prediction errors.

It is of great interest to thoroughly assess the quality of laboratory data as they will be trusted in on-line implementation of the inferential sensor to adjust the real-time predictions. Therefore, it is worthwhile to develop a procedure for reliability analysis of the laboratory data.

2.7 Concluding Remarks and Future Research Challenges

Real-time analysis of process quality variables constitutes an essential prerequisite for advanced monitoring and control of industrial processes. However, on-line acquisition of such quality variables is often restricted by inadequacy of measurement techniques, low reliability of measuring devices, and significant time-delays associated with laboratory analysis. Therefore, there has been a growing interest in the development of inferential sensors to provide frequent on-line estimates of quality variables on the basis of their correlation with real-time process measurements. In this paper, we provided a general

introduction to the main steps involved in design of industrial inferential sensors, discussed some of the challenging issues associated with development and implementation of inferential sensors, and presented an overview of the relevant Bayesian literature. Due to the demonstrated advantages of Bayesian methods as well as the increasing interest in their applications, the main focus of this paper was to introduce the potential of Bayesian methods for inferential modeling practices in the process industry. Adopting Bayesian methods bears several benefits.

1. Process knowledge can be easily incorporated in a Bayesian scheme by specifying proper prior distributions over model parameters, functional forms, and constraints (Tulleken, 1993). Such information obtained from first-principles is forced upon the data-driven models to obtain grey-box models.

2. The model identification problem can be rigorously formulated under a principled framework, which features fewer heuristic design choices. For instance, a Bayesian approach to modeling can naturally deal with complexity control to avoid over-fitting by integrating out the hyperparameters (Hutter, 2007). The significance of this advantage should be highlighted for estimation of the parameters of complex knowledge-driven model structures. In such applications, it is often required to estimate a large number of correlated parameters from scarce and noisy identification data, resulting in an ill-conditioned inferential model (Chu *et al.*, 2009).

3. Within a full Bayesian framework, the uncertainty in model parameters is characterized through posterior PDFs which give rise to a so called **predictive distribution**. Thus, probabilistic predictions are made by marginalizing over the parameters. This feature allows us to quantify the effects of model uncertainty on the reliability of predicted values.

4. General Bayesian learning techniques convert the identification problem into an

equivalent problem of computing expectation or evaluating an integral as opposed to solving a global optimization problem as in likelihood methods. This may be advantageous in many situations as solving a global non-convex optimization problem is avoided.

5. Incomplete data and non-Gaussian distributions can be handled naturally.

As indicated by existing research efforts, development and implementation of inferential sensors entail many challenges. Despite the increasing number of publications dealing with industrial applications, several issues remain open for future investigation. Some of the challenging issues that foreshadow interesting topics for future research are summarized below.

1. Although the problems of process data analysis and model identification are interconnected, most of the existing solutions are disconnected. It is desired to seek for a unified framework that simultaneously considers different aspects of data analysis and inferential modeling. There is potential in formulating the problems of interest as rigorous conditional probabilistic problems within a Bayesian framework.

2. In order to maintain the reliability of an inferential sensor, it is required to track its on-line performance. However, the main body of research in this area has been focused on exploiting advanced strategies for development of inferential sensors. Hence, it is of paramount importance to search for general criteria and techniques for on-line performance assessment of inferential models.

3. Maintenance of inferential sensors is another important topic to be further investigated. There have been several efforts to develop real-time and recursive identification methods as well as local adaptation mechanisms. Yet, proper maintenance of the identification data-set remains a challenging task. Theoretical and

practical developments are required to effectively assess the reliability of operational and laboratory measurements in real-time.

4. There is a growing realization that off-line operation assistance tools can play a significant role in improving plant-wide operations. The main research challenge is to develop information synthesis schemes that can coordinate processing of diverse forms of knowledge. Further research is imperative to effectively synthesize qualitative and quantitative information provided by operations personnel, inferential and physical sensors, laboratory analysis, and many other sources.

5. Long and uncertain time-delays in reference data (*e.g.* lab data) constitute one of the main practical problems in inferential sensor development. Samples are frequently collected from the operational field and the recorded sampling time can deviate significantly from the actual time. Laboratory analysis of each sample can be time consuming, thereby introducing a significant time-delay. Therefore, modeling, filtering, and information synthesis in the presence of long and uncertain time-delays are of great research interest in inferential sensor development.

6. Bias update has been common practice in inferential sensor applications. Advanced updating strategies, as reviewed in this Chapter, include the multi-rate information fusion method and the filtering method. However, due to the slow rate of sampling of lab data, these updates are often associated with an abrupt change of the prediction, introducing undesired bumps to the inferential sensor predictions. Optimal synthesis of multi-rate data is another topic of interest.

7. The main objective of developing inferential sensors is for industrial applications. Almost for certain, all inferential sensors have to be converted to the distributed control system (DCS) language. Initial inferential sensor development is most likely completed in an advanced program environment such as MATLAB. Reliable

implementation of the advanced program into DCS program or simplification of the advanced algorithm for implementation in DCS constitute a significant practical challenge.

Bibliography

Agarwal, D. (2006). Detecting anomalies in cross-classified streams: a Bayesian approach. *Knowledge and Information Systems* **11**(1), 29–44.

Allison, P. D. (2002). *Missing data*. 136 ed.. Sage Publications. Thousand Oaks, USA.

Arulampalam, M. S., S. Maskell, N. Gordon and T. Clapp (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing* **50**(2), 174–188.

Atkeson, C. G., A. W. Moore and S. Schaal (1997). Locally weighted learning. *Artificial Intelligence Review* **11**, 11–73.

Beck, J. L. and K. Yuen (2004). Model selection using response measurements: Bayesian probabilistic approach. *Journal Of Engineering Mechanics* **130**(2), 192–203.

Belsley, D. A. and E. Kuhand R. E. Welsch (2004). *Robust Statistics*. second ed.. John Wiley & Sons. New Jersey, USA.

Bemporad, A., A. Garulli, S. Paoletti and A. Vicino (2005). A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control* **50**(10), 1567–1580.

Ben-Gal, I. (2010). Outlier detection. In: *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers* (O. Maimon and

L. Rockach, Eds.). second ed.. pp. 117–130. Springer Science+Business Media. New York, USA.

Bermak, A. and S. B. Belhouari (2006). Bayesian learning using gaussian process for gas identification. *IEEE Transcation on Instrumentation And Measurement* **55**(3), 787–792.

Bhuyan, M. (2011). *Intelligent Instrumentation: Principals and Applications*. first ed.. Taylor & Francis Group. New York, USA.

Bird, R. B., W. E. Stewart and E. N. Lightfoot (2007). *Transport Phenomena*. second ed.. John Wiley & Sons. New York, USA.

Bishop, C. (1994). Novelty detection and neural network validation. *IEE Proceedings - Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks* **141**(4), 217–222.

Bishop, C. M. and M. E. Tipping (2003). Bayesian regression and classifications. In: *Advances in Learning Theory: Methods, Models and Applications* (J. A. K. Suykens, I. Horvath, S. Basu, C. Micchelli and J. Vandewalle, Eds.). Vol. 190 of *NATO Science Series III: Computer and Systems Sciences*. pp. 267–285. IOS Press. Amsterdam, Netherlands.

Bohlin, T. and S. F. Graebe (1995). Issues in nonlinear stochastic grey box identification. *International Journal of Adaptive Control and Signal Processing* **9**(6), 465–490.

Bolf, N., G. Galinec and M. Ivandić (2009). Soft sensors for kerosene properties estimation and control in crude distillation unit. *Chemical and Biochemical Engineering Quarterly* **23**(3), 11–17.

Bosca, S. and D. Fissore (2011). Design and validation of an innovative soft-sensor for pharmaceuticals freeze-drying monitoring. *Chemical Engineering Science* **66**(21), 5127–5136.

Braga-Neto, U. M. and E. R. Dougherty (2004). Is cross-validation valid for small-sample microarray classification?. *Chemical and Biochemical Engineering Quarterly* **20**(3), 374–380.

Burnham, K. P. and D. R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. second ed.. Springer-Verlag. New York, USA.

Chambers, J., W. Cleveland, B. Kleiner and P. Tukey (1983). *Graphical Methods for Data Analysis*. first ed.. Wadsworth International Group. Belmont, USA.

Chan, K., A. Saltelli and S. Tarantola (1997). Sensitivity analysis of model output: Variance-based methods make the difference. In: *Proceedings of the* 29^{th} *Winter Simulation Conference (WSC)*. Atlanta, USA. pp. 261–268.

Chandola, V., A. Banerjee and V. Kumar (2009). Anomaly detection : A survey. *ACM Computing Surveys* **41**(3), 124–129.

Chen, T., J. Morris and E. Martin (2005). Particle filters for state and parameter estimation in batch processes. *Journal of Process Control* **15**(6), 665–673.

Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. *Statistics* **182**(1), 1–69.

Chitralekha, S. B. (2011). Computational Tools for Soft Sensing and State Estimation. PhD thesis. University of Alberta. Edmonton, Canada.

Chu, Y., Z. Huang and J. Hahn (2009). Improving prediction capabilities of complex dynamic models via parameter selection and estimation. *Chemical Engineering Science* **64**(19), 4178–4185.

Cinar, A., S. J. Parulekar, C. $\ddot{U}$ndey and G. Birol (2003). *Batch Fermentation: Modeling, Monitoring, and Control*. first ed.. Marcel Dekker. New York, USA.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**(368), 829–836.

Dadhe, K., V. Roßmann, K. Durmus and S. Engell (2001). Neural networks as a tool for gray box modelling in reactive distillation. In: *Computational Intelligence. Theory and Applications* (B. Reusch, Ed.). Vol. 2206 of *Lecture Notes in Computer Sciences*. pp. 576–588. Springer-Verlag. Berlin.

Das, K. and J. Schneider (2007). Detecting anomalous records in categorical datasets. In: *Proceedings of the* 13^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press. San Jose, USA. pp. 220–229.

Davies, L. and U. Gather (1993). The identification of multiple outliers. *Journal of the American Statistical Association* **88**(423), 782–792.

Dawsona, C. W., R. J. Abrahartb and L. M. Seec (2007). Hydrotest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. *Environmental Modelling and Software* **22**(7), 10341052.

Dayal, B. S. and J. F. MacGregor (1997). Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *Journal of Process Control* **7**(3), 169–179.

de Vocht, F., N. Cherry and J. Wakefield (2012). A Bayesian mixture modeling approach for assessing the effects of correlated exposures in case-control studies. *Journal of Exposure Science and Environmental Epidemiology* **22**(4), 352–360.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). Maximum likelihood estimation from

incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**(1), 1–38.

Deng, J. and B. Huang (2012). Identification of nonlinear parameter varying systems with missing output data. *AIChE Journal* **58**(11), 3454–3467.

Devanathan, S., S. B. Vardeman and Sr. D. K. Rollins (2005). Likelihood and Bayesian methods for accurate identification of measurement biases in pseudo steady-state processes. *Chemical Engineering Research and Design* **83**(12), 1391–1398.

Domlan, E., B. Huang, F. Xu and A. Espejo (2011). A decoupled multiple model approach for soft sensors design. *Control Engineering Practice* **11**(2), 126–134.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leit ao, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell and S. Lautenbach (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**(1), 27–46.

Doucet, A., N. de Freitas and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. first ed.. Springer-Verlag. New York, USA.

Draper, N. R. and H. Smith (2003). *Applied Regression Analysis*. third ed.. John Wiley & Sons. Singapore.

Dufour, P., S. Bhartiya, P. S. Dhurjati and F. J. Doyle III (2005). Neural network-based software sensor: Training set design and application to a continuous pulp digester. *Control Engineering Practice* **13**(2), 135–143.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. first ed.. Chapman & Hall/CRC. London, UK.

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In: *Proceedings of the* 17^{th} *International Conference on Machine Learning*. Morgan Kaufmann. San Francisco, USA. pp. 255–262.

Evenson, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics* **53**(4), 343367.

Fahrmeir, L., T. Kneib and S. Konrath (2010). Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing and predictor selection. *Statistics And Computing* **20**(2), 203–219.

Ferrari-Trecate, G., M. Muselli, D. Liberati and M. Morari (2003). A clustering technique for the identification of piecewise affine systems. *Automatica* **39**(2), 205–217.

Fortuna, L., S. Graziani, A. Rizzo and M. G. Xibilia (2007). *Soft Sensors for Monitoring and Control of Industrial Processes*. first ed.. Springer-Verlag. London, UK.

Franklin, G. F., J. D. Powell and M. L. Workman (1998). *Digital Control of Dynamic Systems*. third ed.. Prentice-Hall. Englewood Cliffs, USA.

Freire, F. B. and R. Giudici (2004). Temperature oscillation calorimetry by means of a Kalman-like observer: The joint estimation of Qr and UA in a stirred tank polymerization reactor. *Macromolecular Symposia* **206**(1), 15–28.

Friedman, Y. Z., E. A. Neto and C. R. Porfirio (2002). First-principles distillation inference models for product quality prediction. *Hydrocarbon Processing* **81**(2), 54–58.

Frühwirth-Schnatter, S. and H. Wagner (2011). Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data. In: *Bayesian Statistics* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, Eds.). Vol. 9 of *Oxford Science Publications*. pp. 165–200. Oxford University Press. Oxford, uK.

Fuchs, J. J. and S. Maria (2007). A new appraoch to variable selection using the TLS approach. *IEEE Transactions on Signal Processing* **55**(1), 10–19.

Fujiwara, K., H. Sawada and M. Kano (2012). Input variable selection for PLS modeling using nearest correlation spectral clustering. *Chemometrics and Intelligent Laboratory Systems* **118**(1), 109–119.

Fujiwara, M., Z. K. Nagy, J. W. Chew and R. D. Braatz (2005). First-principles and direct design approaches for the control of pharmaceutical crystallization. *Journal of Process Control* **15**(5), 493–504.

Ge, Z. and Z. Song (2010). Mixture Bayesian regularization method of PPCA for multimode process monitoring. *AIChE Journal* **56**(11), 2838–2849.

Ge, Z. and Z. Song (2011). Semisupervised Bayesian method for soft sensor modeling with unlabeled data samples. *AIChE Journal* **57**(8), 2109–2119.

Gelman, A. and J. Hil (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. first ed.. Cambridge University Press. New York, USA.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (2003). *Bayesian Data Analysis*. second ed.. Chapman & Hall/CRC. Boca Raton, USA.

George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**(423), 881–889.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Ghosh-Dastidar, B. and J. L. Schafer (2006). Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics* **22**(3), 487–506.

Gonzagaa, J. C. B., L. A. C. Meleirob, C. Kianga and R. Maciel Filho (2009). ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process. *Computers and Chemical Engineering* **33**(1), 43–49.

Gonzalez, G. D. (1999). Soft sensors for processing plants. In: *Proceedings of the Second International Conference on Intelligent Processing and Manufacturing of Materials (IPMM)*. IEEE. pp. 59–69.

Gonzalez, R., B. Huang, F. Xu and A. Espejo (2011). Estimation of instrument variance and bias using Bayesian methods. *Industrial and Engineering Chemistry Research* **50**(10), 6229–6239.

Gonzalez, R., B. Huang, F. Xu and A. Espejo (2012). Dynamic Bayesian approach to gross error detection and compensation with application toward an oil sands process. *Chemical Engineering Science* **67**(1), 44–56.

Gopaluni, R. B. (2008). A particle filter approach to identification of nonlinear processes under missing observations. *The Canadian Journal of Chemical Engineering* **86**(6), 1081–1092.

Gordon, N. J., D. J. Salmond and A. F. M. Smith (1993). A novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing* **140**(2), 107–113.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology* **60**(1), 549–576.

Grantham, S. D. and L. H. Ungar (1990). A first principles approach to automated troubleshooting of chemical plants. *Computers and Chemical Engineering* **14**(7), 783–798.

Greene, W. H. (2007). *Econometric Analysis*. sixth ed.. Prentice-Hall Inc.. New Jersey, USA.

Griffin, J. E. and P. J. Brown (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**(1), 171–188.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* **11**(1), 1–21.

Gui, Q., Y. Gong, G. Li and B. Li (2007). A Bayesian approach to the detection of gross errors based on posterior probability. *Journal Of Geodesy* **81**(10), 651–659.

Gunawan, R., M. Y. L. Jung, E. G. Seebauer and R. D. Braatz (2003). Maximum *A Posteriori* estimation of transient enhanced diffusion energetics. *AIChE Journal* **49**(8), 2114–2123.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**(7-8), 1157–1182.

Hagenmeyer, V. and M. Nohr (2008). Flatness-based two-degree-of-freedom control of industrial semi-batch reactors using a new observation model for an extended Kalman filter approach. *International Journal of Control* **81**(3), 428–438.

Hangos, K. M. and I.T. Cameron (2001). *Process Modelling and Model Analysis*. first ed.. Academic Press. San Diego, USA.

Himmelblau, D. M. and J. B. Riggs (2004). *Basic Principles And Calculations In Chemical Engineering*. seventh ed.. Prentice Hall. Upper Saddle River, USA.

Hodge, V. J. and J. Austin (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review* **22**(2), 85–126.

Huopaniemi, I., T. Suvitaival, J. Nikkilä, M. Orešič and S. Kaski (2009). Two-way analysis of high-dimensional collinear data. *Data Mining And Knowledge Discovery* **19**(2), 261–276.

Hutter, M. (2007). Exact Bayesian regression of piecewise constant functions. *Bayesian Analysis* **2**(4), 635–664.

Imtiaz, S. A. and S. L. Shah (2008). Treatment of missing values in process data analysis. *The Canadian Journal of Chemical Engineering* **86**(5), 838–858.

Ingrassia, S. and I. Morlini (2005). Neural network modeling for small datasets. *Technometrics* **47**(3), 297–311.

Jampanaa, P., S. L. Shah and R. Kadali (2010). Computer vision based interface level control in separation cells. *Control Engineering Practice* **18**(4), 349–357.

Jiaa, R., Z. Maoa, Y. Changa and L. Zhao (2011). Soft-sensor for copper extraction process in cobalt hydrometallurgy based on adaptive hybrid model. *Chemical Engineering Research and Design* **89**(6), 722–728.

Jin, X. and B. Huang (2010). Robust identification of piecewise/switching Autoregressive eXogenous process. *AIChE Journal* **56**(7), 1829–1844.

Jin, X., S. Wang, B. Huang and F. Forbes (2012). Multiple model based LPV soft sensor development with irregular/missing process output measurement. *Control Engineering Practice* **20**(2), 165–172.

Julier, S. and J. K. Uhlmann (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE* **92**(3), 401–422.

Juloski, A. and S. Weiland (2006). A Bayesian approach to the identification of piecewise linear output error models. In: *Proceedings of the* 14^{th} *IFAC Symposium on System*

Identification (B. Ninness and H. Hjalmarsson, Eds.). number 14. IFAC. Newcastle, Australia. pp. 374–379.

Juloski, A. L., S. Weiland and W. P. M. H. Heemels (2005). A Bayesian approach to identification of hybrid systems. *IEEE Transactions on Automatic Control* **50**(10), 1520–1533.

Kadlec, P. (2009). On Robust and Adaptive Soft Sensors. PhD thesis. Bournemouth University. Bournemouth, UK.

Kadlec, P., B. Gabrys and S. Strandt (2009). Data-driven soft sensors in the process industry. *Computers and Chemical Engineering* **33**(4), 795–814.

Kadlec, P., R. Grbić and B. Gabrys (2011). Review of adaptation mechanisms for data-driven soft sensors. *Computers and Chemical Engineering* **35**(1), 1–24.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering* **82**(Series D), 35–45.

Kaneko, H. and K. Funatsu (2011). Improvement and estimation of prediction accuracy of soft sensor models based on time difference. In: *Modern Approaches in Applied Intelligence* (K. G. Mehrotra, C. K. Mohan, J. C. Oh, P. K. Varshney and M. Ali, Eds.). Vol. 6703 of *Lecture Notes in Computer Science*. pp. 115–124. Springer-Verlag. Berlin, Germany.

Kaneko, H. and K. Funatsu (2012). A new process variable and dynamics selection method based on a genetic algorithm-based wavelength selection method. *AIChE Journal* **58**(6), 1829–1840.

Kaneko, H., M. Arakawa and K. Funatsu (2010). Applicability domains and accuracy of prediction of soft sensor models. *AIChE Journal* **57**(6), 1506–1513.

Kano, M. and K. Fujiwara (2013). Virtual sensing technology in process industries: Trends and challenges revealed by recent industrial applications. *Journal of Chemical Engineering of Japan* **46**(1), 1–17.

Kano, M. and Y. Nakagawa (2008). Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry. *Computers and Chemical Engineering* **32**(1-2), 12–24.

Khatibisepehr, S. and B. Huang (2008). Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Industrial and Engineering Chemistry Research* **47**(22), 8713–8723.

Khatibisepehr, S. and B. Huang (2012). A Bayesian approach to design of adaptive multi-model inferential sensors with application in oil sand industry. *Journal of Process Control* **22**(10), 1913–1929.

Khatibisepehr, S. and B. Huang (2013). A Bayesian approach to robust process identification with ARX models. *AIChE Journal* **59**(3), 845–859.

Khatibisepehr, S., B. Huang and S. Khare (2013). Design of inferential sensors in the process industry: A review of bayesian methods. *Journal of Process Control* p. in press.

Khawaja, T. S. (2010). A Bayesian Least Squares Support Vector Machines Based Framework For Fault Diagnosis And Failure Prognosis. PhD thesis. Georgia Institute of Technology. Georgia, USA.

Kim, M., Y. H. Lee, I. S. Han and C. Han (2005). Clustering-based hybrid soft sensor for an industrial polypropylene process with grade changeover operation. *Industrial and Engineering Chemistry Research* **44**(2), 334–342.

Komulainena, T., M. Souranderb and SL Jämsä-Jounelaa (2004). An online application

of dynamic PlS to a dearomatization process. *Computers and Chemical Engineering* **28**(2), 2611–2619.

Korb, K. B. and A. E. Nicholson (2004). *Bayesian Artificial Intelligence*. first ed.. Chapman & Hall/CRC. London, UK.

Kuo, L. and B. Mallick (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics* **60**, 65–81.

Kwok, J. T. (2000). The evidence framework applied to support vector machines. *IEEE Transactions on Neural Network* **11**(5), 1162–1173.

Lalor, G. C. and C. Zhang (2001). Multivariate outlier detection and remediation in geochemical databases. *The Science of The Total Environment* **281**(1-3), 99–109.

Lambers, J. Hille Ris, B. Aukema, J. Diez, M. Evans and A. Latimer (2006). Effects of global change on inflorescence production: A Bayesian hierarchical analysis. In: *Hierarchical Modelling For The Environmental Sciences - Statistical Methods And Applications* (J. S. Clark and A. Gelfand, Eds.). pp. 59–73. Oxford University Press. Cary, USA.

Lauer, F. (2008). From Support Vector Machines to Hybrid System Identification. PhD thesis. Nancy University. Lorraine, France.

Leamer, E. E. (1973). Multicollinearity: A Bayesian interpretation. *The Review of Economics and Statistics* **55**(3), 371–380.

Lee, J., B. Kang and S. Kang (2011). Integrating independent component analysis and local outlier factor for plant-wide process monitoring. *Journal of Process Control* **21**(7), 1519–1528.

Li, J. and Y. Huang (2006). Bayesian-based on-line applicability evaluation of neural network models in modeling automotive paint spray operations. *Computers and Chemical Engineering* **30**, 1392–1399.

Li, X. L., H. Su and J. Chu (2009). Multiple model soft sensor based on affinity propagation, Gaussian process and Bayesian committee machine. *Chinese Journal of Chemical Engineering* **17**(1), 95–99.

Lin, B. and S. B. Jørgensen (2011). Soft sensor design by multivariate fusion of image features and process measurements. *Journal of Process Control* **21**(4), 547–553.

Lin, B., B. Recke, J. K. H. Knudsen and S. B. Jørgensen (2007). A systematic approach for soft sensor development. *Computers and Chemical Engineering* **31**(5-6), 419–425.

Lin, T. H. (2010). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality and Quantity* **44**(2), 277–287.

Lind, I. and L. Ljung (2005). Regressor selection with the analysis of variance method. *Automatica* **41**(4), 693–700.

Lind, I. and L. Ljung (2008). Regressor and structure selection in NARX models using a structured anova approach. *Automatica* **44**(2), 383–395.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis With Missing Data*. second ed.. John Wiley & Sons. New York, USA.

Liu, X., K. Li, M. McAfee, B. K. Nguyen and G. M. McNally (2012). Dynamic grey-box modelling for online monitoring of extrusion viscosity. *Polymer Engineering and Science* **52**(6), 1332–1341.

Ljung, L. (1999). *System Identification - Theory For the User*. second ed.. Prentice Hall. Upper Saddle River, USA.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*. second ed.. Springer-Verlag. New York, USA.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation* **4**(3), 415–447.

MacKay, D. J. C. (1995). Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* **6**(3), 469–505.

MacKay, D. J. C. (2002). *Information Theory, Inference, and Learning Algorithm*. first ed.. Cambridge University Press. New York, USA.

Marco, S. and A. Gutierrez-Galvez (2012). Signal and data processing for machine olfaction and chemical sensing: A review. *IEEE Sensors Journal* **12**(11), 3189–3214.

Marjanovic, O., B. Lennox, D. Sandoz, K. Smith and M. Crofts (2006). Real-time monitoring of an industrial batch process. *Computers and Chemical Engineering* **30**(10-12), 1476–1481.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* **12**(3), 591–612.

Mason, A., S. Richardson, I. Plewis and N. Best (2012). Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. *Journal of Official Statistics* **28**(2), 279–302.

McCabe, W. L., J. C. Smith and P. Harriott (2005). *Unit Operations Of Chemical Engineering*. seventh ed.. McGraw-Hill. Boston, USA.

Meuwissen, T. H. E. and M. E. Goddard (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution* **36**(3), 261–279.

Mirzaee, A. and K. Salahshoor (2012). Fault diagnosis and accommodation of nonlinear systems based on multiple-model adaptive unscented Kalman filter and switched MPC and H-infinity loop-shaping controller. *Journal of Process Control* **22**(3), 626–634.

Miyabayashi, K., O. Tonomura, M. Kano and S. Hasebe (2012). Comparative study of state estimation of tubular microreactors using ukf and ekf. In: *Proceedings of the* 8^{th} *IFAC Symposium on Advanced Control of Chemical Processes*. IFAC. Singapore. pp. 513–518.

Mobaraki, N. and B. Hemmateenejad (2011). Structural characterization of carbonyl compounds by IR spectroscopy and chemometrics data analysis. *Chemometrics and Intelligent Laboratory Systems* **109**(2), 171–177.

Murphy, K. P. (2002). Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis. University of California. Berkely, USA.

Murray-Smith, R. and T. R. Johansen (1997). *Multiple Model Approaches to Modelling and Control*. first ed.. Taylor & Francis. London, UK.

Nair, A. T., T. K. Radhakrishnan, K. Srinivasan and S. R. Valsalam (2011). Kalman filter based state estimation of a thermal power plant. In: *Proceedings of the International Conference on Process Automation, Control and Computing (PACC)*. IEEE. Coimbatore, India. pp. 1–5.

Narasimhan, S. and C. Jordache (2000). *Data Reconciliation & Gross Error Detection: An Intelligent Use of Process Data*. first ed.. Gulf Publishing Company. Houston, USA.

NIST/SEMATECH (2011). *e-Handbook of Statistical Methods*. http://www.itl.nist.gov/div898/handbook/.

Nomikos, P. and J. F. MacGregor (1995). Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems* **30**(1), 97–108.

Norgaard, M., M. Poulsen and O. Ravn (2000). New developments in state estimation for nonlinear systems. *Automatica* **36**(11), 1627–1638.

Nounou, M. N. and B. R. Bakshi (2004). Process modeling by Bayesian latent variable regression. *AIChE Journal* **48**(8), 1775–1793.

Nounou, M. N., B. R. Bakshi, P. K. Goel and X. Shen (2002). Bayesian principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **16**(11), 576–595.

Oakley, J. E. and A. O'Hagan (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **66**(3), 751–769.

O'Hara, R. B. and M. J. Sillanpää (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis* **4**(1), 85–118.

Olanrewaju, M., B. Huang and A. Afacan (2010). Online composition estimation and experiment validation of distillation processes with switching dynamics. *Chemical Engineering Science* **65**(5), 1597–1608.

Osborne, J. W. (2008). *Best Practices in Quantitative Methods*. first ed.. Sage Publications. Thousand Oaks, USA.

Pani, A. K. and H. K. Mohanta (2011). A survey of data treatment techniques for soft sensor design. *Chemical Product and Process Modeling*.

Paoletti, S., A. L. Juloski, G. Ferrari-Trecate and R. Vidal (2007). Identification of hybrid systems: A tutorial. *European Journal of Control* **13**(2-3), 242–260.

Patwardhan, S. C., S. Narasimhan, P. Jagadeesan, B. Gopaluni and S. L. Shah (2012).

Nonlinear Bayesian state estimation: A review of recent developments. *Control Engineering Practice* **20**(10), 933–953.

Penny, K. I. and I. T. Jolliffe (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. *Journal of the Royal Statistical Society: Series D (The Statistician)* **50**(3), 295–307.

Perry, R. H. and D. W. Green (2008). *Perry's Chemical Engineers' Handbook*. eighth ed.. McGraw-Hill. New York, USA.

Pillai, K. and V. S. S. Nair (1997). A model for software development effort and cost estimation. *IEEE Transactions on Software Engineering* **23**(8), 485–497.

Prakash, J., S. C. Patwardhan and S. Narasimhan (2002). A supervisory approach to fault-tolerant control of linear multivariable systems. *Industrial and Engineering Chemistry Research* **41**(9), 2270–2281.

Prasad, V., M. Schley, L. P. Russo and B. Wayne Bequette (2002). Product property and production rate control of styrene polymerization. *Journal of Process Control* **12**(3), 353–372.

Prívara, S., J. Cigler, Z. Váňa, L. Ferkl and M. Šebek (2010). Subspace identification of poorly excited industrial systems. In: *Proceedings of the* 49^{th} *IEEE Conference on Decision and Control*. Atlanta, USA. pp. 4405–4410.

Prunescu, R. M., M. Blanke, J. M. Jensen and G. Sin (2012). Temperature modelling of the biomass pretreatment process. In: *Proceedings of the* 17^{th} *Nordic Process Control Workshop* (J. K. Huusom J. B. Jørgensen and G. Sin, Eds.). Technical University of Denmark. Kogens Lyngby, Denmark. pp. 8–17.

Qi, F. and B. Huang (2011). Bayesian methods for control loop diagnosis in presence of temporal dependent evidences. *Automatica* **47**(7), 1349–356.

Qi, F., B. Huang and E. C. Tamayo (2010). A Bayesian approach for control loop diagnosis with missing data. *AIChE Journal* **56**(1), 179–195.

Qu, C. C. and J. Hahn (2009). Computation of arrival cost for moving horizon estimation via unscented Kalman filtering. *Journal of Process Control* **19**(2), 358–363.

Raghavan, H., A. K. Tangirala, R. B. Gopaluni and S. L. Shah (2006). Identification of chemical processes with irregular output sampling. *Control Engineering Practice* **14**(5), 467–480.

Ristic, B., S. Arulampalam and N. Gordon (2004). *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. first ed.. Artech House. Boston, USA.

Ritter, G. and M. T. Gallegos (1997). Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters* **18**(6), 525–539.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods*. second ed.. Springer-Verlag. New York, USA.

Roberts, S. (1999). Novelty detection using extreme value statistics. *Proceedings of the IEE Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks* **146**(3), 124–129.

Romagnoli, J. A. and M. C. Sánchez (2000). *Data Processing and Reconciliation for Chemical Process Operations*. first ed.. Academic Press. San Diego, USA.

Rousseeuw, P. and A. Leroy (1996). *Robust Regression and Outlier Detection*. second ed.. Wiley. New York, USA.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**(3), 581–592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. first ed.. John Wiley & Sons. New York, USA.

Salahshoor, K., M. H. Hajisalehi and M. Haghighat Sefat (2012). Nonlinear model identification and adaptive control of Co2 sequestration process in saline aquifers using artificial neural networks. *Applied Soft Computing* **12**(11), 3379–3389.

Saltelli, A., K. Chan and M. Scott (2000). *Sensitivity Analysis*. first ed.. John Wiley & Sons. New York, USA.

Schafer, J. L. and J. W. Graham (2002). Missing data: Our view of the state of the art. *Psychological Methods* **7**(2), 147–177.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**(2), 461–464.

Shao, X., B. Huang, J. M. Lee, F. Xu and A. Espejo (2011). Bayesian method for multirate data synthesis and model calibration. *AIChE Journal* **57**(6), 1514–1525.

Shao, X., F. Xu, B. Huang and A. Espejo (2012). Estimation of bitumen froth quality using Bayesian information synthesis: An application to froth transportation process. *The Canadian Journal of Chemical Engineering* **90**(6), 1393–1399.

Shenoy, A. V., V. Prasad and S. L. Shah (2010). Comparison of unconstrained nonlinear state estimation techniques on a mma polymer reactor. In: *Proceedings of the* 9^{th} *International Symposium on Dynamics and Control of Process Systems (DYCOPS)*. IFAC. Leuven, Belgium. pp. 145–150.

Shittu, O. I. and M. J. Asemota (2009). Comparison of criteria for estimating the order of autoregressive process: A monte carlo approach. *European Journal of Scientific Research* **30**(3), 409–416.

Shumway, R. H. and D. S. Stoffer (2000). *Time Series Analysis and Its Applications*. Springer. New York.

Simon, D. (2006). *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. first ed.. John Wiley & Sons. Hoboken, USA.

Sjöberg, J., Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.Y. Glorennec, H. Hjalmarsson and A. Juditsky (1995). Nonlinear black-box modeling in system identification: a unified overview. *Automatica* **31**(2), 1691–1724.

Smits, G., A. Kordon, E. Jordaan, C. Vladislavleva and M. Kotanchek (2006). Variable selection in industrial data sets using pareto genetic programming. In: *Genetic Programming Theory and Practice III* (T. Yu, R. Riolo and B. Worzel, Eds.). Vol. 9 of *Genetic Programming*. pp. 79–92. Springer. New York, USA.

Sorenson, H. W. (1985). *Kalman Filtering: Theory and Application*. 38 ed.. IEEE Press. New York, USA.

Suykens, J. A. K., T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle (2002). *Least Squares Support Vector Machines*. first ed.. World Scientific. River Edge, USA.

Suzdaleva, E. and I. Nagy (2012). Online soft sensor for hybrid systems with mixed continuous and discrete measurements. *Computers and Chemical Engineering* **36**, 294–300.

Tamhane, A. C. (1988). A Bayesian approach to gross error detection in chemical process data: Part i : Model development. *Chemometrics and Intelligent Laboratory Systems* **4**(1), 33–45.

Tamhane, A. C. and R. S. H. Mah (1985). Data reconcilliation and gross error detection in chemical process networks. *Technometrics* **27**(4), 409–422.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**(398), 528–540.

Tibshirani, R. (1996). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B: Methodological* **58**(1), 267–288.

Tipping, M. E. (2004). Bayesian inference: An introduction to principles and practice in machine learning. In: *Advanced Lectures on Machine Learning* (O. Bousquet, U. von Luxburg and G. Rätsch, Eds.). pp. 41–62. Springer.

Tulleken, H. J. A. F. (1993). Grey-box modelling and identification using physical knowledge and Bayesian techniques. *Automatica* **29**(2), 285–308.

Varbanov, A. (1998). Bayesian approach to outlier detection in multivariate normal samples and linear models. *Communications in Statistics - Theory and Methods* **27**(3), 547–557.

Vidal, R., S. Soatto, Y. Ma and S. Sastry (2003). An algebraic geometric approach to the identification of a class of linear hybrid systems. In: *Proceedings of the* 42^{nd} *IEEE Conference on Decision and Control*. number 1. IEEE. Maui, USA. pp. 167–172.

Villez, K., B. Srinivasan, R. Rengaswamy, S. Narasimhan and V. Venkatasubramanian (2011). Kalman-based strategies for fault detection and identification (FDI): Extensions and critical evaluation for a buffer tank system. *Computers and Chemical Engineering* **35**(5), 806–816.

Wang, D., J. Liu and R. Srinivasan (2010*a*). Data-driven soft sensor approach for quality prediction in a refining process. *IEEE Transactions On Industrial Informatics* **6**(1), 11–17.

Wang, D., R. Srinivasan, J. Liu, P. N. S. Guru and K. M. Leong (2006). Data driven soft sensor approach for quality prediction in a refinery process. In: *Proceedings of the* 4^{th}

IEEE International Conference on Industrial Informatics (INDIN). Singapore. pp. 230–235.

Wang, J., Z. Liqiang and Y. Tao (2010*b*). On-line estimation in fed-batch fermentation process using state space model and unscented Kalman filter. *Chinese Journal of Chemical Engineering* **18**(2), 258–264.

Warne, K., G. Prasad, S. Rezvani and L. Maguire (2004). Statistical and computational intelligence techniques for inferential model development: A comparative evaluation and a novel proposition for fusion. *Engineering Applications of Artificial Intelligence* **17**(8), 871–885.

Western, B. and S. Jackman (1994). Bayesian inference for comparative research. *American Political Science Review* **88**(2), 412–423.

Wu, Y. and X. Luo (2010). A novel calibration approach of soft sensor based on multirate data fusion technology. *Journal of Process Control* **20**(10), 1252–1260.

Yan, W., H. Shao and X. Wang (2004). Soft sensing modeling based on support vector machine and Bayesian model selection. *Computers and Chemical Engineering* **28**(8), 1489–1498.

Ye, N. (2003). *The Handbook of Data Mining*. first ed.. Lawrence Erlbaum Associates. Mahwah, USA.

Yu, J. (2012). A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. *Computers and Chemical Engineering* **41**(11), 134–144.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B: Methodological* **68**(1), 49–67.

Zeng, J. and C. Gao (2009). Improvement of identification of blast furnace ironmaking process by outlier detection and missing value imputation. *Journal of Process Control* **19**(9), 1519–1528.

Zhang, G. P. (2004). *Neural Networks in Business Forecasting*. first ed.. IGI Global. London, UK.

Zhao, Z., X. Shao, B. Huang and F. Liu (2011). On-line estimation of glucose and biomass concentration in penicillin fermentation batch process using particle filter with constraint. In: *Proceedings of the* 4^{th} *International Symposium on Advanced Control of Industrial Processes (ADCONIP)*. IEEE. Hangzhou, China. pp. 391–396.

Zheng, Q. and H. Kimura (2001). Just-in-time modeling for function prediction and its applications. *Asian Journal of Control* **3**(1), 35–44.

Chapter 3

A Classical Framework for Real-time Inferential Modeling and Prediction of Cytotoxicity Induced by Contaminants in Water Resources

3.1 Introduction

Development of a reliable model is a key requirement for investigating the behavior of complex systems. Such descriptive models can help to improve analysis, simulation, design, and control of process systems at both micro and macro levels. Depending on the level of *a priori* knowledge, two different philosophies may guide the choice of modeling strategies, namely **first principles analysis** and **statistical data analysis** (Ljung, 1999). First principles or knowledge-driven models are obtained based on formulating and solving a set of differential and algebraic equations representing physical phenomena. Development of such models requires a deep understanding of transport phenomena, possible reaction pathways, and thermodynamic behavior of the studied systems. The complexity of chemical and biological processes could make first principles modeling infeasible or prohibitively difficult. Therefore, decades of research have been devoted to

A version of this chapter has been published in Computational Biology and Chemistry, Volume 35 (Khatibisepehr *et al.*, 2011*a*). An abbreviated version of this chapter was presented at the 4^{th} International Symposium on Advanced Control of Industrial Processes, May 23-27, 2011, Hangzhou, China (Khatibisepehr *et al.*, 2011*b*).

developing system identification techniques for situations in which complete understanding of the internal mechanisms governing the system dynamics is absent. Data-driven models are thus constructed only based on computational inference of historical relations among system components. The main task is to build a suitable inferential model that is well supported by historical data. In such cases, the main criteria to be considered in model identification are simplicity, generality, and flexibility (Hangos and Cameron, 2001). The more degrees of freedom are allowed in the model structure, the closer the model can approximate the identification data-set. However, adequacy of model fit does not reliably ascertain the performance of the developed inferential sensor, *i.e.*, satisfactory prediction capability on the identification data does not guarantee generalization to other data-sets. Determination of a proper model structure (*e.g.* model order within a specified class) plays a key role in achieving a compromise between accuracy and complexity of the model. In recent years, support vector regression (SVR) (Vapnik, 1999) is gaining popularity due to its many attractive features and promising empirical modeling performance. While the empirical risk minimization (ERM) principle is generally employed in many of the statistical modeling techniques, the SVR implements the structural risk minimization (SRM) principle. This would minimize the upper bound on the generalization error instead of the training error. Based on SRM principle, SVR achieves a balance between the model accuracy and generalization performance. Thus, the over-fitting phenomenon can be avoided and, consequently, a better generalized prediction performance can be achieved.

Another important issue arising in development of data-driven models is related to the non-linearity of underlying mechanisms. In many complex biological processes, a single non-linear model cannot fully capture the dynamics of the system under investigation. In such cases, multi-model inferential sensors can be used to approximate complex processes by concatenating multiple local models. Real-time model identification (Cleveland, 1979; Atkeson *et al.*, 1997), also known as **just-in-time/space modeling** (Zheng and Kimura,

2001), can be viewed as a special modeling technique that results in multi-model inferential sensors with infinite number of local models. The general idea behind this approach is to identify a local model in real-time by prioritizing the identification data-points. The search for the nearest neighbors is carried out from the historical data-set using a predefined notion of similarity. The just-in-time/space modeling techniques can cope with variations in process characteristics and handle non-linearity of underlying mechanisms (Kano and Fujiwara, 2013).

3.1.1 Practical Motivation

Chemical disinfection of water was a major public health triumph of the 20^{th} century. Yet, The ever-increasing number of chemical compounds produced by various industries has prompted the development of research methods for rapid cytotoxicity screening to enhance monitoring the quality of water resources. An essential prerequisite for a successful early-warning system is continuous collection of accurate data describing the risk of toxicant contamination. However, real-time measurement of critical quality variables, such as identity and concentration of potential contaminants, may involve difficulties due to the inadequacy of measurement techniques or low reliability of measuring devices. The key quality indicators are normally available through off-line sample analysis with significant time delays in the order of a few days. Moreover, sampling frequency for routine laboratory analysis might not be adequate to detect intermittent and short-lived contamination. The lack of suitable key variable information in a timely manner can have a severe influence on monitoring the quality of water resources. Therefore, there has been a need to develop on-line sensing tools for timely detection and quantification of potential contaminants with adequate sensitivity, specificity, accuracy, and precision.

Several methods have been applied for detection and identification of the hazardous events in water resources. Analytical-chemical methods have been developed to detect a specific compound or a range of compounds having similar properties. The main

shortcoming of these methods is that they do not necessarily provide any information about the potential toxic effects on the living mechanisms (Brosnan, 1999). In contrast, biological early warning methods are capable of detecting the presence and identifying the consequences of hazardous events, regardless of the type and concentration of contaminants. Despite recent advances in biological monitors and microsensor technologies, effective implementation of these tools has been restricted by the high rate of false positive as well as the need for frequent and high-cost maintenance. To overcome these limitations, applications of inferential models in the assessment of water quality have been widely investigated during the past few decades (Clark *et al.*, 1986; Mazijk, 1996). In recent years, inferential process modeling has been established as a valuable supplement to the classical methods for real-time monitoring the quality of water supplies (Yang *et al.*, 2008).

Water contaminants have two major effects on living cells: 1. toxicity effects resulting in cell death by apoptosis and/or necrosis and 2. cancer effects resulting in uncontrolled cell proliferation. In general, living cells undergo physiological and pathological changes as a result of exposure to toxic compounds. These changes include: morphological dynamics, cell adhesion alterations, cell cycle arrest, DNA damage, and tissue apoptosis and necrosis (Xing *et al.*, 2005). Such cellular changes are dynamic and greatly depend on the cell types as well as the nature, concentration, and exposure duration of toxicant (Botham, 2004). Inferential models can be developed to describe these effects on human cells. Such descriptive models can help to predict cell responses to different type and concentration of water contaminants and, consequently, to assess the biological consequences and cytotoxicity effects of toxicants in environmental contamination (Ibrahim *et al.*, 2010).

3.1.2 Main Contributions

The focus of this work is to provide a classical non-Bayesian framework to capture the non-linearity in the local region around a query point in a real-time manner. The proposed

just-in-time/space modeling approach can cope with variations in process characteristics and handle non-linearity of underlying mechanisms. The developed framework will be implemented to facilitate real-time prediction of cytotoxicity effects on living cells induced by certain water contaminants. Real-time analysis of the intrinsic cell behavior and predicting the trajectory of its progress (growth or death) over a considerable time horizon is a significant contribution in the context of water quality monitoring.

3.2 Problem Statement

For dynamic modeling and prediction of cytotoxicity, both knowledge-driven and data-driven models have been constructed and their results have been presented in the earlier work of Huang and Xing (2006). Not surprisingly, classical knowledge-driven modeling approaches can provide good estimations of dynamic cell responses based on transport equations for cell population. However, the development of a first-principle model becomes practically infeasible if the underlying mechanism is not truly understood. For instance, some phenomena observed in cytotoxicity experiments, such as initial cell fusion, have not been well understood and, thus, cannot be explained from first principles. The difficulties in developing first-principle models for cell-killing mechanisms induced by toxicant were discussed previously in Huang and Xing (2006) in details. Since certain dynamics in cytotoxicity process are very difficult or impossible to model from the first principles due to limited understanding of the complex underlying biochemical and morphological processes, the focus of this work is thus on development of data-driven predictive models.

Some techniques have been developed during the past years for data-driven dynamic modeling of cytotoxicity, most of which are based on time series analysis (Huang and Xing, 2005). However, the nature of cytotoxicity mechanisms is highly non-linear. Being capable of approximating the non-linearity, artificial neural networks (ANNs) with hyperbolic tangent activation functions and non-linear autoregressive with exogenous input (NARX)

model structures have been developed by Huang and Xing (2006). It has been reported that ANN is an effective method for short-term prediction of cell population dynamics in the presence of toxicants. However, it suffers from several limitations:

1. The model performance deteriorates as prediction horizon increases.

2. The accuracy of long-term predictions deteriorate quicker for faster dynamics, such as that of As (III), and for unstable responses resulting from low dose of toxicants.

3. There is no guarantee of convergence and avoidance of local minima.

4. The ANN follows the empirical risk minimization (ERM) approach, which is commonly employed by conventional machine learning methods. In the ERM approach, unknown parameters are adjusted to minimize the prediction errors pertaining to the identification data-set. Since the ERM is based exclusively on the prediction errors for the identification data, a good generalization performance cannot be guaranteed.

5. There are no general methods available to specify the network architecture (Yan *et al.*, 2004).

6. In spite of the ability of ANNs to handle non-linearity, a single non-linear model cannot fully capture the dynamics of such complex biological processes.

To overcome the aforementioned shortcomings, the proven advantages of SVR inspire us to employ it in constructing a data-driven predictive model to improve the effectiveness and efficiency of cytotoxicity monitoring investigated in Huang and Xing (2006). Among the different formulations of the SVR problem, the ν-SVR (Schölkopf *et al.*, 2000) algorithm is adopted to form the core of the inferential model. Moreover, a just-in-time/space modeling technique is developed to better approximate the local process behavior. Thus, it is required to effectively construct an identification sub-set from the most relevant training samples.

Since postulate of conventional statistic theory is infinite numbers of identification samples, for small size training data-sets the principle of ERM cannot deduce the principle of expectation risk minimization (Han *et al.*, 2004). However, support vector regression is based on small-sample statistical learning theory, in which the optimal solution can be obtained from limited identification samples rather than infinite samples in theory (Vapnik, 1998). The formulation of SVRs embodies the structural risk minimization (SRM) principle that minimizes an upper bound on the expected risk, as opposed to the ERM principle that minimizes the prediction error on the identification data. The difficulties of choosing network structure are automatically handled in SVR.

One of the standard MATLAB toolboxes, LIBSVM, is applied to the construction of cell index prediction models. The developed model is found capable of analyzing intrinsic cell behavior and predicting the trajectory of its progress (growth or death) over considerable time horizon.

3.3 Cytotoxicity Experiments

Collecting reliable experimental data is the first and the most important step to ensure the quality of the inferential models. However, dynamic monitoring of the cytotoxicity is not an easy task in most of the conventional cell-based assays. The main reason is that the required chemical and radiation indicators may kill or disturb target cells. In order to detect a broad range of physiological and pathological dynamic responses of living cells to toxicants, an automatic, real-time cell electronic sensing (RT-CES) system has been used to conduct the cytotoxicity experiments as detailed in Xing *et al.* (2005).

The RT-CES system (ACEA Biosciences, CA, U.S.A.) has been used to monitor cellular events by measuring the electronic impedance of sensor electrodes integrated on the bottom of microtiter plates. The RT-CES system has been described in details in Xing *et al.* (2005). Briefly, the system is composed of three main components: 1. electronic sensor analyzer,

2. device station, and 3. $16x$ microelectronic sensor device. Cells were grown onto the surfaces of microelectronic sensors, which are comprised of circle-on-line electrode arrays and are integrated into the bottom surfaces of the microtiter plate. The device station is located inside a tissue culture CO_2 incubator and is capable of electronically switching any one of the wells to the sensor analyzer for impedance measurement. Sensor analyzer can automatically select wells for measurement and continuously transfer measured impedance data to the computer. Based on the measured impedance, a dimensionless parameter termed cell index (CI) is defined to provide quantitative information about the biological status of the cells such as cell number. The CI can be calculated as (Xing *et al.*, 2005)

$$CI = \max_{n=1,\ldots,N} \left[\frac{R_{cell}(f_n)}{R_b(f_n)} - 1 \right] \tag{3.1}$$

where $R_b(f_n)$ and $R_{cell}(f_n)$ are the frequency-dependent electrode impedance (resistance) without or with cells present in the wells, respectively, and N is the number of the frequency points at which the impedance is measured.

Since the cell index is an indirect indication of the biological consequences of toxic contaminants in the aquatic environment, the CI measurements can be used to outline and implement a concept for developing dynamic predictive models. For cytotoxicity assessment, the NIH 3T3 cell-lines were treated with three potential water toxicants, sodium arsenite [As (III)], mercury (II) chloride, and sodium dichromate [chromium (VI)]. The starting cell number was 10,000 cells per sensor wells. When the CI values reached a range between 1.0 and 1.2, the cells were exposed to one of the toxicants at different concentrations. The cell responses were continuously monitored and recorded every hour by the RT-CES system.

3.4 Support Vector Regression

The standard support vector regression algorithm is revisited in this section. First, a brief overview of the principles of SVR is presented. Next, the issues related to solving dynamic

modeling problems and tuning hyperparameters are discussed. Readers are referred to Vapnik (1998) for a more in depth discussion of the associated statistical learning theory.

3.4.1 Historic Background

Support vector machines (SVMs) are a group of supervised learning methods that implement the structural risk minimization inductive principle to obtain good generalization on a limited number of training samples. The support vector (SV) theory is firmly grounded in the framework of statistical learning theory which has been developed over the last three decades by Vapnik and Chervonenkis (1974) and others. In its present form, the SVM was developed by Vapnik and his co-workers on a basis of a separable bipartition problem at the AT & T Bell Laboratories. The SV method has now evolved into an active area of research oriented towards real-world applications. There are two main categories of support vector machines, namely support vector classification (SVC) and support vector regression (SVR).

3.4.2 Basis of Support Vector Regression

The basic idea of the SVR is to non-linearly map the input variables into a high dimensional feature space wherein they are linearly correlated with the output variable. Let $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ denote the identification data-set, where $\mathbf{x}_n \in \mathbb{R}^P$ is the vector of input data and y_n is the corresponding scalar output value. Support vector regression aims at finding the flattest linear regression function that deviates from the identification data by ε at most. That is,

$$f(\mathbf{x}) = \langle \mathbf{w}, \Omega(\mathbf{x}) \rangle + b \tag{3.2}$$

satisfies the following condition:

$$|f(\mathbf{x}_n) - y_n| \leq \varepsilon \quad n = 1, \ldots, N \tag{3.3}$$

where **w** denotes the weight vector connecting the feature space to the output space, b is the bias term, f denotes the feature function, Ω is the mapping function (*i.e.* $\Omega : \mathbb{R}^P \mapsto \mathbb{F}$), and $\langle w, \Omega(\mathbf{x})\rangle$ represents the dot product in the feature space, $\mathbb{F}$.

Support vector regression models are identified by solving the following optimization problem (Smola and Schölkopf, 2004):

$$\min_{\mathbf{w},b,\xi_n,\xi_n^*} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) \tag{3.4}$$

$$\text{Subject to} \begin{cases} y_n - (\langle \mathbf{w}, \Omega(\mathbf{x}_n)\rangle + b) \leq \varepsilon + \xi_n \\ (\langle \mathbf{w}, \Omega(\mathbf{x}_n)\rangle + b) - y_n \leq \varepsilon + \xi_n^* \\ \xi_n, \xi_n^* \geq 0 \end{cases}$$

where ξ_n (and ξ_n^*) are the slack variables of the upper (and lower) training error subject to the ε-insensitive tube (Figure 3.1). The constant C determines the trade-off between the flatness of function f and the acceptable amount of deviations from ε. In other words, C is a bound on Lagrangian multipliers.

The formulation above corresponds to dealing with the ε-insensitive loss function described by (Vapnik, 1998)

$$|\xi|_\varepsilon := \begin{cases} 0 & |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{Otherwise} \end{cases} \tag{3.5}$$

As shown in Figure 3.1, this loss function builds a tube of insensitivity inside which the prediction errors are not of concern. Only points outside of the ε-insensitive tube are penalized to minimize the resulting errors, ξ or ξ^*, in the objective function. Besides the ε-insensitive, other loss functions such as quadratic, Laplace or Huber can be used in SVR algorithm.

Using an ε-insensitive loss function, the non-linear SVR solution is found by minimizing

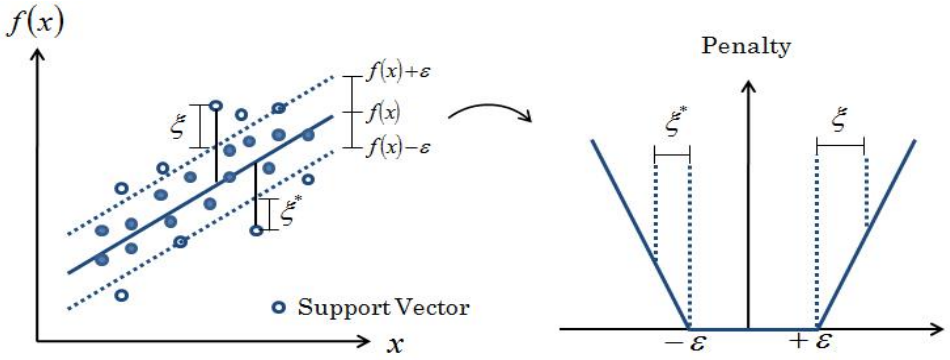


Figure 3.1: Right Panel: Tube of insensitivity. Left Panel: ε-insensitive loss function

the following primal Lagrangian:

$$\begin{aligned} L = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) - \sum_{n=1}^{N}(\eta_n\xi_n + \eta_n^*\xi_n^*) \\ - \sum_{n=1}^{N}\alpha_n(\varepsilon + \xi_n - y_n + \langle \mathbf{w}, \Omega(\mathbf{x}_n)\rangle + b) \\ - \sum_{n=1}^{N}\alpha_n^*(\varepsilon + \xi_n^* + y_n - \langle \mathbf{w}, \Omega(\mathbf{x}_n)\rangle - b) \end{aligned} \tag{3.6}$$

Equation 3.6 is minimized with respect to the primal variables, $\mathbf{w}$, b, ξ_n and ξ_n^*. Moreover, the dual Lagrangian needs to be maximized with respect to the non-negative Lagrangian multipliers, α_n and α_n^*:

$$\begin{aligned} L_D = -\frac{1}{2}\sum_{n,j=1}^{N}(\alpha_n - \alpha_n^*)(\alpha_j - \alpha_j^*)\mathcal{K}(\mathbf{x}_n, \mathbf{x}_j) \\ - \varepsilon\sum_{n=1}^{N}(\alpha_n + \alpha_n^*) + \sum_{n=1}^{N} y_n(\alpha_n - \alpha_n^*) \end{aligned} \tag{3.7}$$

$$\text{Subject to } \textstyle\sum_{n=1}^{N}(\alpha_n - \alpha_n^*) = 0 \qquad 0 \le \alpha_n, \alpha_n^* \le C \qquad n = 1, \ldots, N$$

where $\mathcal{K}(\mathbf{x}_n, \mathbf{x}_j) = \langle \Omega(\mathbf{x}_n), \Omega(\mathbf{x}_j)\rangle$ is the Kernel function. The most common Kernel functions are listed in Table 3.1.

After calculating α_n and α_n^*, optimal desired weights of the regression hyperplane is obtained from

$$\mathbf{w} = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)\Omega(\mathbf{x}_n) \tag{3.8}$$

Table 3.1: Summary of the most common kernel functions

Kernel	$\mathcal{K}(\mathbf{x}_n, \mathbf{x})$	Hyperparameters
Linear	$\langle \mathbf{x}_n, \mathbf{x} \rangle$	None
Polynomial	$\Big(\gamma \langle \mathbf{x}_n, \mathbf{x} \rangle + c\Big)^d$	$d \in \mathbb{N}$, $c \geq 0$
Gaussian RBF	$\exp\Big(\frac{-\|\mathbf{x}_n - \mathbf{x}\|^2}{2\sigma^2}\Big)$	$\sigma > 0$
Sigmoidal	$\tanh\Big(\gamma \langle \mathbf{x}_n, \mathbf{x} \rangle + c\Big)$	$\gamma, c \geq 0$

Finally, the SVR model is given by

$$f(\mathbf{x}) = \sum_{n=1}^{nSV} (\alpha_n - \alpha_n^*) \mathcal{K}(\mathbf{x}_n, \mathbf{x}) + b \tag{3.9}$$

where nSV is the number of support vectors.

A final note has to be made regarding the tuning of hyperparameters of the SVR model. Hyperparameters are high level parameters that may influence the training procedure. They are the constants defining a particular instance of a learning algorithm or involved in the model. They are not usually determined by the learning algorithm, but are instead fixed at the design stage. Several possibilities of hyperparameter determination exist. A very rare possibility is that hyperparameters are known in advance as prior knowledge. Almost always, however, they must be determined during the identification phases (Cherkassky and Ma, 2004). Since the quality of the SVR models depends greatly on the proper tuning of the hyperparameters, the procedures available for tuning the two most relevant hyperparameters are discussed next.

3.4.3 Tuning of Hyperparameters

The tuning of the regularization constant, C, is a delicate task. A larger C implies a smaller training error but possibly a lower generalization performance. On the other hand, a smaller C gives more weight to the regularization term as a result of which a better

generalization performance can be achieved. According to Cherkassky and Ma (2004), the value of hyperparameter C can be directly adjusted on the identification data-set:

$$C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|) \tag{3.10}$$

where $\bar{y}$ and σ_y are the mean and standard deviation of the measured outputs, respectively.

For an SVR-based model, the value of ε determines the level of accuracy of the ε-insensitive loss function. Hyperparameter ε has an effect on both the complexity and the generalization performance of the developed model. It is well-known that the value of hyperparameter ε depends on the level of noise in the identification data, and on the size of the training data-set. However, in many real-world applications, the noise level is unknown. This problem is partially resolved using a new support vector regression algorithm known as ν-SVR (Schölkopf *et al.*, 2000). In this algorithm, a new parameter ν allows us to automatically adjust the width of ε-insensitive tube. Thus, the optimization problem of Equation 3.4 is rewritten as

$$\min_{\mathbf{w},b,\xi_n,\xi_n^*,\varepsilon} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\Big(\sum_{n=1}^{N}(\xi_n + \xi_n^*) + N\nu\varepsilon\Big) \tag{3.11}$$

$$\text{Subject to} \begin{cases} y_n - (\langle \mathbf{w}, \Omega(\mathbf{x}_n)\rangle + b) \leq \varepsilon + \xi_n^* \\ (\langle \mathbf{w}, \Omega(\mathbf{x}_n)\rangle + b) - y_n \leq \varepsilon + \xi_n \\ \nu, \xi_n, \xi_n^* \geq 0 \end{cases}$$

The procedure followed to solve the ν-SVR optimization problem is similar to that of classical SVR. This would lead to the same expression given by Equation 3.9. Schölkopf *et al.* (2000) showed that $\nu \in (0, 1]$ can be interpreted both as an upper bound on the percentage of errors and as a lower bound on the fraction of support vectors at the end of the training. In this way, one can directly control the number of parameters needed to build the regression function. This interesting aspect becomes particularly advantageous when one has to deal with small number of training samples in training phase or rigorous limitations in implementation phase. Note that choosing ν to represent a certain level of accuracy does

of course only guarantee that accuracy on the training set. Often, it is required to choose a larger value of ν in order to achieve a certain level of accuracy overall.

3.4.4 Local ν-Support Vector Regression

The adjacent CI measurements would reflect similar dynamic behavior of the living cells. Given a query sample, the local ν-SVR model identified from the neighboring training samples are thus expected to better capture the local non-linearity and produce more accurate predictions. In this section, a local ν-SVR modeling technique inspired by the work of Fernández (1999) is presented.

Given a query sample, $\mathbf{x}_t$, the search for at most O nearest neighbors is carried out from the identification data-set using a pre-defined notion of similarity or locality. A local ν-SVR model is identified from at most the O nearest neighbors among the training data. In this work, the distance between the query and identification samples is selected as the measure of similarity. The distance metric is defined in terms of the toxicant concentration and latest CI measurements. Having constructed a sub-set of identification data, the ν-SVR algorithm can be used to identify a local model.

Let $\mathcal{D}$ and $\mathcal{D}^V$ denote the identification and validation data-sets, respectively. The implementation procedure of the local ν-SVR approach employed in this study is summarized in Algorithm 3.1.

Algorithm 3.1. *Local ν-Support Vector Regression Modeling*

1. Choose an appropriate Kernel mapping function. A summary of the most common Kernel functions is given in Table 3.1.

2. Determine a search space for the hyperparameters. Find the optimal values of the hyperparameters from the identification data-set (Equation 3.10).

3. Given a query sample $\mathbf{x}_t$, find at most the P spatial and temporal nearest neighbors among the training samples. Construct a new identification data-set based on the

selected O neighbors such that $\mathcal{D}_O^q \in \mathcal{D}$.

4. Identify a local ν-SVR model from $\mathcal{D}_O^q$ to obtain the support vectors and corresponding weight coefficients (Equation 3.8).

5. Obtain a prediction of y_q, denoted as $\widehat{y}_q$, based on the identified local ν-SVR inferential model (Equation 3.9).

6. Repeat the above steps until CI predictions of all query samples in the validation data-set are all acquired. Tune the hyperparameters of the model as required.

3.5 SVR-Based Predictive Model

3.5.1 Data Selection

The entire data-set is divided into two sub-sets, namely the **identification** and **validation** data-sets. The identification data-set is used to develop a SVR-based model, whereas the validation data-set is adopted to monitor the level of agreement between the identified model and the process under investigation. Since the distribution of the identification data within the process operating region is crucial for obtaining reliable predictions, the data for model training should be chosen carefully. Model predictions can be trusted only if a new query sample belongs to the operating region covered by the identification data.

For As (III) experiments, experimental data consist of eight different doses of toxicant including one zero-dosage (also known as controlled experiment). For chromium (VI) and mercury (II) chloride experiments, experimental data consist of seven different doses of toxicant including one zero-dosage. Each dose corresponds to 25 dynamic data points recorded after the injection of the toxicant with one-hour sampling interval. Identification and validation data-sets are chosen such that both consist of a mixture of low-dose and high-dose samples.

3.5.2 Model Development

Two different types of predictions will be investigated, namely k-step-ahead prediction and varying-horizon prediction.

First, let us consider the following one-step-ahead predictor for short-term dynamic prediction of future CI values:

$$\widehat{y}_{t+1|t} = \sum_{n=1}^{nSV} (\alpha_n - \alpha_n^*) \mathcal{K}(\mathbf{x}_n, \mathbf{x}_t) + b \tag{3.12}$$

where $\{\mathbf{x}_1, \ldots, \mathbf{x}_{nSV}\}$ are the support vectors selected from the training data-set and $\mathbf{x}_t$ is the regression vector constructed from three lagged outputs, y_t, y_{t-1}, and y_{t-2}, as well as toxicant concentration at time $t+1$, u_{t+1}. That is, $\mathbf{x}_t = [y_t, y_{t-1}, y_{t-2}, u_{t+1}]$. Finally, $\widehat{y}_{t+1|t}$ denotes the one-step-ahead prediction of y_t obtained from all information available up to and at time t.

The k-step-ahead predictions are obtained by iterating the one-step-ahead predictor up to the desired horizon. In order to perform a k-step-ahead prediction, the regressor should be updated upon the arrival of new measurements. Hence, the regression vector consists of predicted values as well as available actual measurements. A prediction iterated for k times returns a k-step-ahead prediction. Hence, the multi-step-ahead predictive model is defined as

$$\widehat{y}_{t+k|t} = \sum_{n=1}^{nSV} (\alpha_n - \alpha_n^*) \mathcal{K}(\mathbf{x}_n, \mathbf{x}_t) + b \tag{3.13}$$

To illustrate for a model using three lagged outputs, the iterative three-step-ahead prediction, $\widehat{y}_{t+3|t}$, is achieved as follows:

- First stage:

$$\widehat{y}_{t+1|t} = \sum_{n=1}^{nSV} (\alpha_n - \alpha_n^*) \mathcal{K}(\mathbf{x}_n, \mathbf{x}_t) + b$$

$$\text{with } \mathbf{x}_t = [y_t, y_{t-1}, y_{t-2}, u_{t+1}]$$

- Second stage:

$$\widehat{y}_{t+2|t} = \sum_{n=1}^{nSV} (\alpha_n - \alpha_n^*)\mathcal{K}(\mathbf{x}_n, \mathbf{x}_t) + b$$
$$\text{with } \mathbf{x}_t = [\widehat{y}_{t+1|t}, y_t, y_{t-1}, u_{t+2}]$$

- Third stage:

$$\widehat{y}_{t+3|t} = \sum_{n=1}^{nSV} (\alpha_n - \alpha_n^*)\mathcal{K}(\mathbf{x}_n, \mathbf{x}_t) + b$$
$$\text{with } \mathbf{x}_t = [\widehat{y}_{t+2|t}, \widehat{y}_{t+1|t}, y_t, u_{t+3}]$$

For a k-step-ahead predictor, the derived model can pass the validation easier when the prediction horizon k is small and output response is smooth. However, the prediction becomes less accurate as k increases.

Now, let us further investigate the potential of the one-step-ahead predictor. The objective is to evaluate the performance of the derived one-step-ahead predictor to predict the long-term cytotoxicity response only based on the first three measurements such that

$$\widehat{y}_{3+k|3} = f(y_1, y_2, y_3; u_3) \quad \text{for} \quad k = 1 \ldots 22 \tag{3.14}$$

or equivalently,

$$\widehat{y}_t = f(y_1, y_2, y_3; u_3) \quad \text{for} \quad t = 4 \ldots 25 \tag{3.15}$$

The k-step-ahead prediction approaches the infinite-step-ahead prediction, also known as **simulation**, as prediction horizon increases. To illustrate, consider a stable system (*i.e.* $|a_1| < 1$) of the form

$$\widehat{y}_{t+1|t} = a_1 y_t + b_1 u_t \tag{3.16}$$

One approach to find $\widehat{y}_{t+2|t}$ is to back-substitute from the defining equation of the process

so as to eliminate future values of y_t. Thus,

$$\begin{aligned}\widehat{y}_{t+2|t} &= a_1\widehat{y}_{t+1|t} + b_1u_{t+1} \\ &= a_1(a_1y_t + b_1u_t) + b_1u_{t+1} \\ &= a_1^2y_t + a_1b_1u_t + b_1u_{t+1}\end{aligned} \tag{3.17}$$

Similarly, $\widehat{y}_{t+k|t}$ can be calculated as

$$\widehat{y}_{t+k|t} = a_1^k y_t + b_1 z^{k-1}u_t \sum_{j=0}^{k-1} a_1^j z^{-j} \tag{3.18}$$

As k goes to infinity, the first term approaches zero. Assuming the input remains constant, the infinite-step-ahead prediction can be approximated as

$$\begin{aligned}\widehat{y}_{t+k|t} &= b_1 z^{k-1}u_t \sum_{j=0}^{\infty} a_1^j z^{-j} \\ &= \frac{b_1 z^{k-1}}{1 - a_1 z^{-1}}\end{aligned} \tag{3.19}$$

leading to

$$\widehat{y}_{t+\infty|t} = \frac{b_1}{1 - a_1 z^{-1}}u_t \tag{3.20}$$

Taking advantage of the above fact, varying-horizon predictions can be calculated to simulate the dynamic response of living cells to various concentrations of toxic compounds. Given the first three CI measurements, the varying-horizon predictive model is defined as

$$\widehat{y}_t = \sum_{n=1}^{nSV}(\alpha_n - \alpha_n^*)\mathcal{K}(\mathbf{x}_n, \hat{\mathbf{x}}_t) + b \quad \text{for} \quad t = 4 \ldots 25 \tag{3.21}$$

In this framework, the predicted output is fed back as an input to the following prediction. Hence, the regression vector consists of predicted values as opposed to actual measurements, *i.e.* $\hat{\mathbf{x}}_t = [\hat{y}_{t-1}, \hat{y}_{t-2}, \hat{y}_{t-3}, u_t]$.

There are two types of model validations performed, namely **self-validation** and **cross-validation**. Self-validation determines the adequacy of fit by evaluating the prediction performance of the inferential model on the identification data. Cross-validation assesses

Table 3.2: Summary of the optimal values of hyperparameters C and ν

Toxicant	Hyperparameter C	Hyperparameter ν
As (III)	3.7	0.7
Chromium (VI)	2.7	0.9
Mercury (II) Chloride	2.5	0.95

the generalization capability by evaluating the prediction performance of the identified inferential model on an independent data-set that has not been used for the model identification. It is noteworthy that the prediction plots do not have the first k points because the previous k outputs must be known for a k-step-ahead prediction.

3.6 Results and Discussion

3.6.1 Dynamic Prediction

In this section, the accuracy of short-term and long-term predictions is evaluated. For evaluating the short-term prediction performance, the accuracy of one-step-ahead (one-hour) and five-step-ahead (five-hour) predictions are considered. For evaluating the long-term prediction performance, the accuracy of varying-horizon predictions are considered in which only first three measurements are used to predict all future responses. To show the relationship between the prediction performance and prediction horizon, the performance of the one-step-ahead, three-step-ahead, and five-step-ahead predictors are compared.

For As (III) modeling, four experimental data-sets corresponding to 1.25, 6.21, 13.58 and 29.64 μM are used for model training and hyperparameter tuning, while the remaining four are reserved for cross-validation. As mentioned previously, it is of primal importance to find optimal values for the hyperparameters of the SVR-based model. Equation 3.10 is thus used to directly choose hyperparameter C based on the identification data. For a fixed C, the whole parameter range $(0, 1]$ is examined to search for the optimal value of ν. The selected values of hyperparameters C and ν are presented in Table 3.2.

Figure 3.2 shows comparison of model fit based on one-step-ahead and five-step-ahead predictions for As (III) toxicant. Both self-validation and cross-validation results are presented in this figure. The one-step-ahead predictions obtained from the SVR-based model are comparable with the corresponding predictions from ANNs presented in Huang and Xing (2006). However, the five-step-ahead predictions have noticeably improved in this work. Given the first three CI measurements, varying-horizon predictions are presented in Figure 3.3. With reasonable fits overall, one can see that better models and predictions are obtained for larger doses of AS(III). Compared to the simulation results illustrated in Huang and Xing (2006), a significant improvement of the long-term predictions is observed here. To provide an arithmetical basis for comparison, mean absolute error (MAE) of predictions of SVR-Based models and ANNs are summarized in Table 3.3.

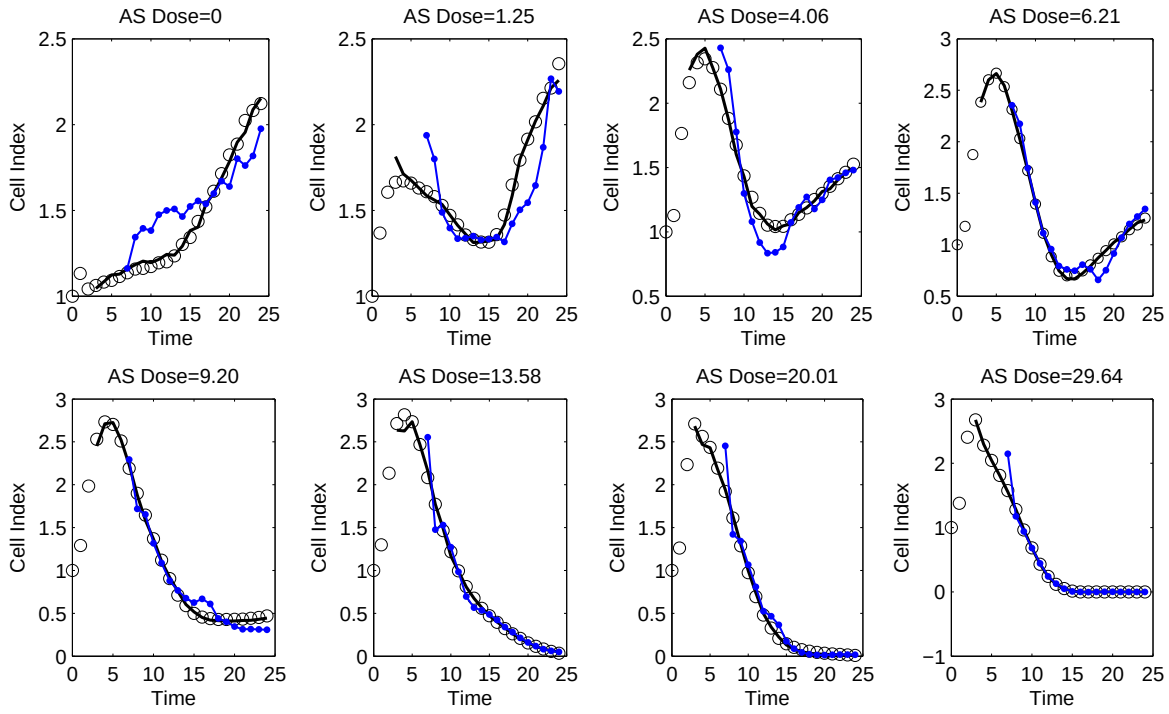


Figure 3.2: Model fit based on short-term predictions for As (III) toxicant; solid line is one-step-ahead prediction; marked line is five-step-ahead prediction; circle is actual measurement of CI

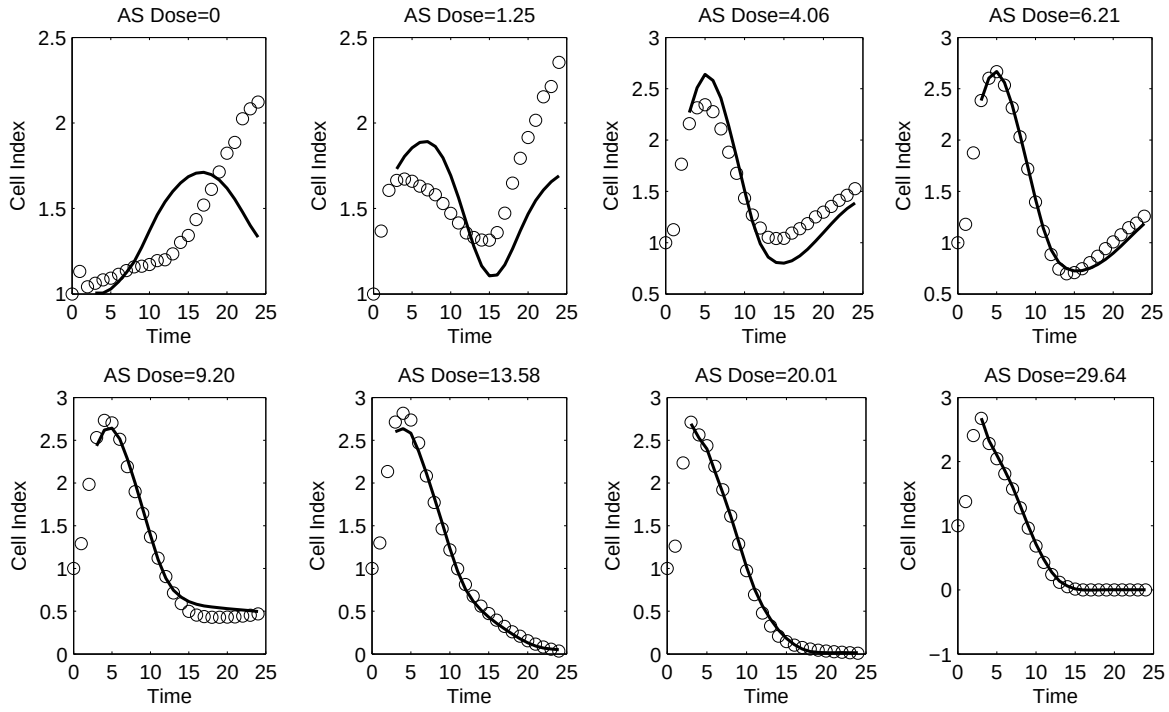


Figure 3.3: Varying-horizon predictions for As (III) toxicant given the first three CI measurements; solid line is prediction; circle is actual measurement of CI.

Table 3.3: Comparison of mean absolute errors resulted from SVR-Based models and ANNs for As (III) toxicant

	0	*1.25*	*4.06*	*6.21*	*9.20*	*13.58*	*20.01*	*29.64*
One-step-ahead Predictions								
SVR	0.024	0.027	0.026	0.011	0.024	0.022	0.021	0.003
ANN	0.032	0.041	0.033	0.025	0.024	0.032	0.024	0.067
Five-step-ahead Predictions								
SVR	0.166	0.152	0.129	0.071	0.096	0.070	0.084	0.042
ANN	0.182	0.257	0.186	0.140	0.134	0.089	0.099	0.325
Varying-horizon Predictions								
SVR	0.246	0.289	0.199	0.049	0.073	0.052	0.032	0.018
ANN	0.218	0.455	0.292	0.252	0.147	0.092	0.103	0.377

For chromium (VI) modeling, four experimental data-sets corresponding to 0, 0.91, 2.89 and 5.78 μM are used for model training and hyperparameter tuning, while the remaining three are retained for cross-validation. The optimal values of hyperparameters are determined as $C = 2.7$ and $\nu = 0.9$ as presented in Table 3.2. Figure 3.4 shows comparison of model fit based on one-step- and five-step-ahead predictions for chromium (VI) toxicant. Given the first three CI measurements, long-term prediction results for varying-horizons of $k = 1 \ldots 22$ are plotted in Figure 3.5. Long-term predictions have been greatly improved in comparison with the five-step-ahead predictions as well as the simulation results presented in Huang and Xing (2006). From Table 3.4 it can be observed that the magnitude of the prediction errors resulting from the SVR-based models are markedly smaller than those of the ANNs.

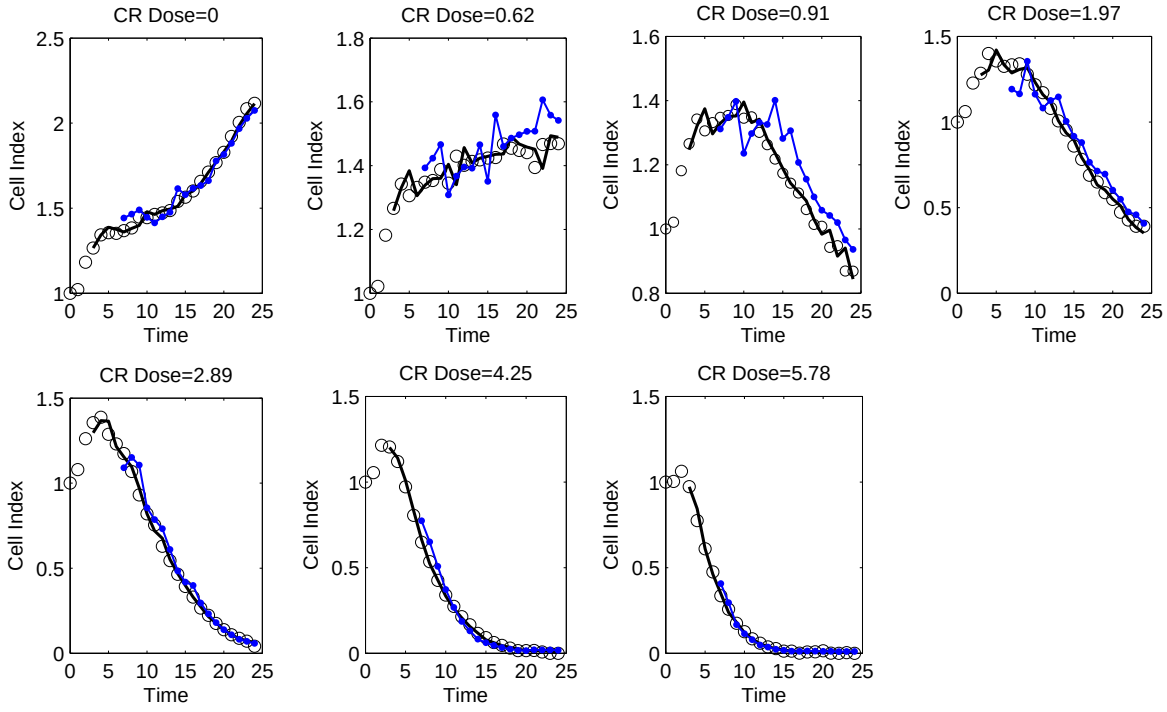


Figure 3.4: Model fit based on short-term predictions for chromium (VI) toxicant; solid line is one-step-ahead prediction; marked line is five-step-ahead prediction; circle is actual measurement of CI.

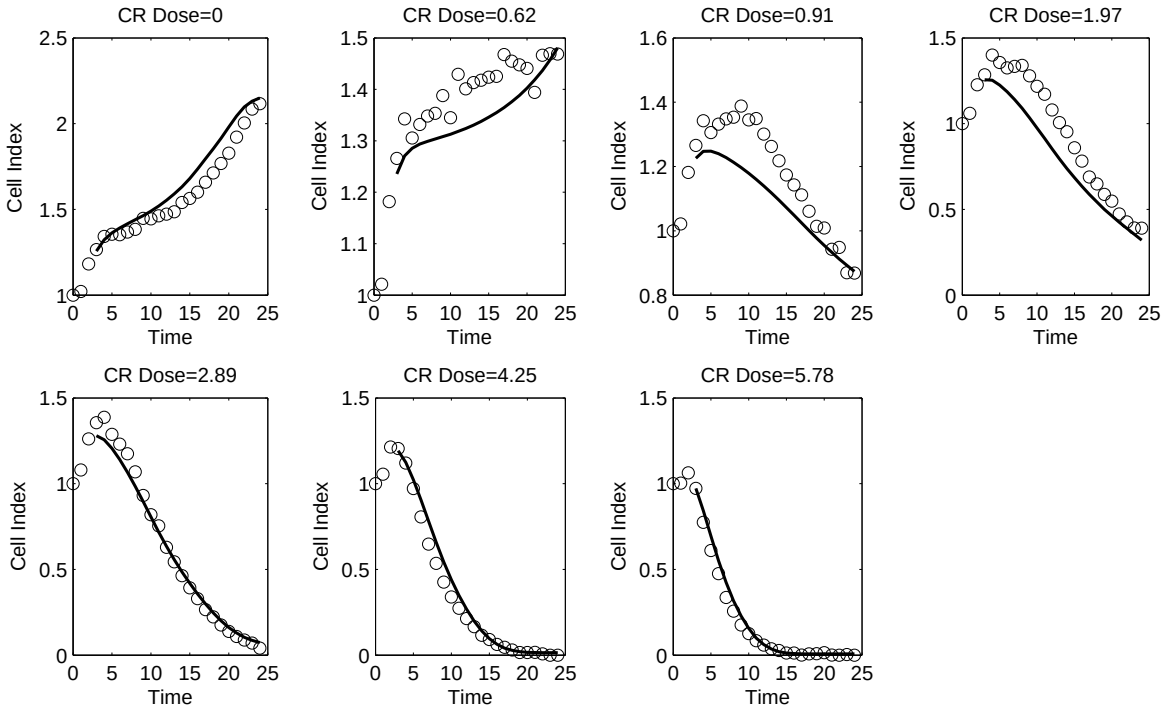


Figure 3.5: Varying-horizon predictions for chromium (VI) toxicant given the first three CI measurements; solid line is prediction; circle is actual measurement of CI.

Table 3.4: Comparison of mean absolute errors resulted from SVR-Based models and ANNs for chromium (VI) toxicant

	0	*0.62*	*0.91*	*1.97*	*2.89*	*4.25*	*5.78*
One-step-ahead Predictions							
SVR	0.014	0.031	0.028	0.026	0.018	0.009	0.008
ANN	0.017	0.032	0.032	0.029	0.020	0.017	0.011
Five-step-ahead Predictions							
SVR	0.037	0.064	0.079	0.080	0.042	0.032	0.012
ANN	0.068	0.069	0.083	0.085	0.067	0.048	0.043
Varying-horizon Predictions							
SVR	0.077	0.056	0.092	0.144	0.043	0.040	0.024
ANN	0.082	0.143	0.289	0.240	0.191	0.062	0.621

For mercury (II) chloride modeling, four experimental data-sets corresponding to 0, 15.2, 32.8 and 71 μM are used for model training and hyperparameter tuning, while the

remaining three are reserved for validation. The optimal values of hyperparameters C and ν are presented in Table 3.2. Figure 3.6 shows comparison of model fit based on one-step- and five-step-ahead predictions for mercury (II) chloride toxicant. Long-term prediction results for varying-horizons of $k = 1 \ldots 22$ based on first three CI measurements are presented in Figure 3.7. The statistical results of the comparison are presented in Table 3.5. It is observed that the SVR-based models, on average, outperform the ANNs in terms of accuracy particularly for longer prediction horizons.

Overall, the one-step-ahead predictions indicate a good fit to the CI data for the developed SVR-based models. The predictive performance of the developed models deteriorates with the increase of the prediction horizon, as expected. To demonstrate the pattern of the prediction performance versus the prediction horizon, one-step-, three-

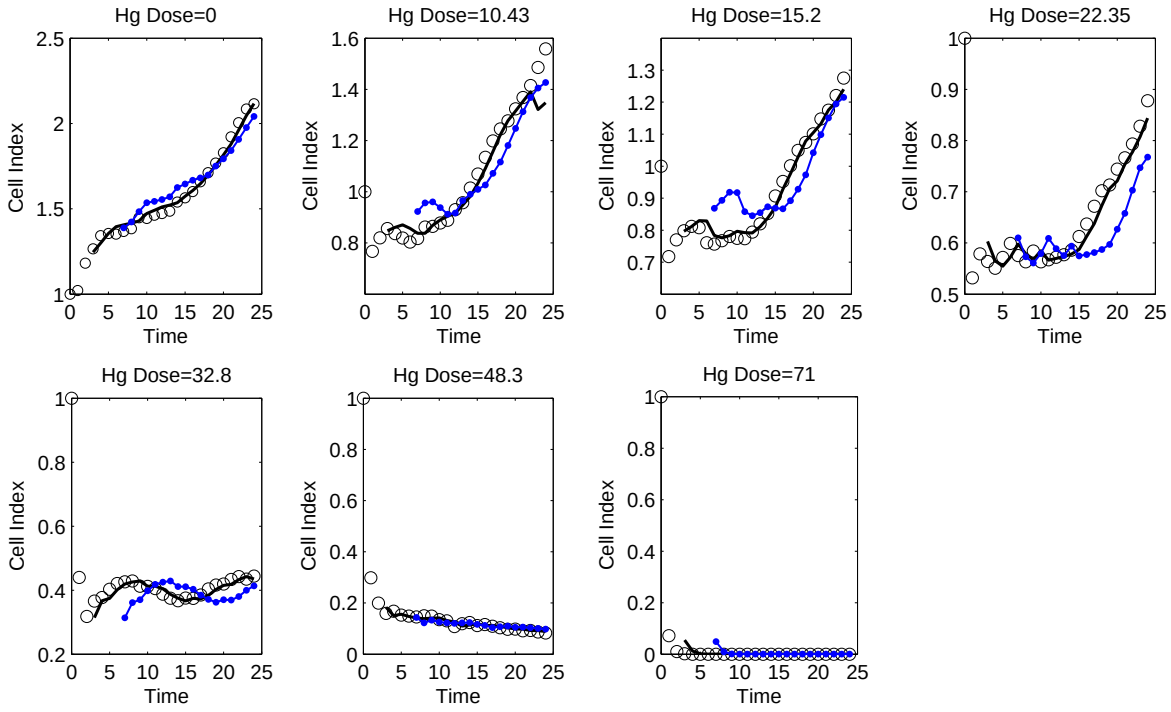


Figure 3.6: Model fit based on short-term predictions for mercury (II) chloride toxicant; solid line is one-step-ahead prediction; marked line is five-step-ahead prediction; circle is actual measurement of CI.

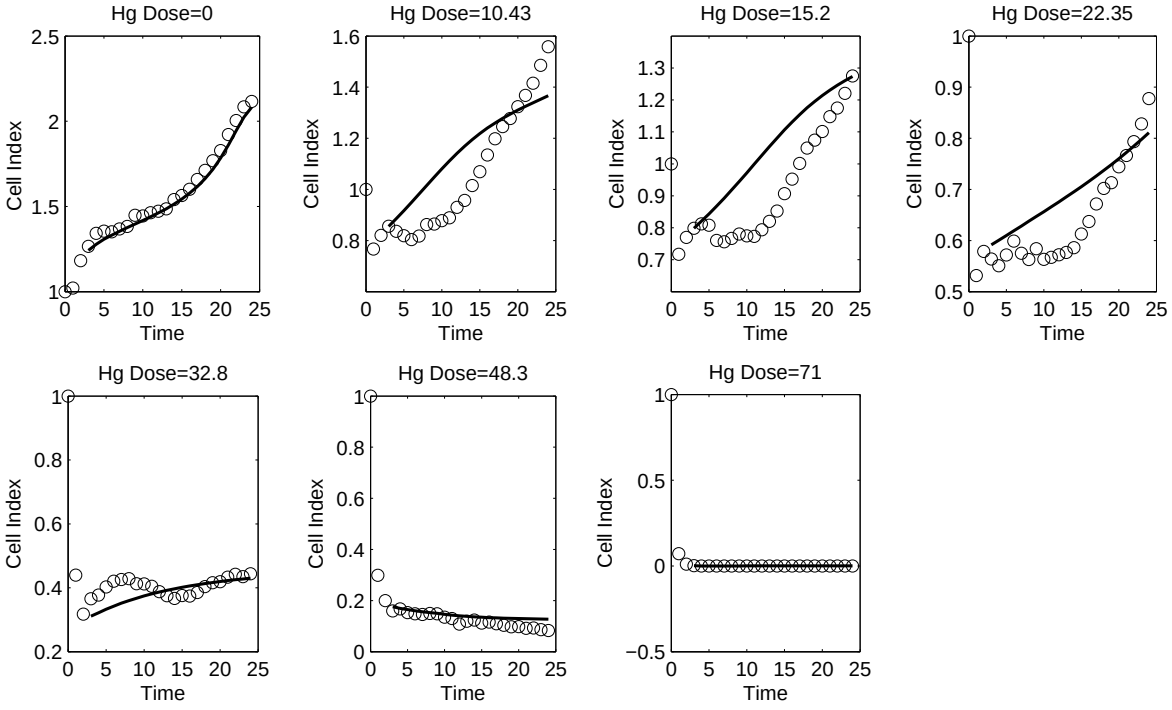


Figure 3.7: Varying-horizon predictions for mercury (II) chloride toxicant given the first three CI measurements; solid line is prediction; circle is actual measurement of CI.

Table 3.5: Comparison of mean absolute errors resulted from SVR-Based models and ANNs for mercury (II) chloride toxicant

	0	*10.43*	*15.2*	*22.35*	*32.8*	*48.3*	*71*
One-step-ahead Predictions							
SVR	0.024	0.038	0.017	0.019	0.013	0.007	0.003
ANN	0.017	0.019	0.014	0.020	0.007	0.033	0.006
Five-step-ahead Predictions							
SVR	0.062	0.074	0.077	0.0600	0.043	0.009	0.003
ANN	0.033	0.094	0.076	0.067	0.028	0.148	0.004
Varying-horizon Predictions							
SVR	0.035	0.120	0.128	0.057	0.032	0.021	0.0004
ANN	0.101	0.661	0.101	0.088	0.023	0.117	0.4121

Table 3.6: Comparison of mean absolute errors based on different prediction horizons

As (III) Toxicant								
	0	*1.25*	*4.06*	*6.21*	*9.20*	*13.58*	*20.01*	*29.64*
$k = 1$	0.0241	0.0270	0.0264	0.0110	0.0237	0.0222	0.0208	0.0026
$k = 3$	0.0833	0.0906	0.0865	0.0385	0.0624	0.0447	0.0469	0.0209
$k = 5$	0.1660	0.1524	0.1291	0.0714	0.0961	0.0703	0.0842	0.0415
Chromium (VI) Toxicant								
	0	*0.62*	*0.91*	*1.97*	*2.89*	*4.25*	*5.78*	
$k = 1$	0.0142	0.0309	0.0281	0.0263	0.0184	0.0095	0.0084	
$k = 3$	0.0313	0.0444	0.0460	0.0475	0.0271	0.0277	0.0154	
$k = 5$	0.0369	0.0635	0.0791	0.0800	0.0422	0.0321	0.0119	
Mercury (II) Chloride Toxicant								
	0	*10.43*	*15.2*	*22.35*	*32.8*	*48.3*	*71*	
$k = 1$	0.0239	0.0380	0.0174	0.0188	0.0127	0.0071	0.0031	
$k = 3$	0.0426	0.0507	0.0473	0.0378	0.0302	0.0087	0.0028	
$k = 5$	0.0621	0.0742	0.0770	0.0599	0.0432	0.0093	0.0034	

step- and five-step-ahead prediction errors are summarized in Table 3.6 for comparison. A deterioration in the predictive performance is clearly observed as the prediction horizon becomes larger. It can be observed that the short-term predictions are more accurate than the varying-horizon predictions. This is expected since the regressor for short-term predictions are continuously updated as and when new CI measurements become available. However, the long-term varying-horizon predictions are obtained based on only the first three CI measurements. That being said, the prediction results indicate that the local ν-SVR is feasible for making one-step-ahead as well as multi-step-ahead predictions. With the prediction horizon k growing, prediction performance does not show remarkable descending tendency on both training and validation data-sets.

3.6.2 Dynamic Cytotoxicity Analysis

The dynamic CI patterns of the NIH 3T3 cells in response to As (III), mercury (II) chloride, and chromium (VI) are distinct. This would indicate that different cell-killing mechanisms

were induced in response to these toxic compounds. The As(III)-treated cells show a significant but transient increase in the CI values during the first 5 hr after the treatment, followed by a gradual decrease in cell population due to cell apoptosis. It is also observed that the As(III)-induced cell fusion is much less dose-dependent at the given dose range. Unlike the As(III)-treated cells, the chromium-treated and mercury-treated ones do not show a sharp initial increase in cell numbers. Yet, the initial toxic effects on the cells are quite different between these two toxic compounds; chromium (VI) causes a relatively slower cell-killing effect right after the toxicant exposure. The chromium-induced gradual cell death has led to the slowly declining CI values after the toxicant treatment, which is in fact dose-dependent. Due to the cell necrosis and quick apoptosis, however, the mercury-induced cytotoxicity has resulted in a quick decrease in the CI values with a strict dose-dependency.

It has been pointed out that apoptosis may undergo several complicated biochemical and morphological processes, which have not been well understood (Huang and Xing, 2006). For instance, the initial cell fusion process observed in As (III) experiments (and partially observed in chromium (VI) experiments) remains largely unexplained. Such unknown mechanisms bring challenges to the modeling and prediction of cytotoxicity responses through first-principle approaches. Therefore, it is desired to search for data-driven techniques that can reasonably capture the effect of unknown mechanisms on cytotoxicity responses. To this end, we have achieved significantly better models for the apoptosis-induced cytotoxicity by applying the proposed local ν-SVR modeling technique. Since the lack of accurate predictions in the initial phase can certainly propagate into long-term predictions, the predictive performance over different prediction horizons has been considered.

The prediction curves presented in Figures 3.3, 3.5 and 3.7 represent cell population dynamics in the presence of different toxicants. Unlike the ANNs and the first-principle

models, the SVR-based models are clearly capable of consistently predicting the dose-dependent dynamic cytotoxicity patterns (growth or death) over a considerable time horizon, indicating that the effect of underlying mechanisms have been captured to a certain degree. Consequently, the SVR-based predictive models enable the prediction of transient as well as ultimate cell behavior in the presence of a given toxic compound without the need to wait for completion of the experiment. Since our ultimate objective is the application in an early warning system, reasonably predicting the dose-dependent dynamic cytotoxicity patterns (growth or death) over a considerable time horizon is a significant contribution in this context.

3.6.3 Model Reproducibility

An essential assumption on the use of data-driven models is that the developed models generate consistent and reproducible results under appropriate conditions. In many cases, the lack of a reproducible result may act to limit the situations to which a model may apply. Therefore, reproducibility of the SVR-based model shall be tested to further check the validity of the proposed approach. The reliability and reproducibility of the SVR-based model is evaluated by comparing the prediction results for two different runs of chromium (VI) toxicant experiment conducted under identical conditions on January 20 and 27, 2010. Two data-sets consisting of experimental data corresponding to 0, 4.94, 7.40, 11.1, 16.66, 24.99 and 37.48 μM of chromium (VI) are considered. One data-set is used to train the model, while the other one is reserved to assess the robustness of the developed model. Figure 3.8 shows comparison of model fit based on one-step-ahead prediction for chromium (VI) toxicant, while Figure 3.9 shows comparison of model fit based on five-step-ahead prediction. Although the two experiments do not produce the identical results, it is observed that the model developed based on first experimental data-set can predict the dynamic behavior of CI in the second experiment. CI prediction results for $t = 4 \ldots 25$ based on first three CI measurements are next compared in Figure 3.10. Even though the

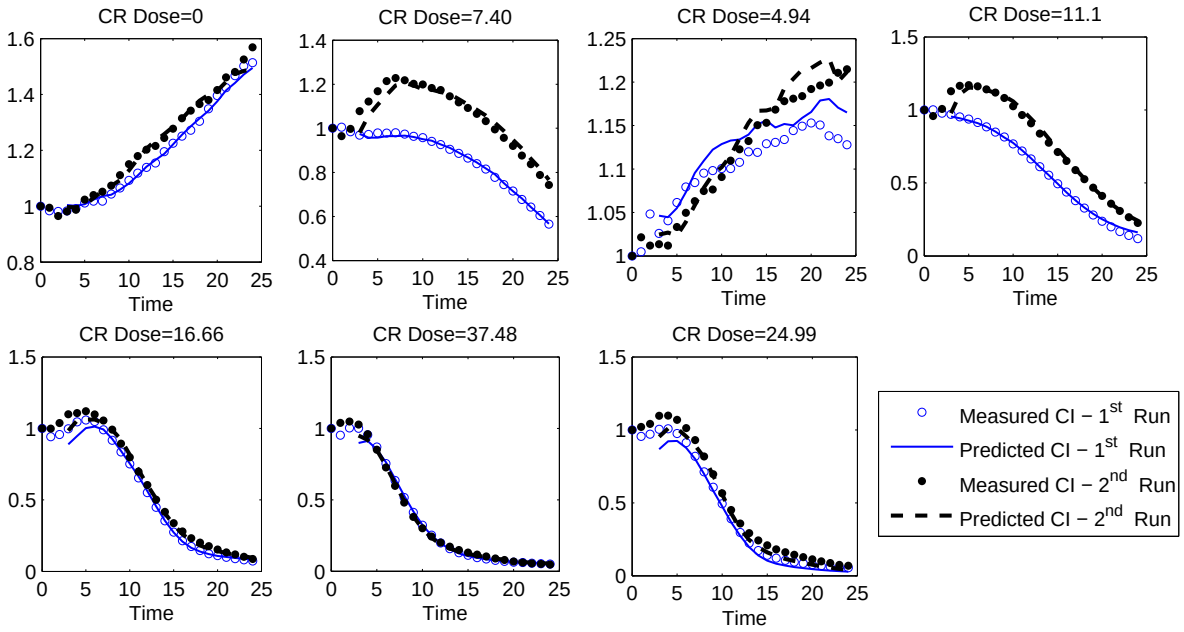


Figure 3.8: Model fit based on one-step-ahead prediction for repeated chromium (VI) toxicant experiment; solid line and dashed line correspond to CI prediction for 1^{st} and 2^{nd} run, respectively; circle and point are actual measurements of CI in 1^{st} and 2^{nd} run, respectively.

model only includes the very first few measurements, a good agreement is achieved for varying-horizon predictions.

3.7 Conclusion

In this study, we have considered dynamic modeling and prediction of cytotoxicity induced by certain water contaminants. A real-time cell electronic sensing (RT-CES) system has been used for conducting cytotoxicity experiments and obtaining CI measurements. Due to the limited understanding of the biochemical and morphological processes involved, the focus of this work was on development of data-driven predictive models. However, the highly non-linear nature of underlying mechanisms would greatly deteriorate the prediction performance of global data-driven models as prediction horizon increases. To address this issue, we developed an inferential framework to capture the non-linearity in the local region around a query point in a real-time manner. The ν-SVR model was selected to form the core

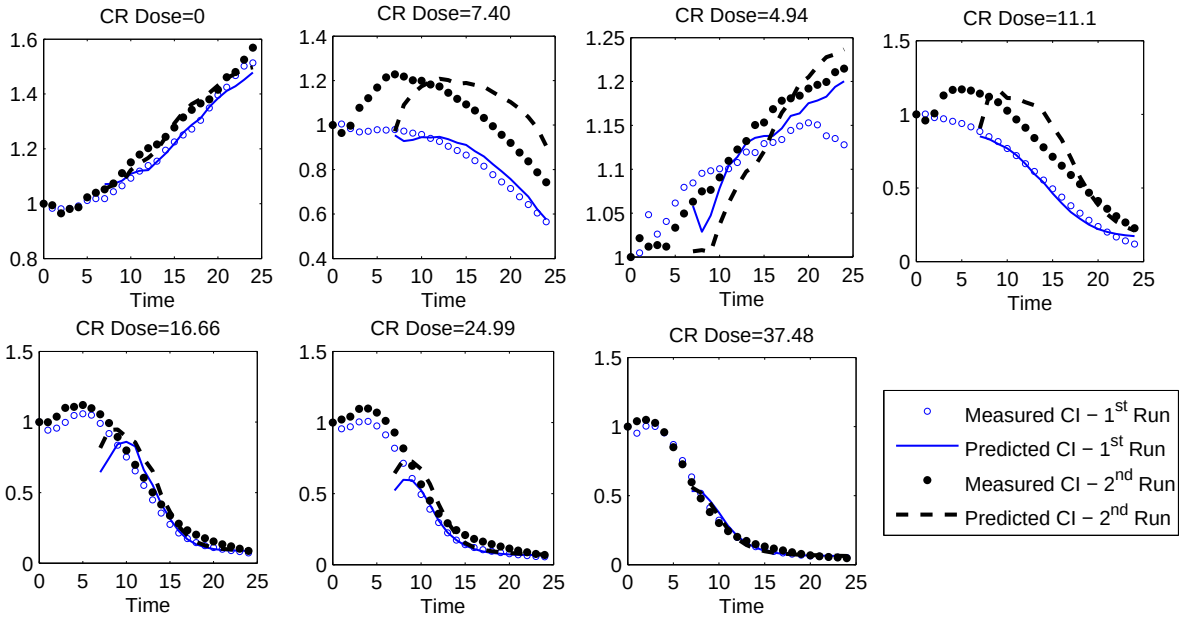


Figure 3.9: Model fit based on five-step-ahead prediction for repeated chromium (VI) toxicant experiment.

of the predictive framework. The ν-SVR algorithm has been chosen, because it is capable to automatically adjusting the width of ε-insensitive tube. The prediction performance of the developed models has been verified on the validation data. Moreover, optimal multi-step-ahead predictions have been obtained and compared with the measured CI values. The long-term dynamic prediction of cytotoxicity based on the first three CI measurements also has been investigated. We examined the reproducibility of the identified SVR-based models on the chromium (VI) experimental data. It has been concluded that the identified models can reproduce the measured CI remarkably well. In summary, the local ν-SVR predictor has some notable advantages in comparison with the ANN approach presented in Huang and Xing (2006). The short-term prediction performance of the local ν-SVR models is superior to that of the ANN. Finally, it has been observed that the SVR-based models are more robust to the increase of the prediction horizon.

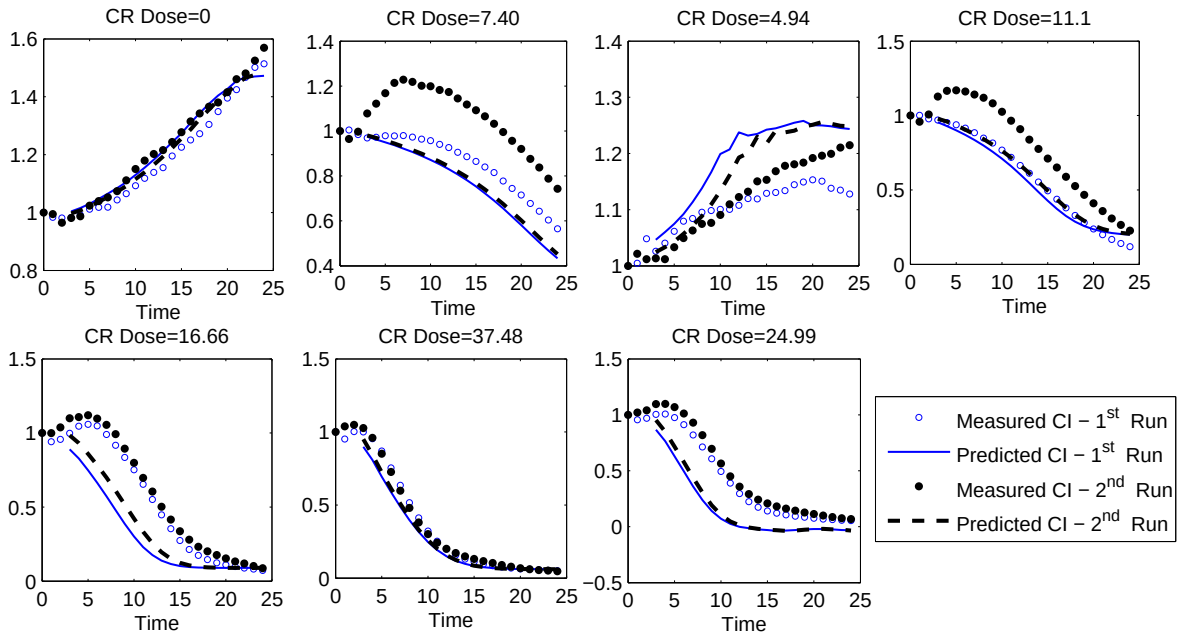


Figure 3.10: Model fit based on first three CI measurements for repeated chromium (VI) toxicant experiment.

Bibliography

Atkeson, C. G., A. W. Moore and S. Schaal (1997). Locally weighted learning. *Artificial Intelligence Review* **11**, 11–73.

Botham, P. A. (2004). Acute systemic toxicity - prospects for tiered testing strategies. *Toxicology in Vitro* **18**(2), 227–230.

Brosnan, T. M. (1999). Early warning monitoring to detect hazardous events in water supplies. Technical report. International Life Sciences Institute. Washington, USA.

Cherkassky, V. and Y. Ma (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks* **17**(1), 113–126.

Clark, R., W. Grayman and J. Goodrich (1986). Toxic screening models for water supply. *Journal of Water Resources Planning and Management* **112**(2), 149–165.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**(368), 829–836.

Fernández, R. (1999). Predicting time series with a local support vector regression machine. In: *Proceedings Of The ECCAI Advanced Course On Artificial Intelligence (ACAI)*. Chania, Greece.

Han, L., L. Ding, J. Yu, Q. Li and Y. Liang (2004). Power plant boiler air preheater hot spots detection system based on least square support vector machines. In: *Advances in*

Neural Networks - ISNN 2004 (F. Yin, J. Wang and C. Guo, Eds.). Vol. 3173 of *Lecture Notes in Computer Science*. pp. 598–604. Springer-Verlag. Berlin, Germany.

Hangos, K. M. and I.T. Cameron (2001). *Process Modelling and Model Analysis*. first ed.. Academic Press. San Diego, USA.

Huang, B. and J. Z. Xing (2005). Dynamic modeling, prediction and analysis of cytotoxicity on microelectronic sensors. In: *Proceedings of the Second International Conference on Fuzzy Systems and Knowledge Discovery - Volume Part II* (L. Wang and Y. Jin, Eds.). Springer-Verlag. Changsha, China. pp. 265–274.

Huang, B. and J. Z. Xing (2006). Dynamic modelling and prediction of cytotoxicity on microelectronic cell sensor array. *The Canadian Journal of Chemical Engineering* **84**(4), 393–405.

Ibrahim, F., B. Huang, J. Z. Xing and S. Gabos (2010). Early determination of toxicant concentration in water supply using MHE. *Water Research* **44**(10), 3252–3260.

Kano, M. and K. Fujiwara (2013). Virtual sensing technology in process industries: Trends and challenges revealed by recent industrial applications. *Journal of Chemical Engineering of Japan* **46**(1), 1–17.

Khatibisepehr, S., B. Huang, F. Ibrahim, J. Z. Xing and W. Roa (2011*a*). Data-based modeling and prediction of cytotoxicity induced by contaminants in water resources. *Computational Biology and Chemistry* **35**(2), 69–80.

Khatibisepehr, S., F. Ibrahim, J. Z. Xing, W. Roa and B. Huang (2011*b*). Data-based modeling and prediction of cytotoxicity on microelectronic sensors. In: *Proceedings of the* 4^{th} *International Symposium on Advanced Control of Industrial Processes (ADCONIP)*. IEEE. Hangzhou, China. pp. 169–174.

Ljung, L. (1999). *System Identification - Theory For the User*. second ed.. Prentice Hall. Upper Saddle River, USA.

Mazijk, A. V. (1996). One-Dimensional Approach Of Transport Phenomena Of Dissolved Matter In Rivers. PhD thesis. Delft University of Technology. Delft, Netherlands.

Schölkopf, B., A. J. Smola, R. C. Williamson and P. L. Bartlett (2000). New support vector algorithms. *Neural Computation* **12**(5), 1207–1245.

Smola, A. J. and B. Schölkopf (2004). A tutorial on support vector regression. *Statistics and Computing* **14**(3), 199–222.

Vapnik, V. (1998). *Statistical Learning Theory*. first ed.. Wiley-Interscience. New York, USA.

Vapnik, V. (1999). *The Nature Of Statistical Learning Theory*. second ed.. Springer-Verlag. New York, USA.

Vapnik, V. and A. Chervonenkis (1974). *Theory of Pattern Recognition (in Russian)*. second ed.. Springer-Verlag. Nauka, Russia.

Xing, J. Z., L. Zhu, J. A. Jackson, S. Gabos, X. J. Sun, X. B. Wang and X. Xu (2005). Dynamic monitoring of cytotoxicity on microelectronic sensors. *The Canadian Journal of Chemical Engineering* **18**(2), 154–161.

Yan, W., H. Shao and X. Wang (2004). Soft sensing modeling based on support vector machine and Bayesian model selection. *Computers and Chemical Engineering* **28**(8), 1489–1498.

Yang, W., J. Nan and D. Sun (2008). An online water quality monitoring and management system developed for the liming river basin in Daqing, China. *Journal of Environmental Management* **88**(2), 318–325.

Zheng, Q. and H. Kimura (2001). Just-in-time modeling for function prediction and its applications. *Asian Journal of Control* **3**(1), 35–44.

Chapter 4

A Bayesian Framework for Model Structure Selection and Hyperparameters Tuning in Locally Weighted Partial Least Squares Regression

4.1 Introduction

4.1.1 Practical Motivation

Process variables are often causally related and, consequently, process measurements are strongly collinear. From the inferential modeling point of view, the collinear measurements of dependent input variables provide little independent information. Partial least squares (PLS) regression can effectively handle the collinear identification data, while classical model identification techniques such as ordinary least squares (OLS) regression may result in an ill-conditioned inferential model (Marjanovic *et al.*, 2006; Mobaraki and Hemmateenejad, 2011; Lin and Jørgensen, 2011). The PLS regression provides a robust solution to the collinearity problem by projecting process variables into a lower number of orthogonal latent variables (Lin *et al.*, 2007). The inherent limitation of this method is the linearity assumption. Therefore, the PLS models become ill-suited for non-linear processes with non-Gaussian disturbances (Yu, 2012). Moreover, if process operations

deviate from the previously identified operating envelope, the prediction performance of the PLS models would deteriorate. Locally weighted partial least squares (LW-PLS) is an adaptive modeling technique that has been widely adopted to cope with variations in process characteristics and handle non-linearity of underlying mechanisms (Schaal *et al.*, 2002; Kim *et al.*, 2011). The basic idea behind the LW-PLS is to identify a local PLS model at a certain operating point by prioritizing the identification samples. Given a query sample, the problem of real-time identification of LW-PLS models involves the following main steps:

1. **Selection of nearest neighbors**: The search for the nearest neighbors is carried out from the historical data-set using a pre-defined notion of similarity or locality. The similarity function is often characterized by a set of hyperparameters that determine the size and shape of region of validity of each local model, also known as receptive field.

2. **Selection of model structure**: Having quantified the similarity between the query and identification samples, one of the key steps in the identification procedure is to find a suitable model structure that best approximates the local behavior of the underlying process. In order to select a proper structure for the LW-PLS models, it is required to find influential input variables and determine the optimal number of latent variables. The choice of the influential input variables and the number of retained latent variables affect the model complexity.

3. **Estimation of model parameters**: Once the structure of the local PLS model is chosen, the LW-PLS algorithm can be used to calculate the loading and score matrices.

It is desired to parameterize the similarity function, select model structure, and estimate model parameters in a real-time manner for identification of LW-PLS models. Several

optimization methods can be applied to obtain the optimal combination of the influential input variables, number of retained latent variables, and hyperparameter values resulting in the lowest root mean squared error of cross-validation (RMSECV) (Perez-Guaita *et al.*, 2013). Many of these methods are often not computationally feasible for real-time implementation (Kim *et al.*, 2013). It is common practice to find the globally optimal structure of the local PLS models as well as estimates of the similarity function hyperparameters in an off-line identification phase. In this way, the computational costs imposed by real-time model structure selection and similarity function parametrization can be avoided. However, variations in process characteristics and non-linearity of underlying mechanisms may not only affect model parameters, but also functional forms and size of receptive fields. Furthermore, evaluating the RMSECV would often lead to the over-fitting phenomenon (Shao, 1993; Ljung, 1999). Kim *et al.* (2013) have pointed out that in industrial applications the trade-off between the model complexity and prediction performance is decided upon by taking into account the experiences and expertise of the plant experts. Thus, there is need to develop a reliable systematic method for selection of the optimal LW-PLS model structure and its region of validity.

4.1.2 Main Contributions

Motivated by the above considerations, this paper presents a novel and computationally feasible Bayesian approach to address the aforementioned issues. It is assumed that the operating space can be partitioned into a finite number of sub-spaces. For each sub-space, the problem of finding the locally optimal LW-PLS model structure and similarity function hyperparameters is formulated under an iterative hierarchical Bayesian optimization framework (MacKay, 2002). The real-time identification problem thus amounts to detecting the underlying operating sub-space and estimating the LW-PLS model parameters. The proposed method has the following attractive features:

1. The Bayesian model comparison allows us to perform objective comparisons between alternative model structures. Therefore, the resulting optimization problem in each sub-space can automatically deal with the model complexity control to avoid over-fitting.

2. Objective criteria for local tuning of the hyperparameters of the similarity function are provided.

3. Real-time model structure selection and similarity function parametrization would become computationally efficient.

4.2 Problem Statement

Let $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ denote the identification data-set. The general form of a PLS model is given by

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_x \tag{4.1}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^T + \mathbf{e}_y \tag{4.2}$$

where $\mathbf{X} \in \mathbb{R}^{N\times K}$ and $\mathbf{y} \in \mathbb{R}^{N\times 1}$ are input and output matrices, respectively. $\mathbf{T} \in \mathbb{R}^{N\times L}$ and $\mathbf{P} \in \mathbb{R}^{K\times L}$ are score and loading matrices, respectively. $\mathbf{q} \in \mathbb{R}^{1\times L}$ is the vector of regression coefficients. $\mathbf{E}_x \in \mathbb{R}^{N\times K}$ and $\mathbf{e}_y \in \mathbb{R}^{N\times 1}$ are additive Gaussian noise terms with time-varying variance.

Given a query sample $\mathbf{x}_q$, a similarity matrix is constructed to prioritize the identification samples:

$$\mathbf{S}_q(\Phi) = \text{diag}(s_{1|q}(\Phi), \cdots, s_{N|q}(\Phi)) \tag{4.3}$$

where $\Phi = \{\phi_1, \cdots, \phi_D\}$ are the hyperparameters of the similarity function and $s_{n|q}$ is the similarity between $\mathbf{x}_q$ and $\mathbf{x}_n$.

Once a similarity matrix is specified, the task is to identify a LW-PLS model. It can be

shown that

$$\begin{aligned}\hat{y}_q &= \mathbf{x}_q \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{q}^T \\ &= \mathbf{x}_q \Theta_q \end{aligned} \tag{4.4}$$

where the columns of $\mathbf{W} \in \mathbb{R}^{K\times L}$ are orthonormal weight vectors.

The l^{th} column of $\mathbf{W}$, $\mathbf{P}$, and $\mathbf{q}$ can be estimated from the following expressions (Kim *et al.*, 2011):

$$\mathbf{w}_l = \frac{\left(\mathbf{X} - \sum_{j=1}^{l-1} \mathbf{t}_j \mathbf{p}_j^T\right)^T \mathbf{S}_q \left(\mathbf{y} - \sum_{j=1}^{l-1} \mathbf{t}_j q_j\right)}{\left\| \left(\mathbf{X} - \sum_{j=1}^{l-1} \mathbf{t}_j \mathbf{p}_j^T\right)^T \mathbf{S}_q \left(\mathbf{y} - \sum_{j=1}^{l-1} \mathbf{t}_j q_j\right) \right\|_2} \tag{4.5}$$

$$\mathbf{p}_l = \frac{\left(\mathbf{X} - \sum_{j=1}^{l-1} \mathbf{t}_j \mathbf{p}_j^T\right)^T \mathbf{S}_q \mathbf{t}_l}{\mathbf{t}_l^T \mathbf{S}_q \mathbf{t}_l} \tag{4.6}$$

$$q_l = \frac{\left(\mathbf{y} - \sum_{j=1}^{l-1} \mathbf{t}_j q_j\right)^T \mathbf{S}_q \mathbf{t}_l}{\mathbf{t}_l^T \mathbf{S}_q \mathbf{t}_l} \tag{4.7}$$

where $\mathbf{t}_l$ denotes the l^{th} column of $\mathbf{T}$ (*i.e.* the l^{th} latent vector) and is calculated as

$$\mathbf{t}_l = \left(\mathbf{X} - \sum_{j=1}^{l-1} \mathbf{t}_j \mathbf{p}_j^T\right) \mathbf{w}_l \tag{4.8}$$

In general, the identification problem is to tune the hyperparameters of the similarity function, Φ, select the structure of the LW-PLS model, $\mathcal{H}$, and estimate the underlying parameters, Θ. There are several algorithms for efficiently estimating the parameters of the local PLS models (Wold, 1966; Chen *et al.*, 2007; Chun and Keleş, 2010; Kim *et al.*, 2011; Perez-Guaita *et al.*, 2013). However, model structure selection and similarity function parametrization are computationally expensive for real-time implementation. Commonly, the structure of the local PLS models as well as the hyperparameters of the similarity function are globally optimized in an off-line identification phase. Algorithm 4.1 presents a widespread optimization technique in which the RMSE of the leave-one-out cross-validation (LOOCV) is regarded as the cost function (Wold *et al.*, 1984; Atkeson *et al.*, 1997; Kim *et al.*, 2013).

Algorithm 4.1. *Classical Cross-validation Procedure for Model Structure Selection and Similarity Function Parametrization*

1. Choose an appropriate similarity function.

2. Determine a search space for the hyperparameters of the similarity function. Often, continuous hyperparameters are discretized to construct a finite set of reasonable values for each hyperparameter. Therefore, the search is performed over $\Phi \in \{\Phi_1, \cdots, \Phi_F\}$ with $\Phi_f = \{\phi_{f,1}, \cdots, \phi_{f,D}\}$.

3. Specify a set of plausible model structures $\mathcal{H} \in \{\mathcal{H}_1, \cdots, \mathcal{H}_L\}$.

4. For each set of $(\Phi_f, \mathcal{H}_l)$, repeat the following steps:

   4.1. Let $\mathcal{D}_{-q}$ denote the identification data-set consisting of all the training samples except $(\mathbf{x}_q, y_q)$, *i.e.* $\mathcal{D}_{-q} \triangleq \mathcal{D} \backslash \{(\mathbf{x}_q, y_q)\}$. In this way, Θ_q parameterizes the LW-PLS model identified from $\mathcal{D}_{-q}$. Complete the following steps for each training sample $(\mathbf{x}_q, y_q) \in \mathcal{D}$: (1) Construct the similarity matrix $\mathbf{S}_q(\Phi_f)$ (Equation 4.3). (2) Identify a LW-PLS model with $\mathcal{H}_l$ structure from $\mathcal{D}_{-q}$ (Equations 4.5-4.8). (3) Use the identified LW-PLS model to calculate Θ_q and obtain a prediction of y_q denoted as $\hat{y}_q^{cv}$ (Equation 4.4).

   4.2. Calculate the corresponding RMSECV as follows:

$$\text{RMSECV}_{f,l} = \sqrt{\frac{\sum_{q=1}^{N}(y_q - \hat{y}_q^{cv})^2}{N}} \tag{4.9}$$

5. Select the model structure and hyperparameters that minimize the RMSECV as the globally optimal setting.

Since evaluating the RMSECV is prone to over-fitting, the globally optimal setting obtained from Algorithm 4.1 is not always reliable in industrial applications. In such cases, the experience of the engineers needs to be taken into account to compromise between

the robustness and prediction performance. This task is performed in a rather *ad hoc* manner. Therefore, a reliable systematic method for LW-PLS model structure selection and similarity function parametrization is yet to be developed.

4.3 Hierarchical Bayesian Optimization Framework

A Bayesian learning approach converts the problem of interest into an equivalent problem of evaluating the joint posterior probability density function (PDF) of the model structure and similarity function hyperparameters, $P\big(\Phi, \mathcal{H}|\mathcal{D}, \{\Theta_q\}_{q=1}^N\big)$. Evaluating the joint posterior PDF provides a systematic method for selecting the structure of the LW-PLS model as well as an objective criterion for parameterizing the similarity function. Such Bayesian approach can automatically deal with the model complexity control issue to avoid over-fitting (Guyon *et al.*, 2010). To circumvent the difficulties associated with direct maximization of $P\big(\Phi, \mathcal{H}|\mathcal{D}, \{\Theta_q\}_{q=1}^N\big)$, the problem is formulated and solved under an iterative hierarchical Bayesian optimization framework (MacKay, 1992, 2002). First, the chain rule of probability theory is used to factorize the joint posterior PDF as

$$P\big(\Phi, \mathcal{H}|\mathcal{D}, \{\Theta_q\}_{q=1}^N\big) \propto P\big(\Phi|\mathcal{H}, \mathcal{D}, \{\Theta_q\}_{q=1}^N\big) P\big(\mathcal{H}|\mathcal{D}, \{\Theta_q\}_{q=1}^N\big) \tag{4.10}$$

Then, the optimization problem is decomposed hierarchically into two layers:

$$\begin{aligned} &\max_{\Phi,\mathcal{H}} P\big(\Phi|\mathcal{H}, \mathcal{D}, \{\Theta_q\}_{q=1}^N\big) P\big(\mathcal{H}|\mathcal{D}, \{\Theta_q\}_{q=1}^N\big) \\ &\qquad = \max_{\mathcal{H}} \Big\{ P\big(\mathcal{H}|\mathcal{D}, \{\Theta_q\}_{q=1}^N\big) \max_{\Phi} \big\{ P\big(\Phi|\mathcal{H}, \mathcal{D}, \{\Theta_q\}_{q=1}^N\big) \big\} \Big\} \end{aligned} \tag{4.11}$$

4.3.1 Inference of Hyperparameters of the Similarity Function

Applying Bayes' rule, the posterior PDF of hyperparameters can be expressed as

$$P\big(\Phi|\mathcal{H}, \mathcal{D}, \{\Theta_q\}_{q=1}^N\big) \propto P\big(\mathcal{D}|\Phi, \mathcal{H}, \{\Theta_q\}_{q=1}^N\big) P\big(\Phi|\mathcal{H}, \{\Theta_q\}_{q=1}^N\big) \tag{4.12}$$

Table 4.1: Some properties of gamma distribution

Property	Explicit Form
Parameter Restriction	$b_d > 0$ and $\beta > 0$
Mode	$(b_d - 1)/\beta$
Mean	b_d/β
Variance	b_d/β^2
Skewness	$2/\sqrt{b_d}$

As priors, it is reasonable to assume that the structure/parameters of the LW-PLS model and hyperparameters of the similarity function are statistically independent:

$$P\big(\Phi|\mathcal{H}, \{\Theta_q\}_{q=1}^{N}\big) = P(\Phi) \tag{4.13}$$

In the absence of explicit background information, non-informative priors are often specified in the form of uniform distributions. To incorporate the available prior knowledge, conjugate priors are commonly assigned for which the resulting posterior distributions can be conveniently evaluated. Since the gamma distribution is the conjugate prior to many likelihood functions, the prior distribution over Φ can be represented by

$$\begin{aligned} P(\phi_d) &= \frac{\beta_d^{b_d}}{\Gamma(b_d)} \phi_d^{b_d-1} \exp\big(-\beta_d\phi_d\big) \\ &\propto \phi_d^{b_d-1} \exp\big(-\beta_d\phi_d\big) \end{aligned} \tag{4.14}$$

where b_d is the shape parameter and β_d is the rate parameter for the d^{th} hyperparameter, ϕ_d.

Hence,

$$P(\Phi) \propto \prod_{d=1}^{D} \phi_d^{b_d-1} \exp\big(-\beta_d\phi_d\big) \tag{4.15}$$

Table 4.1 gives a summary of the properties of gamma distribution.

In the case of conditionally independent observations, the LOOCV predictive likelihood in Equation 4.12, also known as the pseudo-likelihood, can be expressed as (Sundararajan

and Keerthi, 2001)

$$\begin{aligned} P\big(\mathcal{D}|\Phi,\mathcal{H},\{\Theta_q\}_{q=1}^N\big) &\propto \prod_{q=1}^{N} P\big(y_q|\mathbf{x}_q,\mathcal{H},\Phi,\Theta_q\big) \\ &= \prod_{q=1}^{N} \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(y_q-\hat{y}_q^{cv})^2}{2\sigma_q^2}\right) \\ &= \exp\left(-\sum_{q=1}^{N}\frac{(y_q-\hat{y}_q^{cv})^2}{2\sigma_q^2}\right)\prod_{q=1}^{N}\frac{1}{\sqrt{2\pi\sigma_q^2}} \end{aligned} \tag{4.16}$$

where the mean and variance of the predictive distribution for y_q are obtained as follows:

$$\hat{y}_q^{cv} = \mathbf{x}_q\Theta_q \tag{4.17}$$

$$\sigma_q^2(\Phi) = \frac{\sum\limits_{n=1,\ n\neq q}^{N} s_{n|q}(\Phi)\big(y_n-\mathbf{x}_n\Theta_q\big)^2}{\sum\limits_{n=1,\ n\neq q}^{N} s_{n|q}(\Phi)} \tag{4.18}$$

Substituting Equations 4.15 and 4.16 into Equation 4.12, the posterior PDF over hyperparameters of the similarity function is given by

$$\begin{aligned} P\big(\Phi|\mathcal{H},\mathcal{D},\{\Theta_q\}_{q=1}^N\big) &\propto P\big(\mathcal{D}|\Phi,\mathcal{H},\{\Theta_q\}_{q=1}^N\big)P\big(\Phi|\mathcal{H},\{\Theta_q\}_{q=1}^N\big) \\ &\propto \exp\Big(-\sum_{q=1}^{N}\frac{(y_q-\hat{y}_q^{cv})^2}{2\sigma_q^2}-\sum_{d=1}^{D}\beta_d\phi_d\Big)\prod_{q=1}^{N}\sigma_q^{-1}\prod_{d=1}^{D}\phi_d^{b_d-1} \end{aligned} \tag{4.19}$$

The MAP estimates of Φ can be obtained by solving the following optimization problem:

$$\Phi^{\text{MP}} = \underset{\Phi}{\operatorname{argmin}} \sum_{d=1}^{D}\Big[(1-b_d)\log\phi_d+\beta_d\phi_d\Big]+\sum_{q=1}^{N}\Big[\frac{(y_q-\hat{y}_q^{cv})^2}{2\sigma_q^2}+\log\sigma_q\Big] \tag{4.20}$$

The above optimization problem can be solved using either discrete or continuous search methods such as gradient descent and grid search approaches (Atkeson *et al.*, 1997; Kolda *et al.*, 2003; Sra *et al.*, 2011).

4.3.2 Inference of Model Structure

Applying Bayes' rule, the posterior PDF of the LW-PLS model structure can be expressed as

$$P\big(\mathcal{H}|\mathcal{D},\{\Theta_q\}_{q=1}^N\big) \propto P\big(\mathcal{D}|\mathcal{H},\{\Theta_q\}_{q=1}^N\big)P\big(\mathcal{H}|\{\Theta_q\}_{q=1}^N\big) \tag{4.21}$$

Suppose that a set of candidate model structures are given, *i.e.* $\mathcal{H} \in \{\mathcal{H}_1,\cdots,\mathcal{H}_L\}$. The random variable $\mathcal{H}$ is a categorical variable and can be modeled by

$$P(\mathcal{H}) = \prod_{l=1}^{L} P(\mathcal{H}=\mathcal{H}_l)^{[\mathcal{H}=\mathcal{H}_l]} \tag{4.22}$$

where $[\mathcal{H}=\mathcal{H}_l]$ evaluates to 1 if $\mathcal{H}=\mathcal{H}_l$ and evaluates to 0 otherwise. The above prior distribution reflects the prior knowledge of plausibility of the alternative models. In the absence of any prior information, a uniform distribution, *i.e.* $P(\mathcal{H}=\mathcal{H}_1)=\cdots=P(\mathcal{H}=\mathcal{H}_L)$, will suffice (Chipman *et al.*, 2001). Even with uniform prior distribution over plausible model structures, the posterior distribution in Equation 4.21 naturally penalizes model complexity.

The Bayesian inference of model structure requires evaluating the model evidence, $P\big(\mathcal{D}|\mathcal{H},\{\Theta_q\}_{q=1}^N\big)$. This likelihood function, also known as the pseudo-marginal likelihood, can be obtained by integrating over the hyperparameters of the similarity function:

$$P\big(\mathcal{D}|\mathcal{H},\{\Theta_q\}_{q=1}^N\big) = \int P\big(D|\Phi,\mathcal{H},\{\Theta_q\}_{q=1}^N\big)P\big(\Phi|\mathcal{H},\{\Theta_q\}_{q=1}^N\big)d\Phi \tag{4.23}$$

There are a variety of methods available to analytically evaluate or approximate the above integral (MacKay, 2002; Penny *et al.*, 2006). For instance, the model evidence can be approximated using Laplace's method, under certain assumptions (Kass and Raftery, 1995; MacKay, 1999):

$$P\big(\mathcal{D}|\mathcal{H},\{\Theta_q\}_{q=1}^N\big) \approx P\big(D|\Phi^{\text{MP}},\mathcal{H},\{\Theta_q\}_{q=1}^N\big)P\big(\Phi^{\text{MP}}|\mathcal{H},\{\Theta_q\}_{q=1}^N\big)\det(\mathbf{A}/2\pi)^{-\frac{1}{2}} \tag{4.24}$$

Table 4.2: Interpretation of Bayes factors

Bayes Factor	$P(\mathcal{H}=\mathcal{H}_l\|\mathcal{D})$	Evidence Supporting Model $\mathcal{H}_l$
$1-3$	$50-75\%$	Weak
$3-20$	$75-95\%$	Positive
$20-150$	$95-99\%$	Strong
≥ 150	$\geq 99\%$	Very Strong

where $\mathbf{A} = -\nabla\nabla \log P\big(\Phi|\mathcal{H},\mathcal{D},\{\Theta_q\}_{q=1}^N\big)$. In the Laplace's method of approximation it is assumed that $P\big(\Phi|\mathcal{H},\mathcal{D},\{\Theta_q\}_{q=1}^N\big)$ is highly peaked around Φ^{MP}.

Substituting Equations 4.22 and 4.24 into Equation 4.21, the alternative models can be ranked by evaluating the following posterior probability:

$$P\big(\mathcal{H}=\mathcal{H}_l|\mathcal{D},\{\Theta_q\}_{q=1}^N\big) \propto \left[\exp\Big(-\sum_{q=1}^{N}\frac{(y_q-\hat{y}_q^{cv})^2}{2\sigma_q^2}-\sum_{d=1}^{D}\beta_d\phi_d\Big)\prod_{q=1}^{N}\sigma_q^{-1}\prod_{d=1}^{D}\phi_d^{b_d-1}\right]_{\Phi=\Phi^{\text{MP}}} \times \det(\mathbf{A}/2\pi)^{-\frac{1}{2}}P(\mathcal{H}=\mathcal{H}_l) \tag{4.25}$$

The most probable model structure can be selected as the one with the largest posterior probability calculated from Equation 4.25:

$$\mathcal{H}^{\text{MP}} = \underset{\mathcal{H}}{\operatorname{argmax}}\ P\big(\mathcal{H}|\mathcal{D},\{\Theta_q\}_{q=1}^N\big) \tag{4.26}$$

Moreover, pairwise comparison of models $\mathcal{H}_l$ and $\mathcal{H}_j$ can be summarized by the posterior odds:

$$\underbrace{\frac{P(\mathcal{H}_l|\mathcal{D},\{\Theta_q\}_{q=1}^N)}{P(\mathcal{H}_j|\mathcal{D},\{\Theta_q\}_{q=1}^N)}}_{\text{Posterior odds}} = \underbrace{\frac{P(\mathcal{D}|\mathcal{H}_l,\{\Theta_q\}_{q=1}^N)}{P(\mathcal{D}|\mathcal{H}_j,\{\Theta_q\}_{q=1}^N)}}_{\text{pseudo-Bayes factor}}\underbrace{\frac{P(\mathcal{H}_l)}{P(\mathcal{H}_j)}}_{\text{Prior odds}} \tag{4.27}$$

The ratio of the pseudo-marginal likelihoods is a surrogate for the Bayes factor and is thus known as the pseudo-Bayes factor (PsBF) (Geisser and Eddy, 1979; Gelfand and Dey, 1994). The pseudo-Bayes factor is a summary of the information provided by the data about the plausibility of the alternative model structures (Kass and Raftery, 1995).

As shown in Table 4.2, Bayes factors have been classified into different ranges in order to evaluate the strengths of evidence assuming uniform priors over model structures (Penny *et al.*, 2006). It should be highlighted that Bayesian model selection naturally penalizes model complexity, thus preventing the over-fitting phenomenon (MacKay, 2002).

To summarize our discussion, the implementation procedure of the proposed hierarchical Bayesian approach is outlined in Algorithm 4.2.

Algorithm 4.2. *Hierarchical Bayesian Procedure for Model Structure Selection and Similarity Function Parametrization*

1. Choose an appropriate similarity function.

2. Characterize the prior distribution of hyperparameters, $P(\Phi)$, based on the explicit prior knowledge. The available prior information over hyperparameters can be represented by gamma distributions (Equations 4.14 and 4.15). If there is no explicit information available for the hyperparameters, a uniform distribution can then be used to describe appropriate non-informative priors.

3. Specify an ordered set of plausible model structures $\mathcal{H} \in \{\mathcal{H}_1, \cdots, \mathcal{H}_L\}$.

4. Characterize the prior distribution over model structures, $P(\mathcal{H})$. The prior knowledge of plausibility of alternative model structures can be generally well-represented by categorical distributions (Equation 4.22). In the absence of any prior information, a uniform prior distribution can be assumed.

5. Specify a suitable threshold for comparing the plausible model structures in terms of posterior odds which are the ratio of posterior probabilities. Table 4.2 can guide the choice of an appropriate threshold in the case of uniform priors.

6. Select the model $\mathcal{H}_l$. Choose a set of initial values for hyperparameters of the similarity function, $\Phi_l^{[0]}$. Repeat the following steps iteratively until no further improvements are gained:

6.1. Complete the following steps for each training sample $(\mathbf{x}_q, y_q) \in \mathcal{D}$: (1) Construct the similarity matrix $\mathbf{S}_q(\Phi_l^{[k]})$ (Equation 4.3). (2) Identify a LW-PLS model with $\mathcal{H}_l$ structure from $\mathcal{D}_{-q}$ (Equations 4.5-4.8). (3) Use the identified LW-PLS model to calculate $\Theta_q^{\{k\}}$ and obtain predictions of $\{y_1, \cdots, y_q, \cdots, y_N\}$ denoted as $\{\hat{y}_{1|q}, \cdots, \hat{y}_q^{cv}, \cdots, \hat{y}_{N|q}\}$ (Equation 4.4).

6.2. Maximize $P\big(\Phi_l|\mathcal{H}_l, \mathcal{D}, \{\Theta_q^{[k]}\}_{q=1}^N\big)$, or equivalently, minimize its negative logarithm to update the MAP estimates of hyperparameters, $\Phi_l^{[k+1]}$ (Equations 4.19 and 4.20).

7. Repeat Step 6.2 given Φ_l^{MP} to obtain $\{\Theta_q^{\text{MP}}\}_{q=1}^N$.

8. Calculate the model evidence $P\big(\mathcal{D}|\mathcal{H}_l, \{\Theta_q^{\text{MP}}\}_{q=1}^N\big)$ (Equation 4.24).

9. Calculate the posterior probability of $\mathcal{H}_l$, $P\big(\mathcal{H} = \mathcal{H}_l|\mathcal{D}, \{\Theta_q^{\text{MP}}\}_{q=1}^N\big)$ (Equation 4.25). If the posterior odds in pairwise comparison of $\mathcal{H}_{l-1}$ and $\mathcal{H}_l$ is greater than a pre-specified threshold, go to step 6 to evaluate the plausibility of $\mathcal{H}_{l+1}$. Otherwise, select $\mathcal{H}_{l-1}$ and Φ_{l-1}^{MP} as the globally optimal setting.

4.4 Adaptive Locally Weighted Partial Least Squares

There are several reasons to consider real-time model structure selection and similarity function parametrization. These include: (1) sparsity and heteroscedasticity of training samples; and (2) variations in process characteristics and non-linearity from of the underlying mechanisms. However, solving such optimization problem in a real-time manner might be computationally expensive. To obtain a computationally feasible solution, we propose an adaptive scheme in which the optimal hyperparameters and model structures are found for different regions of the operating space in the off-line identification phase. Thus, the real-time identification problem amounts to detecting the underlying operating region and estimating the corresponding model parameters.

Suppose that the operating space can be partitioned into a finite number of sub-spaces, *i.e.* M operating modes. An input-output representation of an adaptive, also called **multi-model**, LW-PLS is given by

$$\begin{cases} \hat{y}_q^{(m)} = \mathcal{H}^{(m)}(\mathbf{x}_q, \Theta_q^{(m)}, \Phi^{(m)}) \quad m = 1, \cdots, M \\ \hat{y}_q = \sum_{m=1}^{M} \psi_q^{(m)} \hat{y}_q^{(m)} \end{cases} \tag{4.28}$$

where a proper interpolation function is defined to assign an importance weight, $\psi_q^{(m)}$, to the output of each sub-model, $\hat{y}_q^{(m)}$.

The problem of identification of the multi-model LW-PLS is divided into two phases:

1. **Off-line identification**: First, the operating space is partitioned into M sub-spaces. Next, training samples are attributed to relevant regions based on descriptive classification criteria; the identification data-set is divided into multiple exclusive sub-sets. Let $\mathcal{D}^{(m)} = \{(\mathbf{x}_q, y_q)\}_{q=1}^{N_m}$ denote the set of N_m identification samples that belong to the m^{th} operating sub-space. Finally, the joint posterior PDF of the LW-PLS model structure and similarity function hyperparameters, $P\big(\Phi, \mathcal{H} | \mathcal{D}^{(m)}, \{\Theta_q\}_{q=1}^{N_m}\big)$, is optimized for $m \in \{1, \cdots, M\}$. Thus, the optimization problem for the m^{th} operating region becomes

$$\{\hat{\Phi}^{(m)}, \hat{\mathcal{H}}^{(m)}\} = \underset{\Phi, \mathcal{H}}{\operatorname{argmax}}\, P\big(\Phi, \mathcal{H} | \mathcal{D}^{(m)}, \{\Theta_q\}_{q=1}^{N_m}\big) \tag{4.29}$$

   where $\hat{\Phi}^{(m)}$ and $\hat{\mathcal{H}}^{(m)}$ are the maximum *a posteriori* (MAP) estimates of Φ and $\mathcal{H}$ for the m^{th} operating region.

2. **Real-time identification** Given a query sample, the parameters of the m^{th} LW-PLS sub-model are estimated based on $\hat{\Phi}^{(m)}$ and $\hat{\mathcal{H}}^{(m)}$. The identified sub-models are used to obtain $\hat{y}_q^{(m)}$ for $m \in \{1, \cdots, M\}$. Appropriate importance weights are assigned to $\hat{y}_q^{(m)}$ for $m \in \{1, \cdots, M\}$ in order to obtain a global prediction, $\hat{y}_q$.

4.4.1 Partitioning of the Operating Space

A discrete-state dynamics can be associated with the system under investigation in order to partition the operating space into a finite number of sub-spaces. The discrete-state dynamics may come from switching controllers, inherent non-linearities in the system, or different production policies. In the absence of relevant process information, the operating space can be partitioned by evaluating the residuals obtained from a global model. In such cases, the task is to partition the operating space such that the identified sub-models would be well-supported by the corresponding identification data sub-sets. It is noteworthy that the choice of the number of sub-spaces is a trade-off between representativeness and complexity of the adaptive LW-PLS framework.

Having partitioned the operating space, it is required to select a representative scheduling variable reflecting changes in the operating region. As shown in Equation 4.28, the interpolation function assigns a proper importance weight to the output of each sub-model in the multi-model LW-PLS scheme. The importance weight assigned to $\hat{y}_q^{(m)}$ can be obtained by evaluating the posterior probability of the m^{th} sub-model capturing the process behavior (Khatibisepehr and Huang, 2012):

$$P(m|\nu_q, \mathcal{D}) \propto P(\nu_q|m, \mathcal{D})P(m) \tag{4.30}$$

where ν_q is a set of scheduling variables parameterizing the interpolation function.

$P(\nu_q|m, \mathcal{D})$ is the likelihood that ν_q belongs to the m^{th} operating region. Marginal and joint PDFs of the scheduling variables, influential input variables, and retained latent variables could be investigated in order to specify such likelihood from identification data. $P(m)$ is the prior probability that the system operates in the m^{th} operating region. Background information about the process operation can be used to specify the prior PDFs. It is noteworthy that the choice of suitable scheduling variables is problem specific. All or a sub-set of the latent variables may be investigated as potential scheduling variables.

Depending on the application, the importance weights can be calculated using one of the

following weight functions:

$$\psi_q^{(m)} = \begin{cases} 1 & m = \text{argmax}\, P(m|\nu_q, \mathcal{D}) \\ 0 & \text{Otherwise} \end{cases} \tag{4.31}$$

or

$$\psi_q^{(m)} = \frac{P(\nu_q|m, \mathcal{D})P(m)}{\sum_{m=1}^{M} P(\nu_q|m, \mathcal{D})P(m)} \quad \text{for } m = 1, \cdots, M \tag{4.32}$$

If the importance weights are assigned according to Equation 4.31, the global prediction is simply equal to the output of the LW-PLS sub-model with the highest posterior probability.

The steps required to characterize an adaptive LW-PLS framework are summarized in Algorithm 4.3.

Algorithm 4.3. *Bayesian Procedure for Characterization of an Adaptive LW-PLS Framework*

1. Partition the operating space into M sub-spaces by taking into account the prior knowledge of the nominal operating conditions or evaluating the performance of the global model.

2. Classify training samples to relevant operating regions in order to construct $\mathcal{D}^{(m)} = \{(\mathbf{x}_q, y_q)\}_{q=1}^{N_m}$ for $m = 1, \cdots, M$.

3. Select a representative scheduling variable, ν, that effectively reflects the operating region at each time instant.

4. Determine the likelihood that ν would be generated by the m^{th} operating mode, $P(\nu|m, \mathcal{D})$. Marginal and joint PDFs of the scheduling variables should be investigated in order to specify such conditional PDFs from historical data.

5. Assign the prior probability of the system operating in the m^{th} mode, $P(m)$. Information about the typical process operation can be used to specify the prior PDFs. $P(m)$ can also be viewed as the prior probability that the m^{th} sub-model

captures the process behavior. If such background information is not available or relevant, a uniform prior distribution can be assumed over the operating space.

6. Implement the hierarchical Bayesian procedure outlined in Algorithm 4.1 to obtain the locally optimal model structure and similarity function hyperparameters for each operating space. This is equivalent to optimizing $P\big(\Phi, \mathcal{H}|\mathcal{D}^{(m)}, \{\Theta_q\}_{q=1}^{N_m}\big)$ for $m \in \{1, \cdots, M\}$.

Finally, the real-time implementation procedure of the proposed adaptive LW-PLS scheme is outlined in Algorithm 4.4.

Algorithm 4.4. *Real-time Implementation Procedure of an Adaptive LW-PLS Scheme*

1. Given a query sample $(\mathbf{x}_q, \nu_q)$, evaluate the posterior probability of the m^{th} operating region to assign an importance weight to the output of the m^{th} sub-model, $\hat{y}_q^{(m)}$ (Equations 4.29-4.32).

2. Estimate the parameters of the m^{th} LW-PLS sub-model based on $\hat{\Phi}^{(m)}$ and $\hat{\mathcal{H}}^{(m)}$ (Equations 4.5-4.8). If the importance weights are assigned according to Equation 4.31, it is only required to identify the sub-model with the highest posterior probability, *i.e.* $m = \operatorname{argmax} P(m|\nu_q, \mathcal{D})$. If the importance weights are assigned according to Equation 4.32, all the M sub-models should be identified.

3. Calculate the output of the m^{th} sub-model, $\hat{y}_q^{(m)}$, in order to obtain a global prediction, $\hat{y}_q$ (Equation 4.28).

4.4.2 A Special Case

To demonstrate the steps required in development of an adaptive LW-PLS framework, a special case is considered in this section.

4.4.2.1 Globally Optimal Model Structure and Similarity Function Hyperparameters

1. The following similarity measure is adopted in this work (Kim *et al.*, 2011):

$$s_{n|q} = \exp\left(- \frac{\phi d_{n|q}}{\sigma_{d|q}} \right) \tag{4.33}$$

where $d_{n|q}$ is the Euclidean distance between x_q and x_n, $\sigma_{d|q}$ is the standard deviation of $\{d_{n|q}\}_{n=1}^{N}$, and ϕ is a localization parameter. In this way, all the identification samples are weighted depending on their distance to the query sample. The similarity between the query and identification samples decreases steeply when ϕ is relatively large.

2. The prior distribution over the localization parameter is specified as follows to assure generality:

$$P(\phi|m) \propto \phi^{b-1} \exp\big(- \beta\phi \big) \tag{4.34}$$

3. To specify a set of plausible model structures, $\mathcal{H} \in \{\mathcal{H}_1, \cdots, \mathcal{H}_L\}$, $\mathcal{H}_l$ is considered to be a PLS model with l latent variables. Therefore, L is the maximum number of retained latent variables.

4. A uniform prior distribution is assumed over model structures:

$$P(\mathcal{H} = \mathcal{H}_l) = \frac{1}{L} \quad \text{for } l = 1, \cdots, L \tag{4.35}$$

5. The threshold for comparing the plausible model structures in terms of posterior odds is set to a certain pre-specified value (*e.g.* 20).

6. The model with l latent variables, $\mathcal{H}_l$, is selected. For this model, the following steps are repeated iteratively until no further improvements are gained:

6.1. The initial value of the localization parameter, $\phi_l^{[0]}$, is set to a certain number (*e.g.* 10). Note that the similarity between x_q and all the identification sample is equal to 1 for $\phi = 0$. Consequently, the LW-PLS model becomes identical to the PLS model.

6.2. The following steps are completed for each training sample $(\mathbf{x}_q, y_q) \in \mathcal{D}$: (1) The similarity matrix, $\mathbf{S}_q\big(\phi_l^{[k]}\big)$, is constructed (Equation 4.3). (2) A LW-PLS model with l retained latent variables is identified from $\mathcal{D}_{-q}$ (Equations 4.5-4.8). This is equivalent to calculating $\Theta_q^{[k]}$. (3) The identified LW-PLS model is used to obtain predictions of $\{y_1, \cdots, y_q, \cdots, y_N\}$ denoted as $\{\hat{y}_{1|q}, \cdots, \hat{y}_q^{cv}, \cdots, \hat{y}_{N|q}\}$ (Equation 4.4).

6.3. The following cost function is minimized to update the MAP estimates of hyperparameters, $\phi_l^{[k+1]}$ (Equation 4.19):

$$\mathcal{J}(\phi) = (1-b)\log\phi + \beta\phi + \frac{1}{2}\sum_{q=1}^{N}\Big[\frac{(y_q - \hat{y}_q^{cv})^2}{\sigma_q^2} + \log\sigma_q^2\Big] \tag{4.36}$$

where

$$\sigma_q^2 = \frac{\displaystyle\sum_{n=1,\ n\neq q}^{N} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)\big(y_n - \mathbf{x}_n\Theta_n^{[k]}\big)^2}{\displaystyle\sum_{n=1,\ n\neq q}^{N} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)} \tag{4.37}$$

7. Given ϕ_l^{MP}, Step 6.2 is repeated to obtain $\{\Theta_q^{\text{MP}}\}_{q=1}^N$.

8. The Hessian of the cost function $\mathcal{J}(\phi)$ is evaluated at ϕ_l^{MP}:

$$\mathbf{A} = \frac{(b-1)}{\phi^2} + \sum_{q=1}^{N}\Big[3\sigma_q^{-4}(y_q - \hat{y}_q^{cv})^2 - \sigma_q^{-2}\Big]\Big(\frac{\partial\sigma_q}{\partial\phi}\Big)^2 - \Big[\sigma_q^{-3}(y_q - \hat{y}_q^{cv})^2 - \sigma_q^{-1}\Big]\frac{\partial^2\sigma_q}{\partial\phi^2} \tag{4.38}$$

where

$$\frac{\partial \sigma_q}{\partial \phi} = \frac{\sigma_q}{2} \frac{\sum_{n=1,\ n\neq q}^{N} -\frac{d_{n|q}}{\sigma_d} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)\left(y_n - \mathbf{x}_n \Theta_n^{\text{MP}}\right)^2}{\sum_{n=1,\ n\neq q}^{N} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)\left(y_n - \mathbf{x}_n \Theta_n^{\text{MP}}\right)^2} - \frac{\sigma_q}{2} \frac{\sum_{n=1,\ n\neq q}^{N} -\frac{d_{n|q}}{\sigma_d} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)}{\sum_{n=1,\ n\neq q}^{N} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)} \tag{4.39}$$

$$\frac{\partial^2 \sigma_q}{\partial \phi^2} = \frac{\sigma_q}{2} \frac{\sum_{n=1,\ n\neq q}^{N} \left(\frac{d_{n|q}}{\sigma_d}\right)^2 \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)\left(y_n - \mathbf{x}_n \Theta_n^{\text{MP}}\right)^2}{\sum_{n=1,\ n\neq q}^{N} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)\left(y_n - \mathbf{x}_n \Theta_n^{\text{MP}}\right)^2} - \frac{\sigma_q}{2} \frac{\sum_{n=1,\ n\neq q}^{N} \left(\frac{d_{n|q}}{\sigma_d}\right)^2 \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)}{\sum_{n=1,\ n\neq q}^{N} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)} - 2\frac{\partial \sigma_q}{\partial \phi} \frac{\sum_{n=1,\ n\neq q}^{N} \frac{d_{n|q}}{\sigma_d} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)}{\sum_{n=1,\ n\neq q}^{N} \exp\left(-\frac{\phi d_{n|q}}{\sigma_d}\right)} - \sigma_q^{-1}\left(\frac{\partial \sigma_q}{\partial \phi}\right)^2 \tag{4.40}$$

Having calculated $\mathbf{A}$, the model evidence $P(\mathcal{D}|\mathcal{H}_l, \{\Theta_q^{\text{MP}}\}_{q=1}^N)$ is calculated (Equation 4.24)

9. The posterior probability of $\mathcal{H}_l$ is evaluated (Equation 4.25). Since the objective of model structure selection in this case is to find the optimal number of retained latent variables, $\{\mathcal{H}_1, \cdots, \mathcal{H}_L\}$ would be an ordered set of plausible models. If the posterior odds in pairwise comparison of $\mathcal{H}_{l-1}$ and $\mathcal{H}_l$ is greater than the pre-specified threshold, one returns to Step 6 to evaluate the plausibility of the model

$l+1$ latent variables. Otherwise, $l-1$ and ϕ_{l-1}^{MP} are selected as the globally optimal number of retained latent variables and localization parameter, respectively.

4.4.2.2 Partitioning and Characterizing the Operating Space

1. The operating space is partitioned into M sub-spaces by evaluating the performance of the global model. Suppose that $\{\mathcal{H}_l^{\text{MP}}, \phi_l^{\text{MP}}\}$ is the globally optimal setting obtained from Section 4.4.2.1.

   1.1. The following steps are completed for each training sample $(\mathbf{x}_q, y_q) \in \mathcal{D}$: (1) The similarity matrix, $\mathbf{S}_q\big(\phi_l^{\text{MP}}\big)$, is constructed (Equation 4.3). (2) A LW-PLS model with l retained latent variables is identified from $\mathcal{D}_{-q}$ (Equations 4.5-4.8). This is equivalent to calculating Θ_q^{MP}. (3) The identified LW-PLS model is used to obtain predictions of y_q denoted as $\hat{y}_q^{cv}$ (Equation 4.4).

   1.2. The prediction errors are calculated, *i.e.* $e_q = y_q - \hat{y}_q^{cv}$.

   1.3. The probability density of the absolute prediction errors can be used to decide on the number of sub-spaces. For instance, the operating space can be partitioned into two sub-spaces based on the PDF shown in Figure 4.1.

2. The training samples are classified into relevant operating regions in order to construct $\mathcal{D}^{(m)} = \{(\mathbf{x}_q, y_q)\}_{q=1}^{N_m}$ for $m = 1, \cdots, M$.

3. All or a sub-set of the latent variables of the global model are selected as the scheduling variables, ν. This can be viewed as partitioning of the latent operating space.

4. The joint PDF of the latent variables in the m^{th} operating mode, $P(\nu|\mathcal{D}^{(m)})$, is approximated from the identification data. This is equivalent to specifying the likelihood that ν would be generated by the m^{th} operating mode.

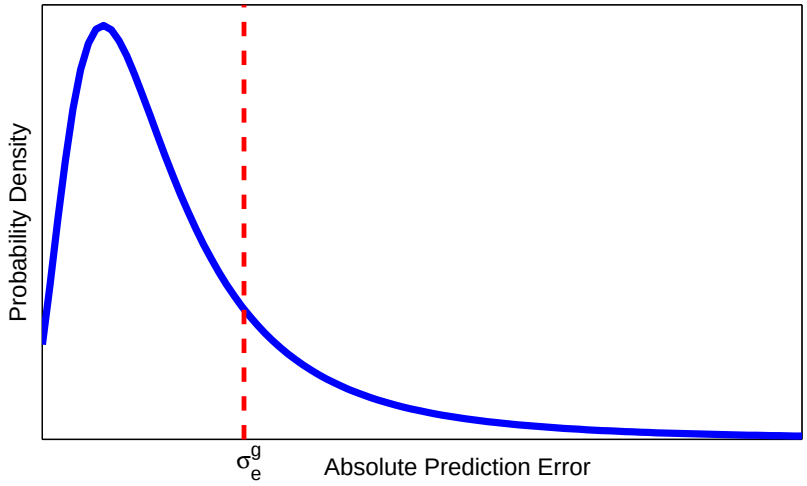


Figure 4.1: Probability density of the absolute prediction errors resulting from the LW-PLS model with the globally optimal setting

5. A uniform prior distribution is assumed over the operating space:

$$P(\text{mode} = m) = \frac{1}{M} \quad \text{for } m = 1, \cdots, M \tag{4.41}$$

4.4.2.3 Locally Optimal Model Structure and Similarity Function Hyperparameters

The hierarchical Bayesian procedure outlined in Section 4.4.2.1 is followed to obtain the locally optimal model structure and similarity function hyperparameters for each operating space. This is equivalent to optimizing $P\big(\Phi, \mathcal{H}|\mathcal{D}^{(m)}, \{\Theta_q\}_{q=1}^{N_m}\big)$ for $m \in \{1, \cdots, M\}$.

4.5 Case Studies

4.5.1 Active Substance in Pharmaceutical Tablets

To illustrate the advantages of the proposed hierarchical Bayesian optimization framework, we consider the problem of chemometric quantization of the active substance of a pharmaceutical tablet using near-infrared (NIR) transmittance spectra. The objective is to develop a LW-PLS model for predicting the active substance content, *i.e.* weight percent, of a pharmaceutical tablet from NIR transmittance spectra (404 points in the range of

Table 4.3: Tablet specifications (Dyrby *et al.*, 2002)

Nominal content of active substance per tablet (mg)	Nominal tablet weight (mg)	Nominal weight percent ($\%$)
5.0	90	5.6
10.0	125	8.0
15.0	188	8.0
20.0	250	8.0
4.3-5.7	90	4.8-6.3
8.3-11.4	125	6.9-9.1
12.9-17.1	188	6.9-9.1
17.3-22.8	250	6.9-9.1

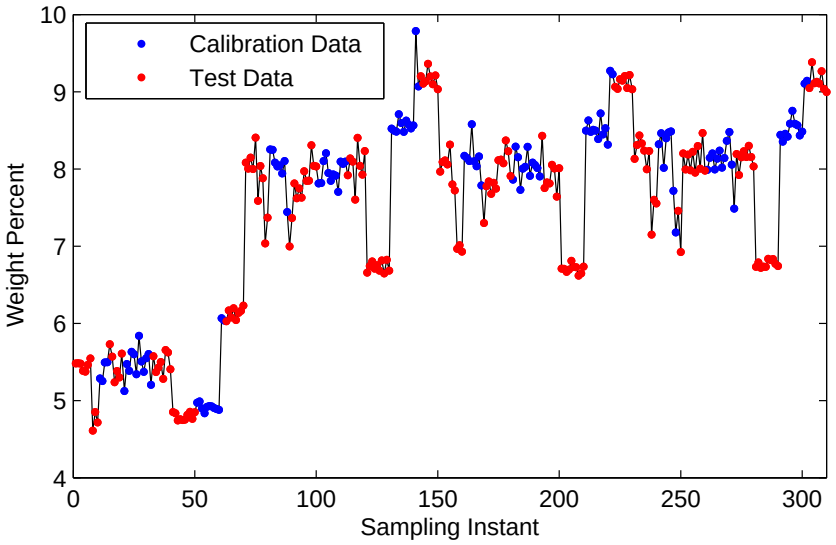
$7,400 - 10,500\ cm^{-1}$). The real-world data-set used in this case study is taken from Dyrby *et al.* (2002).

As shown in Table 4.3, the data was collected for different dosage values of this pharmaceutical drug, ranging from 4.3 to $22.8\ mg$ tablets. Two different cases are considered for constructing the calibration and test data-sets:

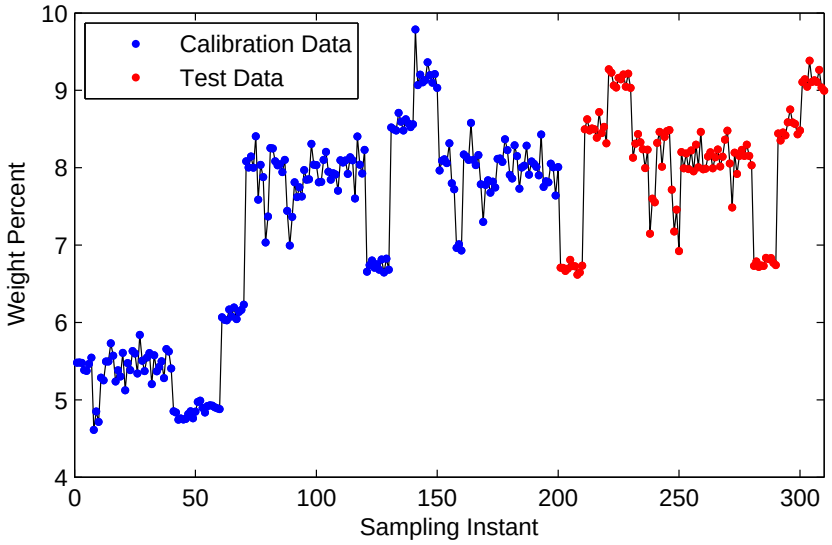
- **Case I.** Calibration samples covering range $85 - 115\%$ of the nominal content are available for all dosages (Figure 4.2.a). The NIR spectra are subject to the standard normal variate transformation (Barnes *et al.*, 1989) in the pre-processing step.

- **Case II.** Calibration samples covering range $85 - 115\%$ of the nominal content are available only for some dosages (Figure 4.2.b). In the pre-processing step, the NIR spectra are subject to the standard normal variate transformation as well as the first-order differentiation using Savitzky-Golay filter (Savitzky and Golay, 1964).

The procedure outlined in Section 4.4.2.1 is followed to find the globally optimal number of the retained latent variables as well as estimate of the localization parameter of the similarity function defined in Equation 4.33. A uniform prior distribution is assumed over the number of retained latent variables. If the pseudo-Bayes factor in pairwise comparison



(a) Case I



(b) Case II

Figure 4.2: Calibration and test samples of the content (weight percent) of the active substance

of $\mathcal{H}_{l-1}$ and $\mathcal{H}_l$ is less than 20 (Table 4.2), $\mathcal{H}_{l-1}$ with $l-1$ latent variable is selected as the best model structure. Moreover, a constrained uniform prior distribution, within the ranges reported in Table 4.4, is assumed over the localization parameter. The results are compared by the ones obtained from the classical global cross-validation method. That is, the optimal combination of the number of retained latent variables and the value of

localization parameter is found by implementing Algorithm 4.1.

The comparison results are presented in Table 4.4 and Table 4.5 for Case I and Case II, respectively. It can be observed that the optimal global values obtained from the classical cross-validation approach are sensitive to the specified search range. For instance, if the maximum number of latent variables changes from 5 to 8, the number of the latent variables resulting in the lowest RMSECV increases from 5 to 8 for Case I and from 3 to 7 for Case II. Furthermore, as the search range increases, the RMSE of prediction decreases for the training data-set but increases for the test data-set. It can be concluded that evaluating the RMSECV resulted by LOOCV technique has led to the over-fitting phenomenon. The proposed hierarchical Bayesian optimization framework, in contrast, can deal with the model complexity control to avoid over-fitting.

Table 4.4: Comparing the prediction performance of the LW-PLS models characterized by the hierarchical Bayesian optimization and classical cross-validation methods using calibration data-set I

	Hierarchical Bayesian	Classical LOOCV
Maximum number of LVs = 5; Range of localization parameter = [0.5,10]		
Selected number of retained LVs	4	5
Selected value of localization parameter	0.91	1
RMSE of prediction for training data	0.3278	0.3093
RMSE of prediction for test data	0.4010	0.4096
Maximum number of LVs = 8; Range of localization parameter = [0.5,10]		
Selected number of retained LVs	4	8
Selected value of localization parameter	0.91	0.5
RMSE of prediction for training data	0.3278	0.2707
RMSE of prediction for test data	0.4010	0.5345
Maximum number of LVs = 8; Range of localization parameter = [0.1,10]		
Selected number of retained LVs	4	8
Selected value of localization parameter	0.94	0.18
RMSE of prediction for training data	0.3278	0.2653
RMSE of prediction for test data	0.4010	0.5353

Table 4.5: Comparing the prediction performance of the LW-PLS models characterized by the hierarchical Bayesian optimization and classical cross-validation methods using calibration data-set II

	Hierarchical Bayesian	Classical LOOCV
Maximum number of LVs = 5; Range of localization parameter = [0.5,10]		
Selected number of retained LVs	1	3
Selected value of localization parameter	3.7	5
RMSE of prediction for training data	0.2341	0.2290
RMSE of prediction for test data	0.3873	0.3748
Maximum number of LVs = 8; Range of localization parameter = [0.5,10]		
Selected number of retained LVs	1	7
Selected value of localization parameter	3.7	3.3
RMSE of prediction for training data	0.2341	0.2243
RMSE of prediction for test data	0.3873	0.4828
Maximum number of LVs = 8; Range of localization parameter = [0.1,10]		
Selected number of retained LVs	1	7
Selected value of localization parameter	3.7	3.3
RMSE of prediction for training data	0.2341	0.2243
RMSE of prediction for test data	0.3873	0.4828

To further improve the results, the procedure outlined in Section 4.4.2.2 is followed to partition the operating space into two sub-spaces by evaluating the residuals of the globally identified LW-PLS model. For each sub-space, the optimal number of the retained latent variables as well as the optimal value of the localization parameter are obtained by implementing the Bayesian hierarchical optimization scheme. The maximum number of the latent variables is set to 5. A constrained uniform prior distribution in the range of $[0.5, 2]$ is assumed over the localization parameter. The results are reported in Table 4.6. The smaller values of the RMSE indicate that the LW-PLS with multiple sets of the locally optimal number of the latent variables and the localization parameter estimate has a better prediction performance than the classical LW-PLS with a single globally optimal set.

Table 4.6: Prediction performance of the multi-model LW-PLS

	Sub-model 1	Sub-model 2	Multi-model
Case I			
Selected number of retained LVs	4	2	-
Selected value of localization parameter	2	0.77	-
RMSE of prediction for test data	-	-	0.3700
Case II			
Selected number of retained LVs	2	2	-
Selected value of localization parameter	2	0.5	-
RMSE of prediction for test data	-	-	0.2993

4.5.2 Reid Vapor Pressure of Gasoline

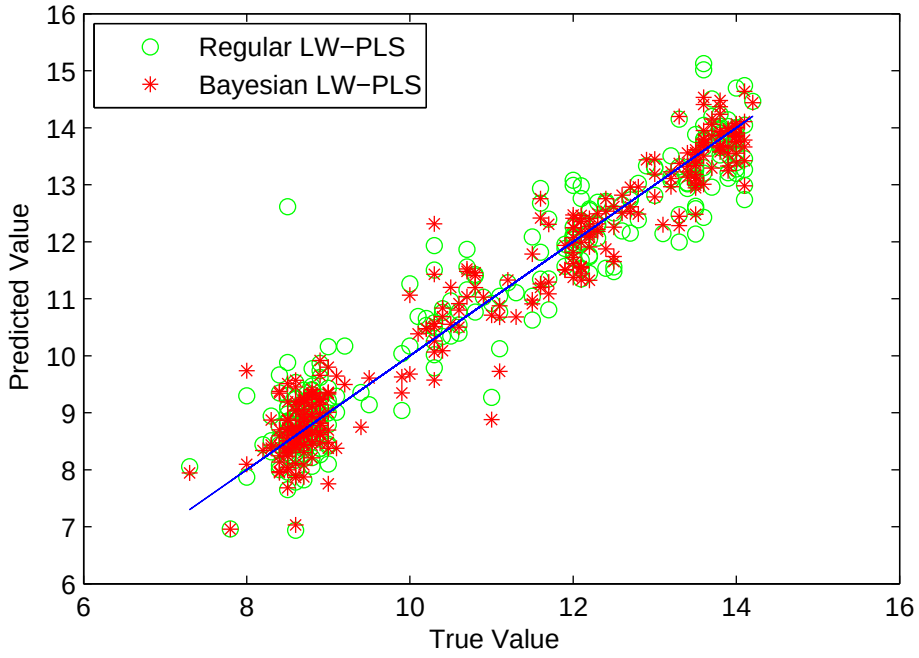
In this case study, the objective is to develop a LW-PLS model for real-time prediction of Reid vapor pressure (RVP) of gasoline that is a measure of the volatility of gasoline. The industrial data have been provided by a refinery located in Edmonton, Canada. A total of 423 gasoline samples were collected from on-line operation between August 2007 and July 2012. The NIR spectra of the collected samples were recorded using an NIR spectrometer having the wavelength range of $800 - 1,700\ nm$ and nominal spectral resolution of $1\ nm$. The reference data for the RVP were obtained using standard ASTM testing methodologies. The NIR spectra are subject to the standard normal variate transformation in the pre-processing step. The minimum and maximum number of the latent variables are set to 12 and 20, respectively. A constrained uniform prior distribution in the range of $[0.5, 1]$ is assumed over the localization parameter. The comparison results are reported in Table 4.7 and illustrated in Figure 4.3. From both arithmetical and graphical comparisons, it can be observed that the proposed Bayesian framework results in the LW-PLS models with better prediction performance.

Table 4.7: Comparing the prediction performance of the LW-PLS models characterized by the hierarchical Bayesian optimization and classical cross-validation methods

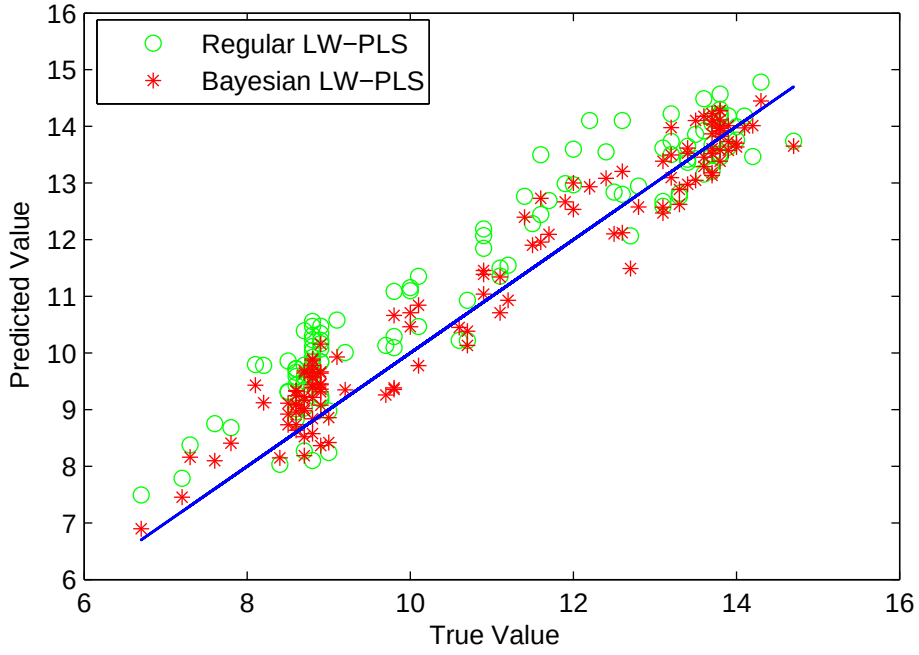
	Hierarchical Bayesian	Classical LOOCV
Selected number of retained LVs	15	12
Selected value of localization parameter	0.59	1
RMSE of prediction for training data	0.4937	0.6044
RMSE of prediction for test data	0.5678	0.8825

4.6 Concluding Remarks

The objective of this study was twofold. First, it was desired to develop a systematic approach for selection of the LW-PLS model structure and its region of validity. Second, it was required to develop a computationally feasible method through which the effect of the system non-linearity on the functional forms and size of the receptive fields can be taken into account for real-time identification of the LW-PLS models. To achieve the aforementioned objectives, the proposed method consists of two main steps. First, the operating space is partitioned into a finite number of sub-spaces during the off-line identification phase. Next, the problem of finding the locally optimal LW-PLS model structure and similarity function hyperparameters is formulated and solved under an iterative hierarchical Bayesian optimization framework for each sub-space. In this way, the real-time identification problem only amounts to detecting the underlying operating sub-space and estimating the LW-PLS model parameters. Therefore, the real-time model structure selection and similarity function parametrization become more computationally efficient. In the proposed optimization scheme the leave-one-out predictive densities are evaluated to perform objective comparison between alternative model structures. It also provides objective criteria for obtaining the locally optimal hyperparameters of the similarity function. Two industrial case-studies were considered to demonstrate the



(a) Auto-validation



(b) Cross-validation

Figure 4.3: Prediction performance of the LW-PLS

effectiveness of the proposed method: 1. real-time prediction of Reid vapor pressure of gasoline in a petrochemical refinery, and 2. real-time prediction of the active substance

content of a pharmaceutical tablet. The method was successfully applied to identify inferential LW-PLS models for real-time prediction of these quality variables using near-infrared (NIR) transmittance spectra.

Bibliography

Atkeson, C. G., A. W. Moore and S. Schaal (1997). Locally weighted learning. *Artificial Intelligence Review* **11**, 11–73.

Barnes, R. J., M. S. Dhanoa and S. J. Lister (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy* **43**(5), 772–777.

Chen, H., B. R. Bakshi and P. K. Goel (2007). Bayesian latent variable regression via gibbs sampling: Methodology and practical aspects. *Journal of Chemometrics* **21**, 578–591.

Chipman, H., E. I. George and R. E. McCulloch (2001). The practical implementation of bayesian model selection. In: *IMC Lecture Notes–Monograph Series* (P. Lahiri, Ed.). Vol. 38. pp. 65–116. Institute of Mathematical Statistics. Beachwood, USA.

Chun, H. and S. Keleş (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society - Series B* **72**(1), 3–25.

Dyrby, M., S. B. Engelsen, L. Nørgaard, M. Bruhn and L. Lundsberg-Nielsen (2002). Chemometric quantitation of the active substance (containing $c \equiv n$) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-raman spectra. *Applied Spectroscopy* **56**(5), 579–585.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**(365), 153–160.

Gelfand, A. E. and D. K. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **56**(3), 501514.

Guyon, I., A. Saffari, G. Dror and G. Cawley (2010). Model selection: Beyond the bayesian/frequentist divide. *Journal of Machine Learning Research* **11**, 61–87.

Kass, R. A. and A. E. Raftery (1995). Bayes' factor. *Journal of the American Statistical Association* **90**(430), 773–795.

Khatibisepehr, S. and B. Huang (2012). A Bayesian approach to design of adaptive multi-model inferential sensors with application in oil sand industry. *Journal of Process Control* **22**(10), 1913–1929.

Kim, S., M. Kano, H. Nakagawa and S. Hasebe (2011). Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *International Journal of Pharmaceutics* **421**(1), 269–274.

Kim, S., M. Kano, S. Hasebe, A. Takinami and T. Seki (2013). Long-term industrial applications of inferential control based on just-in-time soft-sensors: Economical impact and challenges. *Industrial and Engineering Chemistry Research*.

Kolda, T., R. Lewis and V. Torczon (2003). Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review* **45**(3), 385–482.

Lin, B. and S. B. Jørgensen (2011). Soft sensor design by multivariate fusion of image features and process measurements. *Journal of Process Control* **21**(4), 547–553.

Lin, B., B. Recke, J. K. H. Knudsen and S. B. Jørgensen (2007). A systematic approach for soft sensor development. *Computers and Chemical Engineering* **31**(5-6), 419–425.

Ljung, L. (1999). *System Identification - Theory For the User*. second ed.. Prentice Hall. Upper Saddle River, USA.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation* **4**(3), 415–447.

MacKay, D. J. C. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation* **11**(5), 1035–1068.

MacKay, D. J. C. (2002). *Information Theory, Inference, and Learning Algorithm*. first ed.. Cambridge University Press. New York, USA.

Marjanovic, O., B. Lennox, D. Sandoz, K. Smith and M. Crofts (2006). Real-time monitoring of an industrial batch process. *Computers and Chemical Engineering* **30**(10-12), 1476–1481.

Mobaraki, N. and B. Hemmateenejad (2011). Structural characterization of carbonyl compounds by IR spectroscopy and chemometrics data analysis. *Chemometrics and Intelligent Laboratory Systems* **109**(2), 171–177.

Penny, W., J. Mattout and N. Trujillo-Barreto (2006). Bayesian model selection and averaging. In: *Statistical Parametric Mapping: The Analysis Of Functional Brain Images* (K. Friston, J. Ashburner, S. Kiebel, T. Nichols and W. Penny, Eds.). Chap. 35. Elsevier. London, UK.

Perez-Guaita, D., J. Kuligowski, G. Quintás, S. Garrigues and M. de la Guardia (2013). Modified locally weightedpartial leas tsquares regression improving clinical predictions from infrared spectra of human serum samples. *Talanta* **107**(1), 368–375.

Savitzky, A. and M. J. E. Golay (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**(8), 1627–1639.

Schaal, S., C. G. Atkeson and S. V. Vijayakumar (2002). Scalable techniques from nonparametric statistics for real time robot learning. *Applied Intelligence* **17**(1), 49–60.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**(1), 486–494.

Sra, S., S. Nowozin and S. J. Wright (eds.) (2011). *Optimization for Machine Learning*. MIT Press.

Sundararajan, S. and S. S. Keerthi (2001). Predictive approaches for choosing hyperparameters in gaussian processes. *Neural Computation* **13**(5), 1103–1118.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In: *IMC Lecture Notes–Monograph Series* (P. R. Krishnaiaah, Ed.). Vol. Multivariate Analysis. Academic Press. New York.

Wold, S., A. Ruhe, H. Wold and W. J. Dunn (1984). The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing* **5**(3), 735–743.

Yu, J. (2012). A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. *Computers and Chemical Engineering* **41**(11), 134–144.

Chapter 5

A Hierarchical Bayesian Framework for Robust Identification of Inferential Models from Contaminated Data-set

5.1 Introduction

5.1.1 Practical Motivation

Reliable process models are key requirements for investigating the behavior of industrial processes. Such descriptive models can help to improve process productivity, achieve safety of operation, and develop tight control policies (Fortuna *et al.*, 2007).

Depending on the level of *a priori* knowledge, different strategies have been proposed in the literature to model chemical processes. Traditionally, knowledge-driven models are developed on the basis of first principles analysis, which requires complete understanding of underlying mechanisms (Prasad *et al.*, 2002; Muller *et al.*, 2011; Sabbe *et al.*, 2011). Not surprisingly, development of first principle models can often be prohibitively difficult and time-consuming due to the complexity of industrial processes. Therefore, decades of research have been devoted to developing empirical process models without complete *a priori* knowledge of the internal mechanisms governing the process dynamics. The

A version of this chapter has been published in AIChE Journal, Volume 59 (Khatibisepehr and Huang, 2013). An abbreviated version of this chapter was presented at the 2012 American Control Conference, June 27-29, 2012, Montreal, Canada (Khatibisepehr and Huang, 2012).

empirical models are usually constructed based on the limited process knowledge as well as the great amount of historical data acquired for monitoring purposes (Kano and Nakagawa, 2008; Kadlec *et al.*, 2009). In the context of process industries, various data-driven model structures can be used to describe the behavior of unit operations. Owing to the capability of autoregressive with exogenous input (ARX) models in approximating complex linear dynamic systems, ARX model structure is commonly adopted in industrial applications (Fortuna *et al.*, 2007). Regardless of the model structure selected, the procedures applied to identify empirical process models are often highly sensitive to the quality of identification data, *i.e.* the varying quality of process data can greatly deteriorate the performance of data-driven identification methods.

Outlying measurements, also called **outliers**, are one of the common factors that may affect the quality of operational and laboratory data (Chiang *et al.*, 2003; Liu *et al.*, 2004; Khatibisepehr and Huang, 2008). Outliers are observations which appear to deviate markedly from the typical ranges of other observations (Grubbs, 1969). The outliers in operational data mostly represent a random error caused by such issues as process disturbances, instrument degradation, and transmission problems (Zeng and Gao, 2009; Lee *et al.*, 2011). Moreover, the outlying laboratory measurements may be generated due to potential human errors that may occur in collecting samples, conducting experiments, and recording results. Statistical analysis of process data contaminated with outliers may lead to biased parameter estimation and plant-model mismatch. Therefore, the problem of process model identification in the presence of outliers has received great attention during the last two decades and a wide variety of so-called outlier identification approaches have been proposed (Hodge and Austin, 2004; Chandola *et al.*, 2009). As pointed out by Kadlec *et al.* (2009), however, this issue is currently solved in a rather *ad hoc* manner. Therefore, there is a great need to seek for more advanced and more general solutions.

5.1.2 Main Contributions

The main contribution of this work is to formulate and solve the ARX model identification problem in the presence of outliers under a novel robust Bayesian framework consisting of consecutive levels of optimization. First, we adopt a contaminated distribution to describe the observed data and introduce a set of indicator variables to denote the quality of each data point. Next, we propose a unified objective function for model identification in the presence of outliers. The resulting optimization problem is hierarchically decomposed and a layered optimization strategy is implemented. In order to obtain explicit solutions, we adopt an iterative hierarchical Bayesian approach through which the solutions obtained in subsequent layers of optimization are coordinated. The proposed method has the following attractive features:

1. The outlined optimization strategy not only yields maximum *a posteriori* (MAP) estimates of model parameters, but also provides an automated mechanism for determining the hyperparameters and for investigating the quality of each observation.

2. The developed framework allows us to incorporate the prior knowledge of the noise distribution and to include the relevant information contained in identification data. Thereby, restrictive assumptions made in traditional robust methods about contaminating distributions (*e.g.* symmetric noise distribution) can be relaxed.

3. The identification procedures employed in classical statistical estimation techniques often result in a set of single-valued parameter estimates. In contrast, the full Bayesian model identification results in posterior distributions over parameters to reveal how uncertain the estimated values would be (Khatibisepehr and Huang, 2008).

4. In the classical approaches available for identification of inferential models from

contaminated data, the data quality assessment and model identification steps are often disjoint. Within the proposed hierarchical framework, however, the Bayesian inference at a particular level takes into account the uncertainty in the estimates of the previous level. This is a great feature that allows us to link different levels of Bayesian inference together and, consequently, interconnect the solutions obtained in subsequent layers of optimization. In this way, the data quality assessment and model identification steps become integrated.

5.1.3 Chapter Outline

The remainder of this Chapter is organized as follows. A brief overview of the existing outlier identification techniques is presented in Section 5.2. In Section 5.3, the problem of ARX model identification in the presence of outliers is discussed. Our proposed objective function resulting in the consecutive layers of optimization is described in Section 5.4. The idea of hierarchical Bayesian inference approach adopted to solve the layered optimization problem is also explained in this Section. The most common outlier models are introduced in Section 5.5. In Section 5.6, the problem of ARX model parameter estimation is formulated in a unified Bayesian framework and the details of the identification procedure are presented. In Section 5.7, the application of the developed framework is demonstrated on numerical simulation and experimental examples. These case studies will show robustness of the proposed parameter estimation method in the presence of outliers, which is an attractive feature for applying the proposed method to real world problems. Finally, this Chapter is summarized with the concluding remarks in Section 5.8.

5.2 An Overview of the Existing Outlier Identification Methods

Outlier identification constitutes an essential prerequisite for identification of process models and thus several outlier handling approaches have been developed in the past few

decades. Since the focus of this work is on the Bayesian methods, we limit our literature review to the most common **statistical** approaches. A comprehensive review of the outlier detection problem and several outlier detection algorithms is given by Hodge and Austin (2004); Kadlec *et al.* (2009); Chandola *et al.* (2009).

Statistical analysis of residuals described in Fortuna *et al.* (2007) is one of the common outlier detection approach. This is based on the use of a regression model between dependent and independent variables. First, the least square method is applied to obtain an estimation of model parameters for normal operating condition. Outliers can then be detected if the model residuals of new data lie outside a specified confidence interval. Since outliers can significantly deteriorate least-squares solutions, robust regression can be applied to handle them while fitting regression models. In general, robust regression methods are designed to iteratively downweight the influence of outliers. The most common robust regression analysis is performed with M (maximum likelihood) estimators, introduced by Huber (1981). The general M-estimator minimizes the objective function $\sum_{t=1}^{N} \rho(\varepsilon_n)$, where the function ρ gives the contribution of each residual to the objective function. Least-squares estimation would be an special case for which $\rho(\varepsilon_n) = \varepsilon_n^2$). Differentiating the objective function with respect to the parameters and setting the partial derivatives to 0, the estimating equations may be written as $\sum_{t=1}^{N} w_n \varepsilon_n X_n = 0$, where the robustness weight assigned to the t^{th} observation, $w_n = w(\varepsilon_n)$, is obtained from the weight function defined as $w(\varepsilon) = \rho'/\varepsilon$. For instance, the robustness weights in the Huber robust regression technique are determined using the Huber weighting function defined as

$$w_n = \left[\max\left(1, \left| \frac{\varepsilon_n}{c \times s} \right| \right) \right]^{-1} \tag{5.1}$$

where ε_n is the residual calculated from the previous iteration, c is the tuning constant, and $s = \text{MAD}/0.6745$ is an estimate of the standard deviation of the error term.

Several solutions have been proposed for solving the outlier detection problem by estimating a probability density of the normal data. For instance, in Bishop (1994) the

density distribution of the input space is first estimated by a standard Parzen window approach with Gaussian kernel functions. Next, a suitable threshold is specified based on the identification data-set which is known to be representative of normal data. The new observation is then flagged as an outlier, if the value of the density function is above the threshold. Yu (2012) proposed a Bayesian approach to first estimate the posterior probabilities of all samples within the model input space and specify appropriate confidence levels. A calibration procedure is then followed to correct the observations identified as outliers. An alternative approach for probability density estimation is to model normal instances as a mixture of parametric distributions. Bishop (1994) and Agarwal (2006) used Gaussian mixture models for such techniques. In Ritter and Gallegos (1997), both normal instances and outliers are modeled as separate parametric distributions. First, the ellipsoidal multivariate trimming (MVT) (Rousseeuw and Leroy, 1996) technique is used to detect outliers and to estimate distribution parameters of both outliers and regular observations. Next, a Bayesian classifier is designed to compare certain linear combinations of posterior densities of each data vector with respect to the estimated distributions. Several variations of Bayesian classification technique have further been proposed by Varbanov (1998); Ghosh-Dastidar and Schafer (2006); Das and Schneider (2007), and many others.

In this research, we take a hierarchical Bayesian approach to address the problem of model identification in the presence of outliers. We develop a robust Bayesian inference framework consisting of three consecutive steps: 1) Given an identification data-set, the posterior probability of each observation acting as an outlier is evaluated; a set of indicator variables is specified to denote the quality of each data point. 2) The hyperparameters are then estimated by solving an optimization problem that maximizes the posterior probability distribution of hyperparameters conditional upon the indicator variables. 3) Given current estimates of hyperparameters and indicator variables, the posterior probability distribution of model parameters is maximized to obtain MAP estimates. These three steps will be

repeated until the estimates change within a given tolerance.

5.3 Problem Statement

In the identification of an empirical model, the overall objective is to find a model that best fits the identification data-set, $\mathcal{D} = \{(\mathbf{r}_t, y_t)\}_{t=1}^{N}$. Bayesian models are a compact way to represent probabilistic relationships between a set of random variables in a system. Before going into details of how to learn Bayesian models, we need a more detailed definition of what the model includes. A model is defined by its functional form, f, and a set of parameters, Θ. Let us consider a general form of a non-linear model:

$$y_t = f(\mathbf{r}_t; \Theta) + e_t \tag{5.2}$$

where $y_t \in \mathbb{R}$ is the output, $\mathbf{r}_t \in \mathbb{R}^P$ is the regressor constructed from past inputs and outputs, and e_t is the noise/error term.

Suppose e_t is modeled as a zero-mean Gaussian noise with constant standard deviation σ_e. Given the model structure, $\mathcal{H}$, and the model parameters, Θ, the likelihood of the data can be expressed as

$$p(\mathcal{D}|\Theta, \zeta, \mathcal{H}) = \Big(\frac{\zeta}{2\pi}\Big)^{N/2} \exp\big(-\zeta E_D(\mathcal{D}|\Theta, \zeta, \mathcal{H})\big) \tag{5.3}$$

where ζ defines a noise level with $\sigma_e^2 = \zeta^{-1}$, and E_D is the error term defined as

$$E_D = \frac{1}{2}\sum_{t=1}^{N} e_t^2 = \frac{1}{2}\sum_{t=1}^{N} \big(y_t - f(\mathbf{r}_t; \Theta)\big)^2 \tag{5.4}$$

It is well-known that finding the maximum likelihood estimates of the parameters, Θ_{ML}, may be an ill-posed problem. Since the Θ that minimizes E_D may depend sensitively on the details of the noise in the data, the maximum likelihood estimates would oscillate widely so as to fit the noise (MacKay, 1992). Bayesian methods solve this type of ill-posed problem by combining information contained in the observed data with available information concerning the distribution of the parameters. Introducing a regularizing

constant, α, such a prior can be expressed on the parameters; $p(\Theta|\alpha, \mathcal{H})$ represents the current state of knowledge about the plausible values of model parameters. Therefore, the prior distribution of model parameters is parameterized by a set of variables called hyperparameters*. Both ζ and α are considered as hyperparameters, because they describe the overall characteristics of the priors. If a hyperparameter is not known *a priori*, its probability distributions can be estimated in an intermediate step of the model identification process.

To develop a Bayesian formulation of inferential models that is robust to inconsistent data, we need to be able to efficiently perform different levels of Bayesian inference even if the data-set is contaminated with outlying observations. Given an identification data-set $\mathcal{D}$, we can consider a set of hyperparameters $\{\zeta_1, ..., \zeta_N\}$. Thus, the hyperparameter ζ_t defines a noise level $\sigma_{e_t}^2 = \zeta_t^{-1}$ on the t^{th} sample in the given training data-set. When having non-constant values of ζ_t, the outliers will be automatically handled by assigning less weights to the observations with relatively larger $\sigma_{e_t}^2$. However, the underlying formulation involves a heavy non-linear optimization problem in dealing with large data-sets.

To obtain a computationally feasible formulation, we adopt a contaminated distribution to describe the observed data and then solve the problem under a unified Bayesian framework. The error distribution function is thus expressed as $F(e) = \delta G(e) + (1 - \delta)H(e)$, where δ is the unknown prior probability of appearance of an outlier, $H(e) = \mathcal{N}(0, \sigma_e^2)$ is a normal distribution, and $G(e)$ is a contaminating distribution. This model arises for instance if the observations are assumed to be normal with variance σ_y^2, but a fraction δ of them is affected by gross errors (Huber, 1981). Moreover, a set of indicator variables $\{q_1, \cdots, q_N\}$ is introduced to denote the quality of the observed data; the indicator variable associated with each data point determines whether that observation comes from the regular or contaminating distribution. However, the indicator variables are usually not known *a priori* and should be estimated in an intermediate step of the model

*The term is used to distinguish them from model parameters.

identification process.

5.4 Outlier Models

In general, we need to tackle two types of outliers, namely scale outliers and location outliers. As the names suggested, scale and location outliers are generated by a shift in the scale (variability) or in the location (mean) of measurement noise. The process measurements that violate the physical limitations of the involved unit operations can be modeled as scale outliers, while the ones that violate the technological limitations of the measuring devices can often be considered as symmetric location outliers. Moreover, the outlying measurements made by a jammed instrument may be modeled as asymmetric location outliers.

In this section, we present our proposed scale and location outlier models which later will be needed to develop a robust Bayesian framework.

5.4.1 Scale Outlier Model

The error distribution affected by scale outliers is a mixture of two multivariate normal distributions centered at the same mean but with different covariance matrices, one being proportionately larger than the other. Therefore, it is assumed that the noise term, e_t, is distributed as

$$e_t \sim \delta\mathcal{N}(0, \rho^{-1}\sigma_e^2) + (1-\delta)\mathcal{N}(0, \sigma_e^2) \tag{5.5}$$

where $0 < \rho < 1$ is the variance inflation factor that indicates the magnitude of the errors leading to an outlying observation. Note that the proposed Bayesian framework does not require any knowledge of the noise distribution parameters (*e.g.* δ, σ_e and ρ); these parameters are iteratively estimated in the identification process using the observations identified as outliers.

Introduce a set of indicator variables, $\mathbf{q}_{1:N} = \{q_1, \cdots, q_N\}$, to denote identity of each

data point, *i.e.* $q_t = \rho$ if e_t is distributed as $\mathcal{N}(0, \rho^{-1}\sigma_e^2)$ and $q_t = 1$ if e_t is distributed as $\mathcal{N}(0, \sigma_e^2)$. Therefore, q_t is Bernoulli distributed with parameter δ. That is,

$$p(q_t; \delta) = \delta^{\Big(1 - \frac{q_t - \rho}{1 - q_t\rho}\Big)}(1 - \delta)^{\Big(\frac{q_t - \rho}{1 - q_t\rho}\Big)} \tag{5.6}$$

5.4.2 Location Outlier Model

Now, suppose the contaminating distribution consists of two multivariate normals such that $G(e) = \mathcal{N}(-\Delta, \sigma_e^2) + \mathcal{N}(\Delta, \sigma_e^2)$. To capture the presence of location outliers, it is thus assumed that the noise term, e_t, is distributed as

$$e_t \sim \delta\big[\mathcal{N}(\Delta, \sigma_e^2) + \mathcal{N}(-\Delta, \sigma_e^2)\big] + (1 - \delta)\mathcal{N}(0, \sigma_e^2) \tag{5.7}$$

where Δ indicates the location shift in the outlying observations. As mentioned previously, the proposed Bayesian framework does not require any knowledge of the noise distribution parameters (*e.g.* δ, σ_e and Δ); these parameters are iteratively estimated in the identification process using the observations identified as outliers.

Introduce a set of indicator variables, $\mathbf{q}_{1:N} = \{q_1, \cdots, q_N\}$, to denote identity of each data point, *i.e.* $q_t = +\Delta$ if e_t is generated from $\mathcal{N}(+\Delta, \sigma_e^2)$, $q_t = -\Delta$ if e_t is generated from $\mathcal{N}(-\Delta, \sigma_e^2)$, and $q_t = 0$ if e_t is distributed as $\mathcal{N}(0, \sigma_e^2)$. Therefore, q_t has a categorical distribution expressed as

$$p(q_t; \delta) = (0.5\delta)^{\Big(\frac{|q_t| - q_t}{2\Delta}\Big)}(0.5\delta)^{\Big(\frac{|q_t| + q_t}{2\Delta}\Big)}(1 - \delta)^{\Big(1 - \frac{|q_t|}{\Delta}\Big)} \tag{5.8}$$

or equivalently, $|q_t|$ has a Bernoulli distribution:

$$p(|q_t|; \delta) = \delta^{\Big(\frac{|q_t|}{\Delta}\Big)}(1 - \delta)^{\Big(1 - \frac{|q_t|}{\Delta}\Big)} \tag{5.9}$$

5.5 Hierarchical Optimization Framework

In general, the identification problem is to estimate the model parameters, Θ, the hyperparameters of the prior distribution of model parameters, Φ, and the indicator

variables, Q, using the process data-set, $\mathcal{D}$. To obtain MAP estimates simultaneously, the joint probability density function, $p(\Theta, \Phi, Q|\mathcal{D})$ should be optimized. However, evaluating such posterior density functions requires a complex non-linear optimization problem to be solved. To circumvent the difficulties associated with the direct maximization of $p(\Theta, \Phi, Q|\mathcal{D})$, the identification problem is formulated under a layered optimization framework, as we will show in the following.

First, the chain rule of probability theory is used to factorize the joint probability density function (JPDF) as

$$p(\Theta, \Phi, Q|\mathcal{D}) = p(\Theta|\Phi, Q, \mathcal{D})p(\Phi|Q, \mathcal{D})p(Q|\mathcal{D}) \tag{5.10}$$

Then, the optimization problem is decomposed hierarchically into three layers:

$$\begin{aligned}
&\max_{\Theta,\Phi,Q} p(\Theta|\Phi, Q, \mathcal{D})p(\Phi|Q, \mathcal{D})p(Q|\mathcal{D}) \\
&\quad = \max_{\Phi,Q} \left\{ p(Q|\mathcal{D})p(\Phi|Q, \mathcal{D}) \max_{\Theta} \left\{ p(\Theta|\Phi, Q, \mathcal{D}) \right\} \right\} \\
&\quad = \max_{Q} \left\{ p(Q|\mathcal{D}) \max_{\Phi} \left\{ p(\Phi|Q, \mathcal{D}) \max_{\Theta} \left\{ p(\Theta|\Phi, Q, \mathcal{D}) \right\} \right\} \right\}
\end{aligned} \tag{5.11}$$

The three-layer optimization problem is formulated as follows:

1. Inference of model parameters Θ by maximizing the following posterior density function
$$p(\Theta|\mathcal{D}, \Phi, Q) = \frac{p(\mathcal{D}|\Theta, \Phi, Q)p(\Theta|\Phi, Q)}{p(\mathcal{D}|\Phi, Q)} \tag{5.12}$$
2. Inference of hyperparameters Φ by maximizing the following posterior density function
$$p(\Phi|\mathcal{D}, Q) = \frac{p(\mathcal{D}|\Phi, Q)p(\Phi|Q)}{p(\mathcal{D}|Q)} \tag{5.13}$$
3. Inference of outlier indicator variables Q by maximizing the following posterior density function
$$p(Q|\mathcal{D}) = \frac{p(\mathcal{D}|Q)p(Q)}{p(\mathcal{D})} \tag{5.14}$$

In this Bayesian formulation, the likelihood function at a particular level corresponds to the evidence function at the previous level. For example, the likelihood at Level 2, $p(\mathcal{D}|\Phi, Q)$, is equal to the evidence at Level 1. Through this pattern, the optimization variables are gradually integrated out at different levels of Bayesian inference. Consequently, the optimal solutions obtained in subsequent layers of optimization are coordinated. However, direct optimization of all these three layers is still not a tractable problem and further simplification is required.

In order to obtain a tractable explicit solution to the above layered optimization problem, we adopt a hierarchical Bayesian approach through which the posterior probability density functions are sequentially approximated in each layer and the procedure is iterated. The hierarchical Bayesian approach has been applied to a great variety of problems. For instance, MacKay (1992) is the first author who proposed the heuristic Bayesian evidence framework and later on applied it to neural network modeling (MacKay, 1995). Molina *et al.* (2008) and Galatsanos *et al.* (2000) used the hierarchical Bayesian paradigm to address the image modeling and restoration problem. Kwok (2000) and Suykens *et al.* (2002) derived a probabilistic formulation of the least squares support vector machine (SVM) within a hierarchical Bayesian evidence framework.

5.6 Formulation of Inferential Modeling Problem in a Bayesian Framework

To derive analytical expressions for all levels of inference, here we use the popular autoregressive with exogenous input (ARX) model to illustrate the design of a robust unified Bayesian framework. The application of the ideas presented in this section is not limited to ARX models. The derivations can be directly extended to other classes of dynamic models, though numerical optimization may be required.

For fixed model orders na and nb, an ARX model is defined by introducing the

regression vector $\mathbf{r}_t \in \mathbb{R}^P$:

$$\mathbf{r}_t = [y_{t-1}, \ldots, y_{t-na}, \mathbf{u}_{t-1}^T, \ldots, \mathbf{u}_{t-nb}^T]^T \tag{5.15}$$

where $\mathbf{u}_t \in \mathbb{R}^M$ is the input and $P = na + M \cdot nb$.

The output can then be expressed as a linear function of r_t such that

$$y_t = \Theta^T \begin{bmatrix} \mathbf{r}_t \\ 1 \end{bmatrix} + e_t \tag{5.16}$$

where y_t is the output, e_t is a zero-mean Gaussian noise with non-constant variance and $\Theta = [\theta_1, \ldots, \theta_j, \theta_{P+1}]^T \in \mathbb{R}^{P+1}$ denotes the parameter vector including a subset of model parameters, $\Theta_{1:P} = [\theta_1, \ldots, \theta_j]^T$, and a bias term, θ_{P+1}.The reason for keeping $\Theta_{1:P}$ and θ_{P+1} distinct will become clear in deriving analytical expressions for the location outlier model.

Given the identification data-set that is contaminated by the presence of outliers, the objective is to identify model parameters Θ. The proposed hierarchical Bayesian optimization framework allows us to obtain MAP estimates of model parameters with an automated mechanism for determining the hyperparameters and investigating the quality of each data point.

5.6.1 Inference of Model Parameters Θ

Given the identification data-set $\mathcal{D} = \{(\mathbf{r}_t, y_t)\}_{t=1}^N = \{Z_t\}_{t=1}^N$ along with a set of indicator variables $\mathbf{q}_{1:N} = \{q_1, \cdots, q_t\}$ and the hyperparameters $\alpha_{1:P+1} = \{\alpha_1, \cdots, \alpha_{P+1}\} = \{\sigma_{\theta_1}^{-2}, \cdots, \sigma_{\theta_{P+1}}^{-2}\}$ and $\zeta = \sigma_e^{-2}$, the MAP estimates of model parameters are obtained by maximizing the posterior $p(\Theta|\mathcal{D}, \alpha_{1:P+1}, \zeta, \mathbf{q}_{1:N})$. Thus, the formulation of Bayes' Theorem in the first level of optimization becomes

$$p(\Theta|\mathcal{D}, \alpha_{1:P+1}, \zeta, \mathbf{q}_{1:N}) = \frac{p(\mathcal{D}|\Theta, \alpha_{1:P+1}, \zeta, \mathbf{q}_{1:N}) p(\Theta|\alpha_{1:P+1}, \zeta, \mathbf{q}_{1:N})}{p(\mathcal{D}|\alpha_{1:P+1}, \zeta, \mathbf{q}_{1:N})} \tag{5.17}$$

It is reasonable to assume that the prior distribution of each parameter $\theta_j \in \Theta$ is independent of hyperparameter ζ and indicator variables $\mathbf{q}_{1:N}$, *i.e.* $p(\theta_j|\alpha_j, \zeta, \mathbf{q}_{1:N}) =$

$p(\theta_j|\alpha_j)$. In the absence of other prior information, the prior distribution of Θ is taken as independent Gaussian with zero-mean and variance of $\sigma_{\theta_j}^2 = \alpha_j^{-1}$:

$$\begin{aligned} p(\Theta|\alpha_{1:P+1}) &= \prod_{j=1}^{P+1} p(\theta_j|\alpha_j) \\ &= \prod_{j=1}^{P+1} \sqrt{\frac{\alpha_j}{2\pi}} \exp\Big(-\frac{1}{2}\alpha_j\theta_j^2\Big) \end{aligned} \tag{5.18}$$

It is noteworthy that a set of independent hyperparameters $\{\alpha_1, ..., \alpha_{P+1}\}$ is specified in order to obtain sparsity. Considering that the bias could be any value, an uniform prior is chosen for θ_{P+1}; that is, $\alpha_{P+1} \to 0$ to approximate a uniform distribution, which can also be considered as a Gaussian distribution in the limit. Plugging in our assumptions, the prior is then expressed as follows:

$$p(\Theta|\alpha_{1:P}, \zeta, \mathbf{q}_{1:N}) \propto \prod_{j=1}^{P} \sqrt{\frac{\alpha_j}{2\pi}} \exp\Big(-\frac{1}{2}\alpha_j\theta_j^2\Big) \tag{5.19}$$

The chain rule of probability theory allows us to factorize joint probabilities as

$$\begin{aligned} p(\mathcal{D}) &= p(Z_1, Z_2, ..., Z_N) \\ &= \prod_{t=1}^{N} p(Z_t|Z_{1:t-1}) \end{aligned} \tag{5.20}$$

Given Θ, the sampled data $\mathcal{D}$ would be independent of hyperparameters $\alpha_{1:P}$ (inverse of the variance of the prior distribution of model parameters), *i.e.* $p(\mathcal{D}|\Theta, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N}) = p(\mathcal{D}|\Theta, \zeta, \mathbf{q}_{1:N})$. Applying the chain rule, therefore, the likelihood can be further expressed as

$$\begin{aligned} p(\mathcal{D}|\Theta, \zeta, \mathbf{q}_{1:N}) &= \prod_{t=1}^{N} p(Z_t|Z_{1:t-1}, \Theta, \zeta, q_t) \\ &\propto \prod_{t=1}^{N} p(e_t|\Theta, \zeta, q_t) \end{aligned} \tag{5.21}$$

where

$$p(e_t|\Theta, \zeta, q_t) = \sqrt{\frac{\zeta q_t}{2\pi}} \exp\Big(-\zeta q_t \frac{1}{2} e_t^2\Big) \tag{5.22}$$

if the identification data-set is contaminated with scale outliers and

$$p(e_t|\Theta,\zeta,q_t) = \sqrt{\frac{\zeta}{2\pi}}\exp\Big(-\zeta\frac{1}{2}(e_t-q_t)^2\Big) \tag{5.23}$$

if the identification data-set is contaminated with location outliers.

To be able to carry forward the derivations, we need to take the underlying outlier model into account.

5.6.1.1 Scale Outlier Model

Combining Equations 5.19 and 5.21 (along with 5.22), the posterior probability of the model parameters is then

$$\begin{aligned} p(\Theta|\mathcal{D},\alpha_{1:P},\zeta,\mathbf{q}_{1:N}) &\propto \exp\Big(-\frac{1}{2}\sum_{j=1}^{P}\alpha_j\theta_j^2 - \frac{1}{2}\sum_{t=1}^{N}\zeta q_t e_t^2\Big) \\ &= \exp\Big(-\sum_{j=1}^{P}\alpha_j E_{\theta_j} - \zeta\sum_{t=1}^{N} q_t E_{e_t}\Big) \\ &= \exp\Big(-\mathcal{J}_1(\Theta)\Big) \end{aligned} \tag{5.24}$$

where $E_{\theta_j} = \theta_j^2/2$ and $E_{e_t} = e_t^2/2$. All constants are neglected in Equation 5.24, because the optimal solution will not be affected by constant terms in the objective function.

One then proceeds to estimate the most probable values of the model parameters, Θ^{MP}, by maximizing the posterior probability, or equivalently, by minimizing the negative logarithm of Equation 5.24. The gradient of the cost function $\mathcal{J}_1(\Theta)$ is

$$\frac{\partial\mathcal{J}_1}{\partial\Theta_{1:P}} = \mathbf{D}_\alpha\Theta_{1:P} - \zeta\mathbf{R}\mathbf{D}_q\mathbf{y} + \zeta\mathbf{R}\mathbf{D}_q\mathbf{R}^T\Theta_{1:P} + \zeta\mathbf{R}\mathbf{D}_q\vec{\mathbf{1}}_N\theta_{P+1} \tag{5.25}$$

$$\frac{\partial\mathcal{J}_1}{\partial\theta_{P+1}} = \zeta\vec{\mathbf{1}}_N^T\mathbf{D}_q\mathbf{y} - \zeta\vec{\mathbf{1}}_N^T\mathbf{D}_q\mathbf{R}^T\Theta_{1:P} - \zeta s_q\theta_{P+1} \tag{5.26}$$

where $\vec{\mathbf{1}}_N = [1,...,1]^T \in \mathbb{R}^N$, $\mathbf{y} = [y_1,...,y_t]^T \in \mathbb{R}^N$, $\mathbf{R} = [\mathbf{r}_1,...,\mathbf{r}_N] \in \mathbb{R}^{P\times N}$, $\mathbf{D}_\alpha = \text{diag}(\alpha_1,...,\alpha_j) \in \mathbb{R}^{P\times P}$, $\mathbf{D}_q = \text{diag}(q_1,...,q_t) \in \mathbb{R}^{N\times N}$ and $s_q = \sum_{t=1}^{N} q_t$. Note that $\mathbf{D}_q$ may be viewed as a weighting matrix constructed to downplay the effect of scale outliers on the parameter estimates.

Making the partial derivatives expressed in Equations 5.25 and 5.26 equal to zero, the analytical expressions for $\Theta_{1:P}^{\text{MP}}$ and θ_{P+1}^{MP} can be derived:

$$\Theta_{1:P}^{\text{MP}} = \Big(\mathbf{RCR}^T + \frac{1}{\zeta}\mathbf{D}_\alpha\Big)^{-1}\mathbf{RCy} \tag{5.27}$$

$$\theta_{P+1}^{\text{MP}} = \frac{1}{s_q}\Big(\vec{\mathbf{1}}_N^T\mathbf{D}_q\mathbf{y} - \vec{\mathbf{1}}_N^T\mathbf{D}_q R^T\Theta_{1:P}^{\text{MP}}\Big) \tag{5.28}$$

where $\mathbf{C} = \mathbf{D}_q - {s_q}^{-1}\mathbf{D}_q\vec{\mathbf{1}}_N\vec{\mathbf{1}}_N^T\mathbf{D}_q$.

The posterior given by Equation 5.24 is complex in general and cannot be directly used for the three-layer optimization of Equation 5.11. The key to the hierarchical Bayesian approach is to obtain an approximation of the posterior. This approach of MacKay (2002) is adopted here to obtain an approximated solution first and then the optimization problem is solved through iteration. Approximating the logarithm of the posterior distribution by its second order Taylor expansion around $\Theta_{1:P+1}^{\text{MP}}$, we obtain

$$\begin{aligned}\log p(\Theta|\mathcal{D}, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N}) \approx \log\ & p(\Theta|\mathcal{D}, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N})\big|_{\Theta^{\text{MP}}} \\ & + \nabla \log p(\Theta|\mathcal{D}, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N})\big|_{\Theta^{\text{MP}}}\mathbf{m} \\ & + \frac{1}{2}\mathbf{m}^T\nabla\nabla \log p(\Theta|\mathcal{D}, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N})\big|_{\Theta^{\text{MP}}}\mathbf{m}\end{aligned} \tag{5.29}$$

where $\mathbf{m} = \big[\Theta - \Theta^{\text{MP}}\big]$.

Since Θ^{MP} corresponds to a maximum of the logarithm of the posterior, the second term on the right hand side of Equation 5.29 evaluates to zero. A Gaussian approximation of the posterior distribution can therefore be obtained as

$$\begin{aligned}p(\Theta|\mathcal{D}, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N}) &\approx p(\Theta^{\text{MP}}|\mathcal{D}, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N}) \exp\Big(-\frac{1}{2}\mathbf{m}^T\mathbf{Hm}\Big) \\ &= \frac{1}{\sqrt{(2\pi)^{(P+1)}\det\mathbf{H}^{-1}}} \exp\Big(-\frac{1}{2}\mathbf{m}^T\mathbf{Hm}\Big)\end{aligned} \tag{5.30}$$

where $\mathbf{H}$ is the Hessian of the cost function $\mathcal{J}_1(\Theta)$ evaluated at Θ^{MP}. The Hessian of the

cost function $\mathcal{J}_1(\Theta)$ is defined as

$$\begin{aligned}\mathbf{H} &= \begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{12}^T & \mathbf{H}_{22}\end{bmatrix}\\ &= \begin{bmatrix}\dfrac{\partial^2 \mathcal{J}_1}{\partial \Theta_{1:P}^2} & \dfrac{\partial^2 \mathcal{J}_1}{\partial \Theta_{1:P}\partial \theta_{P+1}}\\ \dfrac{\partial^2 \mathcal{J}_1}{\partial \theta_{P+1}\partial \Theta_{1:P}} & \dfrac{\partial^2 \mathcal{J}_1}{\partial \theta_{P+1}^2}\end{bmatrix}\end{aligned} \tag{5.31}$$

where

$$\mathbf{H}_{11} = \mathbf{D}_\alpha + \zeta \mathbf{R}\mathbf{D}_q\mathbf{R}^T \tag{5.32}$$

$$\mathbf{H}_{12} = \zeta \mathbf{R}\mathbf{D}_q\vec{\mathbf{1}}_N \tag{5.33}$$

$$\mathbf{H}_{22} = \zeta s_q \tag{5.34}$$

Using the Schur complement of the Hessian matrix, we obtain (Suykens *et al.*, 2002)

$$\mathbf{H} = \begin{bmatrix}\mathbf{I}_n & \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathbf{H}_{11} - \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{12}^T & 0\\ 0 & \mathbf{H}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{I}_n & 0\\ \mathbf{H}_{22}^{-1}\mathbf{H}_{12}^T & 1\end{bmatrix} \tag{5.35}$$

Hence,

$$\begin{aligned}\det H &= \det\begin{bmatrix}\mathbf{H}_{11} - \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{12}^T & 0\\ 0 & \mathbf{H}_{22}\end{bmatrix}\\ &= \mathbf{H}_{22}\det(\mathbf{H}_{11} - \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{12}^T)\\ &= \zeta s_q \det(\mathbf{D}_\alpha + \zeta\mathbf{G})\\ &= \zeta s_q \prod_{j=1}^{P}(\alpha_j + \zeta\lambda_{\mathbf{G},j})\end{aligned} \tag{5.36}$$

where $\lambda_{\mathbf{G},j}$ are the eigenvalues of the symmetric matrix $\mathbf{G} = \mathbf{R}\mathbf{C}\mathbf{R}^T$; the eigenvalue problem is

$$\mathbf{R}(\mathbf{D}_q - \frac{1}{s_q}\mathbf{D}_q\vec{\mathbf{1}}_N\vec{\mathbf{1}}_N^T\mathbf{D}_q)\mathbf{R}^T\nu_{\mathbf{G},j} = \lambda_{\mathbf{G},j}\nu_{\mathbf{G},j} \tag{5.37}$$

5.6.1.2 Location Outlier Model

Combining Equations 5.19 and 5.21 (along with 5.23) and neglecting all constants, the posterior probability of the model parameters is then

$$\begin{aligned} p(\Theta|\mathcal{D}, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N}) &\propto \exp\Big(-\frac{1}{2}\sum_{j=1}^{P}\alpha_j\theta_j^2 - \frac{1}{2}\sum_{t=1}^{N}\zeta(e_t - q_t)^2\Big) \\ &= \exp\Big(-\sum_{j=1}^{P}\alpha_j E_{\theta_j} - \zeta\sum_{t=1}^{N} E_{e_t'}\Big) \\ &= \exp\Big(-\mathcal{J}_2(\Theta)\Big) \end{aligned} \tag{5.38}$$

where $E_{e_t'} = E_{e_t} + q_t^2/2 - e_t q_t$.

The MAP estimates of the model parameters, Θ^{MP}, are obtained by maximizing the posterior probability, or equivalently, by minimizing the negative logarithm of Equation 5.38. The gradient of the cost function $\mathcal{J}_2(\Theta)$ is

$$\frac{\partial \mathcal{J}_2}{\partial \Theta_{1:P}} = \mathbf{D}_\alpha\Theta_{1:P} - \zeta\mathbf{R}\mathbf{y} + \zeta\mathbf{R}\mathbf{R}^T\Theta_{1:P} + \zeta R\vec{\mathbf{1}}_N\theta_{P+1} + \zeta\mathbf{R}\mathbf{D}_q\vec{\mathbf{1}}_N \tag{5.39}$$

$$\frac{\partial \mathcal{J}_2}{\partial \theta_{P+1}} = \zeta\vec{\mathbf{1}}_N^T\mathbf{y} - \zeta\vec{\mathbf{1}}_N^T\mathbf{R}^T\Theta_{1:P} - \zeta N\theta_{P+1} - \zeta\vec{\mathbf{1}}_N^T\mathbf{D}_q\vec{\mathbf{1}}_N \tag{5.40}$$

Note that $\mathbf{D}_q$ may be viewed as a correction matrix constructed to reduce the effect of location outliers on the parameter estimates.

Making the partial derivatives expressed in Equations 5.39 and 5.40 equal to zero, the analytical expression for $\Theta_{1:P}^{\text{MP}}$ and θ_{P+1}^{MP} can be derived:

$$\Theta_{1:P}^{\text{MP}} = \Big(\mathbf{R}\mathbf{C}'\mathbf{R}^T + \frac{1}{\zeta}\mathbf{D}_\alpha\Big)^{-1}\mathbf{R}\mathbf{C}'\Big(\mathbf{y} - \mathbf{D}_q\vec{\mathbf{1}}_N\Big) \tag{5.41}$$

$$\theta_{P+1}^{\text{MP}} = N^{-1}\vec{\mathbf{1}}_N^T\Big(\mathbf{y} - \mathbf{D}_q\vec{\mathbf{1}}_N - \mathbf{R}^T\Theta_{1:P}^{\text{MP}}\Big) \tag{5.42}$$

where $\mathbf{C}' = \mathbf{I}_n - N^{-1}\vec{\mathbf{1}}_N\vec{\mathbf{1}}_N^T$.

As explained previously, a Gaussian approximation of the posterior distribution is given as

$$p(\Theta|\mathcal{D}, \alpha_{1:P}, \zeta, \mathbf{q}_{1:N}) \approx \frac{1}{\sqrt{(2\pi)^{(P+1)}\text{det}\mathbf{H}^{-1}}}\exp\Big(-\frac{1}{2}\mathbf{m}^T\mathbf{H}\mathbf{m}\Big) \tag{5.43}$$

where $\mathbf{H}$ is the Hessian of the cost function $\mathcal{J}_2(\Theta)$ evaluated at Θ^{MP}.

It can be shown that the elements of the Hessian are

$$\mathbf{H}_{11} = \mathbf{D}_\alpha + \zeta \mathbf{R}\mathbf{R}^T \tag{5.44}$$

$$\mathbf{H}_{12} = \zeta \mathbf{R}\vec{\mathbf{1}}_N \tag{5.45}$$

$$\mathbf{H}_{22} = \zeta N \tag{5.46}$$

The Cholesky factorization of the Hessian would be similar to Equation 5.35. In order to obtain an expression for $\det H$, thus, one needs to solve the following eigenvalue problem:

$$\mathbf{R}(\mathbf{I}_n - \frac{1}{N}\vec{\mathbf{1}}_N\vec{\mathbf{1}}_N^T)\mathbf{R}^T\nu_{\mathbf{G},j} = \lambda'_{\mathbf{G},j}\nu_{\mathbf{G},j} \tag{5.47}$$

where $\lambda'_{\mathbf{G},j}$ are the eigenvalues of the symmetric matrix $\mathbf{G}' = \mathbf{R}\mathbf{C}'\mathbf{R}^T$.

Finally, we obtain

$$\begin{aligned} \det \mathbf{H} &= \zeta N \det(\mathbf{D}_\alpha + \zeta \mathbf{G}') \\ &= \zeta N \prod_{j=1}^{P} (\alpha_j + \zeta \lambda'_{\mathbf{G},j}) \end{aligned} \tag{5.48}$$

5.6.2 Inference of Hyperparameters $\alpha_{1:P}$ and ζ

Hyperparameters $\alpha_{1:P}$ and ζ are inferred from the identification data $\mathcal{D}$ by applying Bayes' rule in the second layer of optimization. First, the posterior distribution of the hyperparameters is written as

$$p(\alpha_{1:P}, \zeta|\mathcal{D}, \mathbf{q}_{1:N}) = \frac{p(\mathcal{D}|\alpha_{1:P}, \zeta, \mathbf{q}_{1:N})p(\alpha_{1:P}, \zeta|\mathbf{q}_{1:N})}{p(\mathcal{D}|\mathbf{q}_{1:N})} \tag{5.49}$$

As priors, it is assumed that the hyperparameters are statistically independent, *i.e.* $p(\alpha_{1:P}, \zeta|\mathbf{q}_{1:N}) = p(\zeta|\mathbf{q}_{1:N}) \prod_{j=1}^{P} p(\alpha_j|\mathbf{q}_{1:N})$. If there is no explicit information available for the hyperparameters, a uniform distribution can then be used to describe appropriate

non-informative priors on $\log \alpha_j$ and $\log \zeta$. To incorporate limited prior knowledge, however, **conjugate priors** (Raiffa and Schlaifer, 1961) are commonly assigned for which the resulting posterior distribution can be conveniently evaluated. To assure generality, we consider the following gamma distributions as hyperpriors:

$$\begin{aligned} p(\alpha_j|\mathbf{q}_{1:N}) &= \frac{s_j^{k_j}\alpha_j^{k_j-1}}{\Gamma(k_j)}\exp(-s_j\alpha_j) \\ &\propto \alpha_j^{k_j-1}\exp(-s_j\alpha_j) \end{aligned} \tag{5.50}$$

$$\begin{aligned} p(\zeta|\mathbf{q}_{1:N}) &= \frac{s_0^{k_0}\zeta^{k_0-1}}{\Gamma(k_0)}\exp(-s_0\zeta) \\ &\propto \zeta^{k_0-1}\exp(-s_0\zeta) \end{aligned} \tag{5.51}$$

where k_j is the shape parameter and s_j is the inverse of the scale parameter. Therefore, gamma distribution is a simple peaked distribution for which mean and variance are defined by k_j/s_j and k_j/s_j^2, respectively. The fact that the gamma distribution is the conjugate prior to many likelihood distributions justifies the choice of gamma hyperpriors.

Under the stated assumptions, the prior distribution over hyperparameters is expressed as

$$p(\alpha_{1:P},\zeta|\mathbf{q}_{1:N}) \propto \zeta^{k_0-1}\exp(-s_0\zeta)\prod_{j=1}^{P}\alpha_j^{k_j-1}\exp(-s_j\alpha_j) \tag{5.52}$$

Hereinafter, the underlying outlier model will be taken into account in order to lay out a computational procedure for the inference of hyperparameters.

5.6.2.1 Scale Outlier Model

The likelihood $p(\mathcal{D}|\alpha,\zeta,\mathbf{q}_{1:N})$ is equal to the normalizing constant in Equation 5.17 for the first level of inference. Substituting Equations 5.19, 5.21 (along with 5.22) and 5.30 in Equation 5.17, we can derive the following expression for the likelihood:

$$p(\mathcal{D}|\alpha_{1:P},\zeta,\mathbf{q}_{1:N}) \propto \frac{\prod_{j=1}^{P}\sqrt{\alpha_j}\prod_{t=1}^{N}\sqrt{\zeta q_t}}{\sqrt{\det \mathbf{H}}}\exp\Big(-\mathcal{J}_1(\Theta)+\frac{1}{2}\mathbf{m}^T\mathbf{H}\mathbf{m}\Big)\Big|_{\Theta^{\text{MP}}} \tag{5.53}$$

Substituting Equations 5.52 and 5.53 into Equation 5.49, the posterior probability of the hyperparameters becomes

$$p(\alpha_{1:P}, \zeta|\mathcal{D}, \mathbf{q}_{1:N}) \propto \sqrt{\frac{\zeta^N \prod_{j=1}^{P} \alpha_j \prod_{t=1}^{N} q_t}{\zeta s_q \prod_{j=1}^{P}(\alpha_j + \zeta\lambda_{\mathbf{G},j})}} \exp\Big(- \mathcal{J}_1\big(\Theta^{\text{MP}}\big)\Big) \\ \times \zeta^{k_0-1}\exp(-s_0\zeta)\prod_{j=1}^{P}\alpha_j^{k_j-1}\exp(-s_j\alpha_j) \tag{5.54}$$

Minimizing the negative logarithm of Equation 5.54 leads to the following optimization problem:

$$\min_{\alpha_{1:P},\zeta} \mathcal{J}_1(\alpha_{1:P},\zeta) = \sum_{j=1}^{P}\alpha_j\Big[s_j + E_{\theta_j}(\theta_j^{\text{MP}})\Big] + \zeta\Big[s_0 + \sum_{t=1}^{N} q_t E_{e_t}(\Theta^{\text{MP}})\Big] \\ - \frac{N+2k_0-3}{2}\log\zeta - \frac{1}{2}\sum_{j=1}^{P}(2k_j-1)\log\alpha_j + \frac{1}{2}\sum_{j=1}^{P}\log(\alpha_j + \zeta\lambda_{\mathbf{G},j}) \tag{5.55}$$

The gradient of the cost function $\mathcal{J}_1(\alpha_{1:P},\zeta)$ is

$$\frac{\partial\mathcal{J}_1}{\partial\alpha_{1:P}} = \Big(\mathbf{D}_s + \mathbf{E}_\Theta(\Theta^{\text{MP}}) + \frac{1}{2}(\mathbf{D}_\alpha + \zeta\mathbf{D}_\lambda)^{-1} - \frac{1}{2}\mathbf{D}_\alpha^{-1}(2\mathbf{D}_k - \mathbf{I}_P)\Big)\vec{\mathbf{1}}_P \tag{5.56}$$

$$\frac{\partial\mathcal{J}_1}{\partial\zeta} = s_0 + \vec{\mathbf{1}}_N^T\mathbf{D}_q\mathbf{E}_e(\Theta^{\text{MP}})\vec{\mathbf{1}}_N + \frac{1}{2}\vec{\mathbf{1}}_p^T\mathbf{D}_\lambda(\mathbf{D}_\alpha + \zeta\mathbf{D}_\lambda)^{-1}\vec{\mathbf{1}}_P - \frac{N+2k_0-3}{2\zeta} \tag{5.57}$$

where $\mathbf{D}_s = \text{diag}(s_1,\ldots,s_p) \in \mathbb{R}^{P\times P}$, $\mathbf{E}_\Theta = \text{diag}(E_{\theta_1},\ldots,E_{\theta_j}) \in \mathbb{R}^{P\times P}$, $\mathbf{E}_e = \text{diag}(E_{e_1},\ldots,E_{e_t}) \in \mathbb{R}^{N\times N}$, $\mathbf{D}_\lambda = \text{diag}(\lambda_{\mathbf{G},1},\ldots,\lambda_{\mathbf{G},P}) \in \mathbb{R}^{P\times P}$, $\mathbf{D}_k = \text{diag}(k_1,\ldots,k_P) \in \mathbb{R}^{P\times P}$, and $\vec{\mathbf{1}}_P = [1,...,1]^T \in \mathbb{R}^P$.

Setting the partial derivatives equal to zero and carrying out a few algebraic manipulations, the following expressions are obtained in the optimum of $\mathcal{J}_1(\alpha_{1:P},\zeta)$:

$$\mathbf{D}_\alpha^{\text{MP}} = \Big(\mathbf{D}_s + \mathbf{E}_\Theta(\Theta^{\text{MP}})\Big)^{-1}\Big(\frac{1}{2}\mathbf{D}_\gamma + \mathbf{D}_k - \mathbf{I}_P\Big) \tag{5.58}$$

$$\zeta^{\text{MP}} = \frac{1}{2}\Big(s_0 + \vec{\mathbf{1}}_N^T\mathbf{D}_q\mathbf{E}_e(\Theta^{\text{MP}})\vec{\mathbf{1}}_N\Big)^{-1}\Big(N + 2k_0 - 3 - \vec{\mathbf{1}}_p^T\mathbf{D}_\gamma\vec{\mathbf{1}}_P\Big) \tag{5.59}$$

where $\mathbf{D}_\gamma = \text{diag}(\gamma_1,\ldots,\gamma_P)$. The j^{th} diagonal element of $\mathbf{D}_\gamma$ is defined as

$$\gamma_j = \frac{\zeta^{\text{MP}}\lambda_{\mathbf{G},j}}{\alpha_j^{\text{MP}} + \zeta^{\text{MP}}\lambda_{\mathbf{G},j}} \tag{5.60}$$

where $\lambda_{\mathbf{G},j}$ is obtained by solving the eigenvalue problem of Equation 5.37.Thus, $\gamma_j \in [0,1]$ is a measure of the strength of the likelihood in relation to the prior in determining θ_j. For instance, $\gamma_j \longrightarrow 0$ (*i.e.* $\lambda_j \ll \alpha_j$) indicates that θ_j is poorly measured from the identification data. Consequently,

$$\begin{aligned}\gamma_{eff} &= 1 + \sum_{j=1}^{P} \frac{\zeta^{\text{MP}}\lambda_{\mathbf{G},j}}{\alpha_j^{\text{MP}} + \zeta^{\text{MP}}\lambda_{\mathbf{G},j}} \\ &= 1 + \vec{\mathbf{1}}_p^T \mathbf{D}_\gamma \vec{\mathbf{1}}_P \end{aligned} \tag{5.61}$$

is the number of well-determined parameters (MacKay, 1995).

Since $\alpha_{1:P}$ and ζ are positive scale variables, we can consider a separable Gaussian distribution for $p(\log \alpha_{1:P}, \log \zeta | \mathcal{D}, \mathbf{q}_{1:N})$ such that[†]

$$p(\log \alpha_{1:P}, \log \zeta | \mathcal{D}, \mathbf{q}_{1:N}) \approx \frac{1}{2\pi\sqrt{\det \mathbf{A}^{-1}}} \exp\Big(-\frac{1}{2}\mathbf{d}^T\mathbf{A}\mathbf{d}\Big) \tag{5.62}$$

where $\mathbf{d} = \begin{bmatrix}\log \alpha_{1:P} - \log \alpha_{1:P}^{\text{MP}} & \log \zeta - \log \zeta^{\text{MP}}\end{bmatrix}^T$ and $\mathbf{A}$ is Hessian of the cost function $\mathcal{J}_1(\alpha_{1:P}, \zeta)$ evaluated at $\alpha_{1:P}^{\text{MP}}, \zeta^{\text{MP}}$.

It is pointed out by MacKay (1999) that the Gaussian approximation over $\log \alpha_j^{\text{MP}}$ and $\log \zeta$ holds good if the model parameters are all well-determined in relation to their prior range by the identification data.

Having obtained MAP estimates of hyperparameters, the elements of the A are calculated as follows:

$$\begin{aligned}\mathbf{A}_{11} &= \frac{\partial^2 \mathcal{J}_1(\alpha_{1:P}, \zeta)}{\partial(\log \alpha_{1:P})^2}\Big|_{\alpha_{1:P}^{\text{MP}}, \zeta^{\text{MP}}} \\ &= \Big(\mathbf{D}_s + \mathbf{E}_\Theta(\Theta^{\text{MP}})\Big)\mathbf{D}_\alpha^{\text{MP}} + \frac{1}{2}\zeta^{\text{MP}}\mathbf{D}_\alpha^{\text{MP}}\mathbf{D}_\lambda\Big(\mathbf{D}_\alpha^{\text{MP}} + \zeta^{\text{MP}}\mathbf{D}_\lambda\Big)^{-2} \\ &\approx \Big(\mathbf{D}_s + \mathbf{E}_\Theta(\Theta^{\text{MP}})\Big)\mathbf{D}_\alpha^{\text{MP}} \\ &= \frac{1}{2}\mathbf{D}_\gamma + \mathbf{D}_k - \mathbf{I}_P \end{aligned} \tag{5.63}$$

[†]It is natural to represent the uncertainty associated with positive scale variables on a log scale.

$$\begin{aligned}\mathbf{A}_{22} &= \frac{\partial^2 \mathcal{J}_1(\alpha_{1:P},\zeta)}{\partial(\log\zeta)^2}\Big|_{\alpha_{1:P}^{\text{MP}},\zeta^{\text{MP}}} \\ &= \zeta^{\text{MP}}\Big(s_0 + \vec{\mathbf{1}}_N^T\mathbf{D}_q\mathbf{E}_e(\Theta^{\text{MP}})\vec{\mathbf{1}}_N\Big) + \frac{1}{2}\zeta^{\text{MP}}\vec{\mathbf{1}}_p^T\mathbf{D}_\alpha^{\text{MP}}\mathbf{D}_\lambda\Big(\mathbf{D}_\alpha^{\text{MP}} + \zeta^{\text{MP}}\mathbf{D}_\lambda\Big)^{-2}\vec{\mathbf{1}}_P \\ &\approx \zeta^{\text{MP}}\Big(s_0 + \vec{\mathbf{1}}_N^T\mathbf{D}_q\mathbf{E}_e(\Theta^{\text{MP}})\vec{\mathbf{1}}_N\Big) \\ &= \frac{1}{2}\big(N + 2k_0 - 2 - \gamma_{eff}\big) \end{aligned} \tag{5.64}$$

These approximations are valid if $\gamma_j + 2k_j - 2 \gg 1$ and $N + 2k_0 - 2 - \gamma_{eff} \gg 1$ (MacKay, 1999).

From Equations 5.63 and 5.64 it is straightforward to show that

$$\det\mathbf{A} = \frac{1}{2}(N + 2k_0 - 2 - \gamma_{eff})\prod_{j=1}^{P}(\frac{\gamma_j}{2} + k_j - 1) \tag{5.65}$$

5.6.2.2 Location Outlier Model

When the identification data-set is contaminated with location outliers, Equations 5.19, 5.21 (along with 5.23) and 5.43 are substituted in Equation 5.17 to obtain an expression for the likelihood:

$$p(\mathcal{D}|\alpha_{1:P},\zeta,\mathbf{q}_{1:N}) \propto \frac{\sqrt{\zeta^N}\prod_{j=1}^{P}\sqrt{\alpha_j}}{\sqrt{\det\mathbf{H}}}\exp\Big(-\mathcal{J}_2(\Theta) + \frac{1}{2}\mathbf{m}^T\mathbf{H}\mathbf{m}\Big)\Big|_{\Theta^{\text{MP}}} \tag{5.66}$$

Substituting Equations 5.52 and 5.66 into Equation 5.49, the posterior distribution of the hyperparameters $\alpha_{1:P}$ and ζ becomes

$$\begin{aligned} p(\alpha_{1:P},\zeta|\mathcal{D},\mathbf{q}_{1:N}) \propto &\sqrt{\frac{\zeta^N\prod_{j=1}^{P}\alpha_j}{\zeta N\prod_{j=1}^{P}(\alpha_j + \zeta\lambda'_{\mathbf{G},j})}}\exp\Big(-\mathcal{J}_2\big(\Theta^{\text{MP}}\big)\Big) \\ &\times \zeta^{k_0-1}\exp(-s_0\zeta)\prod_{j=1}^{P}\alpha_j^{k_j-1}\exp(-s_j\alpha_j) \end{aligned} \tag{5.67}$$

One can then proceed to infer the hyperparameters in a similar way as for the scale outlier model. The condition for optimality is thus expressed as

$$\mathbf{D}_\alpha^{\text{MP}} = \Big(\mathbf{D}_s + \mathbf{E}_\Theta(\Theta^{\text{MP}})\Big)^{-1}\Big(\frac{1}{2}\mathbf{D}_{\gamma'} + \mathbf{D}_k - \mathbf{I}_P\Big) \tag{5.68}$$

$$\zeta^{\text{MP}} = \frac{1}{2}\Big(s_0 + \vec{\mathbf{1}}_N^T \mathbf{E}_{e'}(\Theta^{\text{MP}})\vec{\mathbf{1}}_N\Big)^{-1}\Big(N + 2k_0 - 3 - \vec{\mathbf{1}}_p^T \mathbf{D}_{\gamma'}\vec{\mathbf{1}}_P\Big) \tag{5.69}$$

where $\mathbf{D}_{\gamma'} = \text{diag}(\gamma_1^{'}, \ldots, \gamma_P^{'})$. The j^{th} diagonal element of $\mathbf{D}_{\gamma}^{'}$ is defined as

$$\gamma_j = \frac{\zeta^{\text{MP}}\lambda_{\mathbf{G},j}^{'}}{\alpha_j^{\text{MP}} + \zeta^{\text{MP}}\lambda_{\mathbf{G},j}^{'}} \tag{5.70}$$

where $\lambda_{\mathbf{G},j}^{'}$ is obtained by solving the eigenvalue problem of Equation 5.47.

A separable Gaussian approximation of $p(\log \alpha_{1:P}, \log \zeta | \mathcal{D}, \mathbf{q}_{1:N})$ can be obtained as

$$p(\log \alpha_{1:P}, \log \zeta | \mathcal{D}, \mathbf{q}_{1:N}) \approx \frac{1}{2\pi\sqrt{\det \mathbf{A}^{-1}}} \exp\Big(-\frac{1}{2}\mathbf{d}^T\mathbf{A}\mathbf{d}\Big) \tag{5.71}$$

where $\mathbf{A}$ is Hessian of the cost function $\mathcal{J}_2(\alpha_{1:P}, \zeta)$ evaluated at $\alpha_{1:P}^{\text{MP}}, \zeta^{\text{MP}}$.

Finally, it can be shown that

$$\det \mathbf{A} = \frac{1}{2}(N + 2k_0 - 2 - \gamma_{eff}^{'})\prod_{j=1}^{P}(\frac{\gamma_j^{'}}{2} + k_j - 1) \tag{5.72}$$

5.6.3 Inference of Outlier Indicator Variables $\mathbf{q}_{1:N}$

So far in our derivations, we have assumed that the indicator variables $\mathbf{q}_{1:N}$ are known. Since $\mathbf{q}_{1:N}$ are unobserved variables, they still need to be estimated from the identification data-set. Applying Bayes' rule in the third level of optimization, we obtain the following posterior distribution:

$$p(\mathbf{q}_{1:N}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{q}_{1:N})p(\mathbf{q}_{1:N})}{p(\mathcal{D})} \tag{5.73}$$

where the prior distribution of $\mathbf{q}_{1:N}$ is expressed as

$$\begin{aligned} p(\mathbf{q}_{1:N}) &= \prod_{t=1}^{N} p(q_t) \\ &= \prod_{t=1}^{N} \delta^{\Big(1 - \frac{q_t - \rho}{1 - q_t\rho}\Big)}(1-\delta)^{\Big(\frac{q_t - \rho}{1 - q_t\rho}\Big)} \end{aligned} \tag{5.74}$$

and as

$$\begin{aligned} p(\mathbf{q}_{1:N}) &= \prod_{t=1}^{N} p(q_t) \\ &= \prod_{t=1}^{N} (0.5\delta) \Big(\frac{|q_t| - q_t}{2\Delta} \Big) (0.5\delta) \Big(\frac{|q_t| + q_t}{2\Delta} \Big) (1-\delta) \Big(1 - \frac{|q_t|}{\Delta} \Big) \\ &= \prod_{t=1}^{N} (0.5\delta) \Big(\frac{|q_t|}{\Delta} \Big) (1-\delta) \Big(1 - \frac{|q_t|}{\Delta} \Big) \end{aligned} \tag{5.75}$$

for the scale and location outliers, respectively. In deriving Equations 5.74 and 5.75, we assumed that the occurrence of the outliers is completely random.

The likelihood $p(\mathcal{D}|\mathbf{q}_{1:N})$ can be obtained by integrating over $\alpha_{1:P}$ and ζ, and then an approximate solution is obtained (MacKay, 1995):

$$\begin{aligned} p(\mathcal{D}|\mathbf{q}_{1:N}) &= \int p(\mathcal{D}|\mathbf{q}_{1:N}, \alpha_{1:P}, \zeta) p(\alpha_{1:P}, \zeta|\mathbf{q}_{1:N}) d\alpha_{1:P} d\zeta \\ &\approx p(\mathcal{D}|\mathbf{q}_{1:N}, \alpha_{1:P}^{\text{MP}}, \zeta^{\text{MP}}) p(\alpha_{1:P}^{\text{MP}}, \zeta^{\text{MP}}|\mathbf{q}_{1:N}) 2\pi \sqrt{\det \mathbf{A}^{-1}} \end{aligned} \tag{5.76}$$

At this stage, the type of outliers should be determined in order to obtain explicit expressions for evaluating the posterior probability of indicator variables.

5.6.3.1 Scale Outlier Model

Combining Equations 5.52, 5.53 and 5.65 and neglecting all constants, the likelihood of the third level of Bayesian inference is expressed as

$$\begin{aligned} p(\mathcal{D}|\mathbf{q}_{1:N}) \propto & \prod_{j=1}^{P} (\alpha_j^{\text{MP}})^{k_j - 1} \exp(-s_j \alpha_j^{\text{MP}}) \sqrt{\frac{\alpha_j^{\text{MP}}}{(0.5\gamma_j + k_j - 1)(\alpha_j^{\text{MP}} + \zeta^{\text{MP}} \lambda_{\mathbf{G},j})}} \\ & \times \sqrt{\frac{(\zeta^{\text{MP}})^{N+2k_0-3} \prod_{t=1}^{N} q_t}{s_q (N + 2k_0 - 2 - \gamma_{eff})}} \exp\Big(- s_0 \zeta^{\text{MP}} - \mathcal{J}_1\big(\Theta^{\text{MP}}\big) \Big) \end{aligned} \tag{5.77}$$

The posterior probability of the indicator variables is obtained by substituting Equations 5.74 and 5.77 into Equation 5.73:

$$p(\mathbf{q}_{1:N}|\mathcal{D}) \propto \prod_{j=1}^{P}(\alpha_j^{\text{MP}})^{k_j-1}\exp(-s_j\alpha_j^{\text{MP}})\sqrt{\frac{\alpha_j^{\text{MP}}}{(0.5\gamma_j+k_j-1)(\alpha_j^{\text{MP}}+\zeta^{\text{MP}}\lambda_{\mathbf{G},j})}} \times \sqrt{\frac{(\zeta^{\text{MP}})^{N+2k_0-3}}{s_q(N+2k_0-2-\gamma_{eff})}}\exp\Big(-s_0\zeta^{\text{MP}}-\mathcal{J}_1\big(\Theta^{\text{MP}}\big)\Big) \times \prod_{t=1}^{N}\delta^{\Big(1-\frac{q_t-\rho}{1-q_t\rho}\Big)}(1-\delta)^{\Big(\frac{q_t-\rho}{1-q_t\rho}\Big)}\sqrt{q_t} \tag{5.78}$$

To assess the quality of each data pair, $Z_t = (r_i, y_t)$, the posterior probability $p(q_t|\mathcal{D})$ is first evaluated for $q_t \in \{1, \rho\}$. The normalized probabilities are then used to estimate the expected value of q_t as follows:

$$\mathbb{E}[q_t|\mathcal{D}] = p(q_t = 1|\mathcal{D}) + \rho p(q_t = \rho|\mathcal{D}) \tag{5.79}$$

5.6.3.2 Location Outlier Model

Combining Equations 5.52, 5.66 and 5.72, the likelihood of the third level of Bayesian inference is expressed as

$$p(\mathcal{D}|\mathbf{q}_{1:N}) \propto \prod_{j=1}^{P}(\alpha_j^{\text{MP}})^{k_j-1}\exp(-s_j\alpha_j^{\text{MP}})\sqrt{\frac{\alpha_j^{\text{MP}}}{(0.5\gamma_j^{'}+k_j-1)(\alpha_j^{\text{MP}}+\zeta^{\text{MP}}\lambda_{\mathbf{G},j}^{'})}} \times \sqrt{\frac{(\zeta^{\text{MP}})^{N+2k_0-3}}{N(N+2k_0-2-\gamma_{eff}^{'})}}\exp\Big(-s_0\zeta^{\text{MP}}-\mathcal{J}_2\big(\Theta^{\text{MP}}\big)\Big) \tag{5.80}$$

Substituting Equations 5.75 and 5.80 into Equation 5.73, the posterior probability of the indicator variables becomes

$$p(\mathbf{q}_{1:N}|\mathcal{D}) \propto \prod_{j=1}^{P}(\alpha_j^{\text{MP}})^{k_j-1}\exp(-s_j\alpha_j^{\text{MP}})\sqrt{\frac{\alpha_j^{\text{MP}}}{(0.5\gamma_j^{'}+k_j-1)(\alpha_j^{\text{MP}}+\zeta^{\text{MP}}\lambda_{\mathbf{G},j}^{'})}} \times \sqrt{\frac{(\zeta^{\text{MP}})^{N+2k_0-3}}{N(N+2k_0-2-\gamma_{eff}^{'})}}\exp\Big(-s_0\zeta^{\text{MP}}-\mathcal{J}_2\big(\Theta^{\text{MP}}\big)\Big) \times \prod_{t=1}^{N}(0.5\delta)^{\Big(\frac{|q_t|}{\Delta}\Big)}(1-\delta)^{\Big(1-\frac{|q_t|}{\Delta}\Big)} \tag{5.81}$$

For the data pair, $Z_t = (\mathbf{r}_t, y_t)$, the posterior probability $p(q_t|\mathcal{D})$ is first evaluated over the set of possible values $q_t \in \{0, -\Delta, +\Delta\}$. The expected value of q_t is then estimated from the normalized probabilities:

$$\mathbb{E}[q_t|\mathcal{D}] = \Delta p(|q_t| = \Delta|\mathcal{D})\text{sign}\big[p(q_t = +\Delta|\mathcal{D}) - p(q_t = -\Delta|\mathcal{D})\big] \tag{5.82}$$

5.6.4 Robust Model Identification Procedure

To summarize our discussion, the implementation procedure of the proposed robust identification approach is outlined in Algorithm 5.1.

Algorithm 5.1. *Hierarchical Bayesian Optimization Framework for Robust Model Identification*

First, a few preparatory steps are completed to incorporate the relevant prior knowledge. Given a contaminated identification data-set,

1. Specify a set of indicator variables, $\mathbf{q}_{1:N} = \{q_1, \cdots, q_N\}$, to denote the quality of the observed data.

2. Select an appropriate outlier model to describe the contaminating distribution (Equations 5.5 and 5.7).

3. Include the noise distribution information to describe the prior distribution of $p(\mathbf{q}_{1:N})$ (Equations 5.74 and 5.75). In the absence of relevant prior information, the 3σ edit rule is used to detect potential outliers and hence to initialize the estimation of noise distribution parameters *i.e.* $\delta^{[0]}$, ${\sigma_e}^{[0]}$, and $\Delta^{[0]}$ or $\rho^{[0]}$.

4. Characterize the prior distribution of hyperparameters $p(\alpha_{1:P}, \zeta|\mathbf{q}_{1:N})$ based on the explicit prior knowledge. The prior information over hyperparameters can be generally well-represented by gamma distributions (Equations 5.50 and 5.51). If there is no explicit information available for the hyperparameters, a uniform

distribution can then be used to describe appropriate non-informative priors on $\log \alpha_j$ and $\log \zeta$.

5. Determine the prior distribution of model parameters $p(\Theta|\alpha_{1:P}, \zeta)$ based on the available background information. In the absence of other prior information, the prior probability of Θ can be approximated by independent Gaussian distributions (Equation 5.18). Depending on the model structure, it might be reasonable to assume that $\alpha_1 = \alpha_2 = ... = \alpha_j$.

6. Choose a set of initial values for indicator variables, $\mathbf{q}_{1:N}^{[0]}$, and hyperparameters, $\alpha_{1:P}^{[0]}$ and $\zeta^{[0]}$.

Next, the following steps will be repeated iteratively until no further improvements are gained:

1. Maximize $p(\Theta^{[k]}|\mathcal{D}, \alpha_{1:P}^{[k-1]}, \zeta^{[k-1]}, \mathbf{q}_{1:N}^{[k-1]})$ to update the MAP estimates of model parameters, $\Theta^{[k]}$ (Equations 5.27-5.28 and 5.41-5.42).

2. Maximize $p(\alpha_{1:P}^{[k]}, \zeta^{[k]}|\mathcal{D}, \mathbf{q}_{1:N}^{[k-1]})$ to update the MAP estimates of hyperparameters, $\alpha_{1:P}^{[k]}$ and $\zeta^{[k]}$ (Equations 5.58-5.59 and 5.68-5.69).

3. Evaluate the posterior probability of each observation acting as an outlier to update the MAP estimates of indicator variables, $\mathbf{q}_{1:N}^{[k]}$ (Equations 5.78-5.79 and 5.81-5.82); the updated estimates are used in the next iteration.

4. Update the estimated values of the noise distribution parameters, $\delta^{[k]}$, ${\sigma_e}^{[k]}$, and $\Delta^{[k]}$ or $\rho^{[k]}$, using the observations identified as outliers.

Although Gaussian approximations to posterior density functions may not always be adequate, the application of the robust identification procedure proposed in this Chapter is not limited to ARX models. For robust identification of non-linear models with non-Gaussian noise distributions, it is often required to adopt more sophisticated approximation

methods such as variational Bayes methods or Monte Carlo methods with various Bayesian sampling schemes. The derivations can thus be directly extended to other classes of dynamic models, though numerical optimization may be required.

5.7 Simulation and Experimental Study

In this section, we demonstrate the effectiveness of the proposed identification approach through the simulated and experimental data-sets. The purpose is to verify the performance of the Bayesian-based outlier detection algorithm and to evaluate the overall robust behavior of the proposed framework. The robustness of the Bayesian framework is compared with that of the Huber estimator, which is one of the most widely used methods of robust regression.

It is noteworthy that the M-estimation with various weighting functions were performed. In general, the results were similar to those of the Huber robust regression.

5.7.1 Second-order Finite Impulse Response Model

Consider the following linear second-order finite impulse response (FIR) model:

$$y_t = \begin{bmatrix} 6.5 & -2 & -1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_t \\ 1 \end{bmatrix} + e_t \tag{5.83}$$

with $\mathbf{x}_t = \begin{bmatrix} u_1(t) & u_2(t) \end{bmatrix}^T$. Three different scenarios will be considered to simulate noise distribution:

Case I. $e_t \sim 0.85\mathcal{N}(0, 4) + 0.15\mathcal{N}(0, 40)$

Case II. $e_t \sim 0.85\mathcal{N}(0, 2.25) + 0.15[\mathcal{N}(-5, 2.25), \mathcal{N}(5, 2.25)]$

Case III. $e_t \sim 0.85\mathcal{N}(0, 2.25) + 0.15\mathcal{N}(5, 2.25)$

Total number of data points is set to $N = 200$ in which around 15 percents have been generated from the contaminating distribution. The comparison is performed between the following methods using standard implementations:

1. Ordinary least square (OLS) regression: The most straightforward method for identification of ARX models is the OLS method, relying on minimization of the sum of squared errors between measurements and model predictions.

2. Regular Bayesian: The identification data-set used in the proposed Bayesian method is considered to be healthy *i.e.* $\delta = 0$.

3. Robust Regression: The M-estimation with the Huber weight function is performed. The tuning parameters of this algorithm are adjusted based on the recommended settings in MATLAB.

4. Robust Bayesian: The robust Bayesian framework does not require any knowledge of the noise distribution parameters as these parameters are actually iteratively estimated in the identification process. To illustrate how the proposed Bayesian identification framework is implemented, Figure 5.1 shows a detailed flowchart demonstrating the sequence of steps performed for Case I. Basic MATLAB commands can be used to execute each step.

Table 5.1 shows the mean relative estimation error (MRE) and mean squared error (MSE) of prediction averaged over 50 trials with different noise sequences. Not surprisingly, the robust methods improve the parameter estimation performance by detecting and accommodating outlying observations of the identification data-set. The smaller values of MRE indicate that the robust Bayesian framework outperforms the Huber robust regression in terms of accuracy of the parameter estimates. As a result, the models identified using robust Bayesian framework show better predictive performance, with smaller values of MSE. Specifically, the Huber robust regression can suffer from the effect of outliers when the contaminating distribution is asymmetric. In general, traditional robust regression methods assume a symmetric Gaussian distribution for the contaminating distribution and assign robustness weights accordingly. Therefore, in the case of asymmetric contaminating

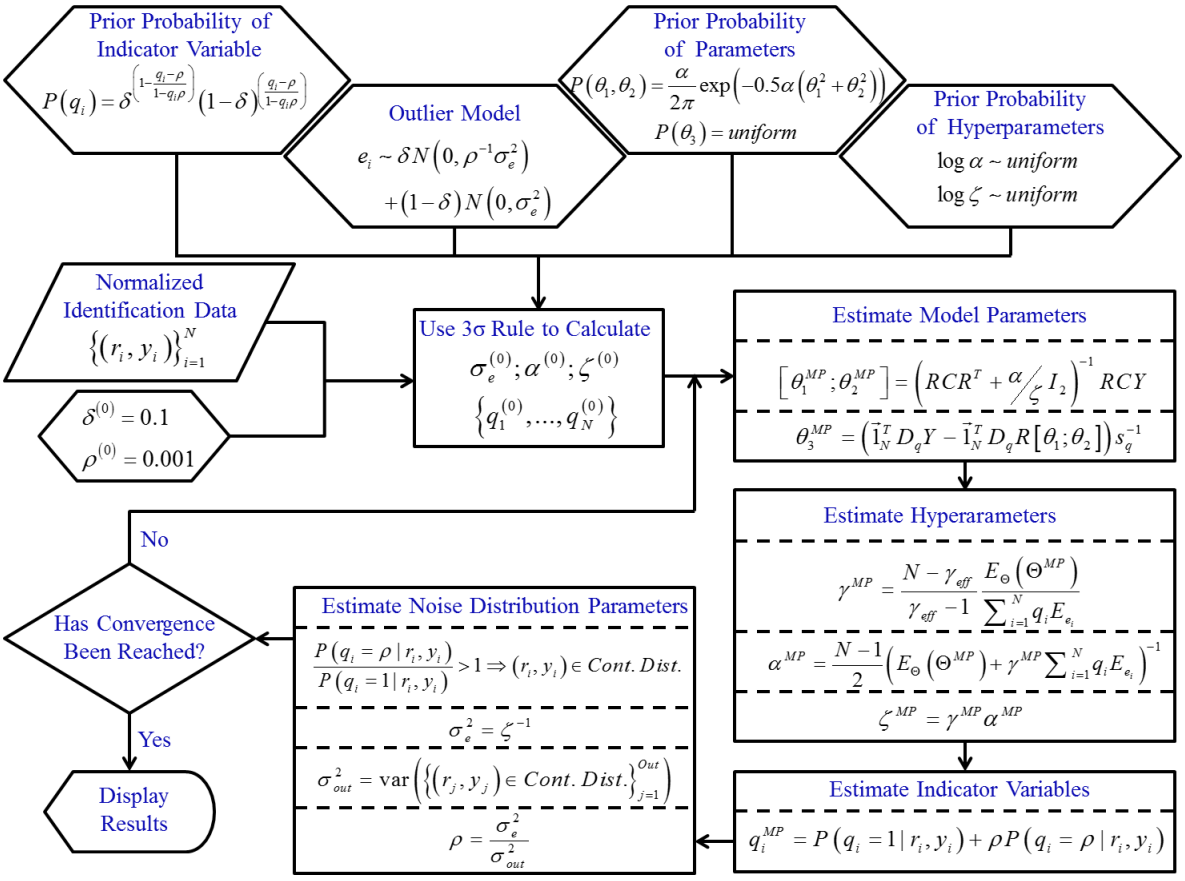


Figure 5.1: The flowchart of the Bayesian procedure followed for robust identification of the second-order FIR model in the presence of scale outlier

Table 5.1: Comparison of estimated parameters of the 2^{nd}-order FIR model

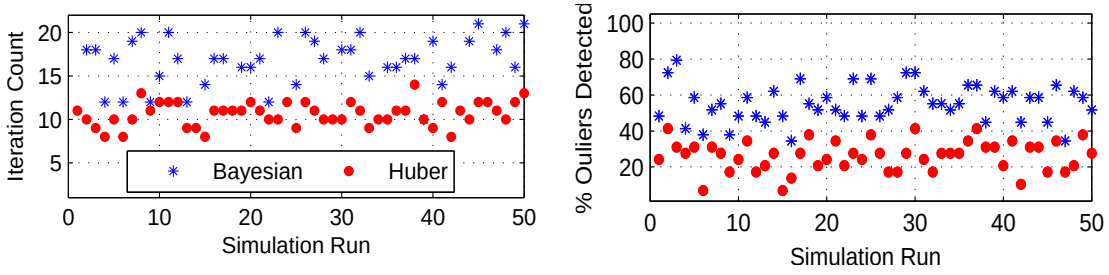
	OLS Regression	Robust Regression	Regular Bayesian	Robust Bayesian
Case I: Scale Outliers in the Identification Data				
MRE of θ_1 (%)	2.48 ± 1.84	2.02 ± 1.46	2.46 ± 1.84	1.86 ± 1.24
MRE of θ_2 (%)	3.96 ± 3.42	3.51 ± 2.42	3.95 ± 3.42	3.01 ± 2.15
MRE of θ_3 (%)	23.18 ± 15.63	16.30 ± 13.32	23.18 ± 15.63	12.27 ± 11.72
MSE of Prediction	0.161 ± 0.135	0.098 ± 0.093	0.160 ± 0.135	0.069 ± 0.074
Case II: Symmetric Location Outliers in the Identification Data				
MRE of θ_1 (%)	2.95 ± 2.04	2.27 ± 1.57	2.95 ± 2.03	1.08 ± 0.88
MRE of θ_2 (%)	4.17 ± 3.57	3.52 ± 2.65	4.16 ± 3.55	2.18 ± 1.98
MRE of θ_3 (%)	19.26 ± 12.79	14.48 ± 11.25	19.26 ± 12.78	7.44 ± 5.44
MSE of Prediction	0.149 ± 0.100	0.092 ± 0.069	0.148 ± 0.100	0.029 ± 0.030
Case III: Asymmetric Location Outliers in the Identification Data				
MRE of θ_1 (%)	1.80 ± 1.33	1.72 ± 1.22	1.76 ± 1.32	1.46 ± 0.97
MRE of θ_2 (%)	3.34 ± 2.70	2.85 ± 2.56	3.34 ± 2.70	2.59 ± 1.99
MRE of θ_3 (%)	65.30 ± 10.21	40.40 ± 11.77	65.3 ± 10.21	7.60 ± 7.04
MSE of Prediction	0.481 ± 0.137	0.215 ± 0.095	0.480 ± 0.136	0.042 ± 0.036

distribution (*e.g.* the noise term, e_t, is distributed as $e_t \sim \delta\mathcal{N}(\Delta, \sigma_e^2) + (1-\delta)\mathcal{N}(0, \sigma_e^2)$), downweighting the outliers causes a strong bias to the estimates.

Figures 5.2.a, 5.3.a, and 5.4.a show the number of iterations required for the convergence of model parameter estimates, while Figures 5.2.b, 5.3.b, and 5.4.b present the percentage of the outliers detected in the individual runs of Monte Carlo simulation. Although the iterations needed for the robust Bayesian and Huber methods are comparable, the former is capable of successfully detecting a higher percentage of outliers.

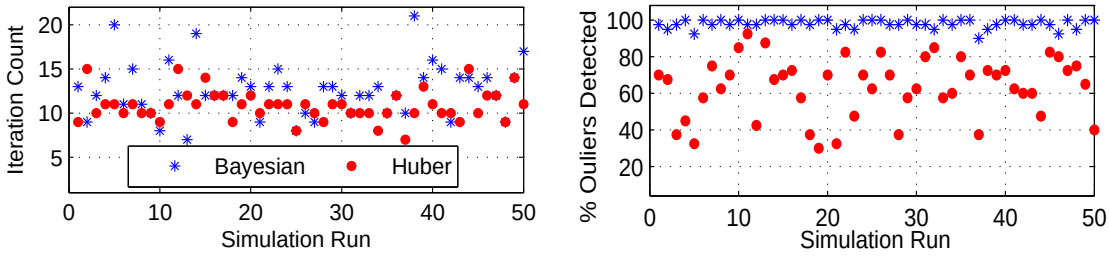
5.7.2 Continuous Fermentation Reactor Simulation

To illustrate potential applications of the proposed method in process industries, identification of a simulated continuous fermentation reactor is considered in this section. The non-linear dynamic behavior of a continuous fermentation reactor (CFR) is described



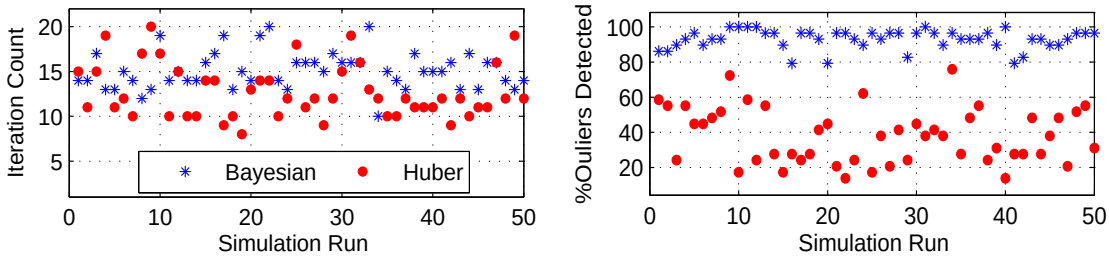
(a) Iteration counts (b) Percentage of outlier detected

Figure 5.2: Scale outlier



(a) Iteration counts (b) Percentage of outlier detected

Figure 5.3: Symmetric location outlier



(a) Iteration counts (b) Percentage of outlier detected

Figure 5.4: Asymmetric location outlier

as follows (Henson and Seborg, 1997):

$$\dot{X} = -DX + \mu X \tag{5.84}$$

$$\dot{S} = D(S_f - S) - \frac{1}{Y_{X/S}}\mu X \tag{5.85}$$

$$\dot{P} = -DP + (\alpha\mu + \beta)X \tag{5.86}$$

where specific growth rate (μ) is defined as

$$\mu = \frac{\mu_m\big(1 - \frac{P}{P_m}\big)S}{K_m + S + \frac{S^2}{K_i}} \tag{5.87}$$

X, S, and P are the state variables of the system representing the biomass concentration, substrate concentration and product concentration, respectively. Dilution rate (D) and feed substrate concentration (S_f) are normally treated as the system inputs. The cell-mass yield ($Y_{X/S}$), the yield parameters (α, β), the maximum specific growth rate (μ_m), the product saturation constant (P_m), the substrate saturation constant (K_m), and the substrate inhibition constant (K_i) are the model parameters. In this study, the case where the CFR has a single stable steady-state is considered for which the parameter settings and the operating conditions are given by Henson and Seborg (1997). The objective is to identify a multiple-input single-output (MISO) model relating the two input variables, dilution rate (u_1) and feed substrate concentration (u_2), with product concentration (y_1); the identification data-set is contaminated with the scale or location outliers. Dilution rate is assumed to vary between $0.13\ \mathrm{hr}^{-1}$ and $0.17\ \mathrm{hr}^{-1}$, while feed substrate concentration is assumed to vary between $18\ \mathrm{kg/m}^3$ and $22\ \mathrm{kg/m}^3$.

For both steady-state and dynamic modeling exercises presented below, Gaussian noise with a relative variance of 10% was added to the outputs. To test the robustness of the proposed Bayesian framework, several outliers are randomly added to the simulated identification data-set. To fairly investigate the performance of different identification procedures in the presence of outliers, Monte-Carlo simulation is performed. The

percentage of the observations generated from the contaminating distribution is fixed as 15%. Three different scenarios will be considered to simulate the contaminating distribution

Case I. Identification data-set is contaminated with scale outliers.

Case II. Identification data-set is contaminated with symmetric location outliers.

Case III. Identification data-set is contaminated with asymmetric location outliers.

The steady-state model to be identified is chosen as the form

$$y_1(k) = \theta_1 u_1(k) + \theta_2 u_2(k) + \theta_3 \tag{5.88}$$

We also consider the dynamic ARX-based identification of the fermentation problem in the neighbor of the nominal operating point to approximately capture the dynamic relationship between the input and output variables. The model to be identified is of the form

$$y_1(k) = \theta_1 u_1(k) + \theta_2 u_2(k) + \theta_3 y_1(k-1) + \theta_4 \tag{5.89}$$

OLS regression, regular Bayesian, Huber robust regression, and robust Bayesian are applied for identification of the steady-state and dynamic models. To evaluate the robustness of these methods, the prediction performance of the identified models is compared in Tables 5.2 and 5.3 for validation data-sets; the results are summarized from 50 simulation runs. Mean squared error (MSE), mean absolute error (MAE), and standard deviation of error (StdE) are the performance metrics evaluated. It can be observed that the models identified using robust Bayesian framework are both more accurate (with smaller MAE) and more reliable (with smaller StdE). Moreover, the relatively smaller values of MSE imply the overall better prediction performance in terms of both accuracy and reliability. The advantage of the proposed robust framework over the traditional robust regression techniques is highlighted specially when the identification data-set is contaminated with the location outliers.

Table 5.2: Comparison of the prediction performance of the identified steady-state models on the validation data

	OLS Regression	Robust Regression	Regular Bayesian	Robust Bayesian
Case I: Scale Outliers in the Identification Data				
MSE of Prediction	0.313 ± 0.067	0.286 ± 0.033	0.313 ± 0.045	0.250 ± 0.039
StdE of Prediction	0.230 ± 0.042	0.220 ± 0.031	0.230 ± 0.041	0.215 ± 0.024
MAE of Prediction	0.254 ± 0.059	0.230 ± 0.045	0.253 ± 0.045	0.198 ± 0.033
Case II: Symmetric Location Outliers in the Identification Data				
RMSE of Prediction	0.308 ± 0.076	0.249 ± 0.036	0.307 ± 0.075	0.208 ± 0.020
StdE of Prediction	0.265 ± 0.057	0.224 ± 0.030	0.265 ± 0.056	0.201 ± 0.018
MAE of Prediction	0.251 ± 0.067	0.202 ± 0.030	0.250 ± 0.066	0.165 ± 0.016
Case III: Asymmetric Location Outliers in the Identification Data				
RMSE of Prediction	0.487 ± 0.086	0.383 ± 0.059	0.486 ± 0.084	0.229 ± 0.039
StdE of Prediction	0.229 ± 0.040	0.227 ± 0.036	0.228 ± 0.038	0.212 ± 0.025
MAE of Prediction	0.432 ± 0.084	0.321 ± 0.054	0.432 ± 0.083	0.182 ± 0.029

Table 5.3: Comparison of the prediction performance of the identified dynamic models on the validation data

	OLS Regression	Robust Regression	Regular Bayesian	Robust Bayesian
Case I: Scale Outliers in the Identification Data				
RMSE of Prediction	0.462 ± 0.124	0.445 ± 0.120	0.460 ± 0.142	0.424 ± 0.119
StdE of Prediction	0.400 ± 0.120	0.394 ± 0.112	0.396 ± 0.129	0.387 ± 0.113
MAE of Prediction	0.363 ± 0.080	0.346 ± 0.074	0.363 ± 0.107	0.324 ± 0.078
Case II: Symmetric Location Outliers in the Identification Data				
RMSE of Prediction	0.478 ± 0.108	0.448 ± 0.081	$0.0.475 \pm 0.109$	0.409 ± 0.072
StdE of Prediction	0.435 ± 0.088	0.415 ± 0.072	0.433 ± 0.088	0.393 ± 0.066
MAE of Prediction	0.376 ± 0.090	$0.343 \pm 0.0.62$	0.374 ± 0.089	0.312 ± 0.054
Case III: Asymmetric Location Outliers in the Identification Data				
RMSE of Prediction	0.745 ± 0.113	0.633 ± 0.097	0.739 ± 0.111	0.438 ± 0.097
StdE of Prediction	0.390 ± 0.087	0.390 ± 0.086	0.390 ± 0.087	0.395 ± 0.089
MAE of Prediction	0.660 ± 0.116	0.541 ± 0.094	0.657 ± 0.112	0.341 ± 0.080

The averaged noise variance (σ_e^2) estimates along with the standard deviation of the estimated values obtained from each of the robust methods are presented in Table 5.4. The

Table 5.4: Comparison of the noise variance estimates obtained using different robust methods

	Simulated Value	Robust Regression	Robust Bayesian
Case I: Scale Outliers in the Identification Data			
Steady-State Model	0.490 ± 0.048	1.092 ± 0.124	0.408 ± 0.098
Dynamic Model	0.490 ± 0.040	1.387 ± 0.109	0.645 ± 0.126
Case II: Symmetric Location Outliers in the Identification Data			
Steady-State Model	0.494 ± 0.053	1.554 ± 0.204	0.451 ± 0.044
Dynamic Model	0.507 ± 0.046	2.201 ± 0.278	0.636 ± 0.073
Case III: Asymmetric Location Outliers in the Identification Data			
Steady-State Model	0.495 ± 0.046	1.219 ± 0.186	0.352 ± 0.066
Dynamic Model	0.512 ± 0.044	2.238 ± 0.293	0.462 ± 0.074

reported results show that the robust Bayesian outperforms the robust regression in terms of the accuracy of the noise variance estimates. Therefore, another advantage of the developed Bayesian framework is that it provides much more accurate estimates of hyperparameters such as the measurement noise variance.

5.7.3 Continuous Stirred Tank Heater Experiment

To further demonstrate the capability of the proposed Bayesian method, identification of an ARX model using the experimental data obtained from a pilot-scale continuous stirred tank heater (CSTH) is considered. The CSTH pilot plant is located in the Computer Process Control Laboratory in the Department of Chemical and Materials Engineering at the University of Alberta. As illustrated in Figure 5.5, the feed stream of the cold water flows into a well-stirred heated tank. The cold water is heated using saturated steam through a heating coil and drained from the tank through a long pipe (Thornhill *et al.*, 2008). Given a fixed volume of water in the tank, it is desired to heat the inlet stream to a higher setpoint temperature. To achieve this control objective the outflow temperature is measured and the steam flow rate is adjusted accordingly.

We consider the problem of identifying a dynamic model relating the steam flow rate

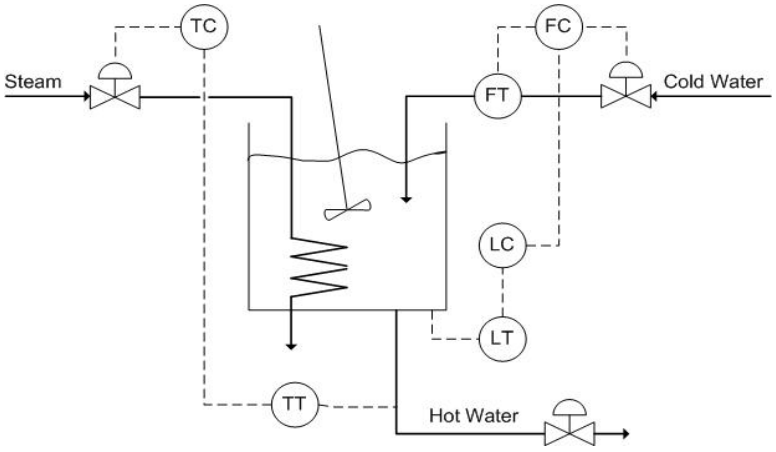


Figure 5.5: A simplified configuration of the CSTH

(input u) to the outlet water temperature (output y); the experimental data was taken from Jin (2010). A random binary sequence (RBS) based variation in the steam flow rate was used to sufficiently excite the process for collecting identification data; the input was varied between 10 kg/hr and 15 kg/hr. It is noteworthy that the level of water in the tank is controlled at 25 cm to isolate the significant effect of level variations on the process dynamics. The input-output data collected from the CSTH pilot plant is shown in Figure 5.6.

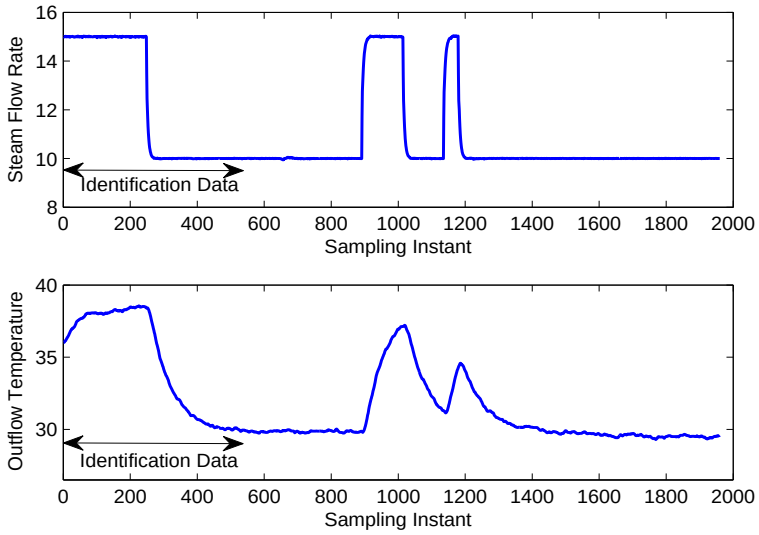


Figure 5.6: Input-output experimental data from a pilot scale CSTH

The identification data-set is used to identify empirical models of the form

$$y(k) = \theta_1 u(k) + \theta_2 y(k-1) + \theta_3 \tag{5.90}$$

OLS regression, regular Bayesian, Huber robust regression, and robust Bayesian are applied to estimate the model parameters. The prediction performance of the identified models is then tested on the validation data-set.

Since the main focus of this study is to investigate the robustness of different identification procedures, an identification data-set contaminated with outliers is of interest. Therefore, several outliers are randomly added to the identification data-set. It is expected that the presence of outliers will decrease the performance of various identification procedures. Thus, the parameter estimates obtained from the original identification data-set are considered as reference values.

Table 5.5 compares the parameter estimates obtained from the original data-set with the ones identified from the contaminated data-sets. In the absence of contamination, parameter estimation results from the investigated identification methods are comparable. Regardless of the form of contamination, however, presence of outliers in the identification data generally destroys the performance of non-robust estimators. Also, it can be clearly observed that the OLS regression and the regular Bayesian methods fail similarly. In contrast, the robust methods provide reasonably accurate parameter estimates, even when the identification data-set is contaminated by either scale or location outlying observations. Having included the contaminating model in the identification procedure, it is evident that the proposed Bayesian approach outperforms the Huber estimator in robustness especially in the presence of location outliers.

In order to evaluate the performance of the identified models, infinite horizon predictions (simulation) are performed on the validation data-set. The results are compared in Figures 5.7, 5.8, and 5.9. In the case of the contaminated identification data, the models identified through the use of non-robust methods exhibit poor prediction performance. However, it

Table 5.5: Comparison of estimated parameters of the CSTH model

	OLS Regression	Robust Regression	Regular Bayesian	Robust Bayesian
Case I: No Outlier in the Identification Data				
θ_1	0.0243	0.0248	0.0243	0.0254
θ_2	0.9866	0.9862	0.9866	0.9859
θ_3	0.1547	0.1599	0.1547	0.1639
Case II: Scale Outliers in the Identification Data				
θ_1	0.0537	0.0252	0.0539	0.0241
θ_2	0.9626	0.9856	0.9624	0.9865
θ_3	0.6236	0.1789	0.6271	0.1608
Case III: Symmetric Location Outliers in the Identification Data				
θ_1	0.0613	0.0285	0.0615	0.0250
θ_2	0.9566	0.9831	0.9564	0.9865
θ_3	0.7393	0.2212	0.7432	0.1513
Case IV: Asymmetric Location Outliers in the Identification Data				
θ_1	0.0642	0.0294	0.0644	0.0249
θ_2	0.9545	0.9823	0.9542	0.9860
θ_3	0.7834	0.2404	0.7838	0.1684

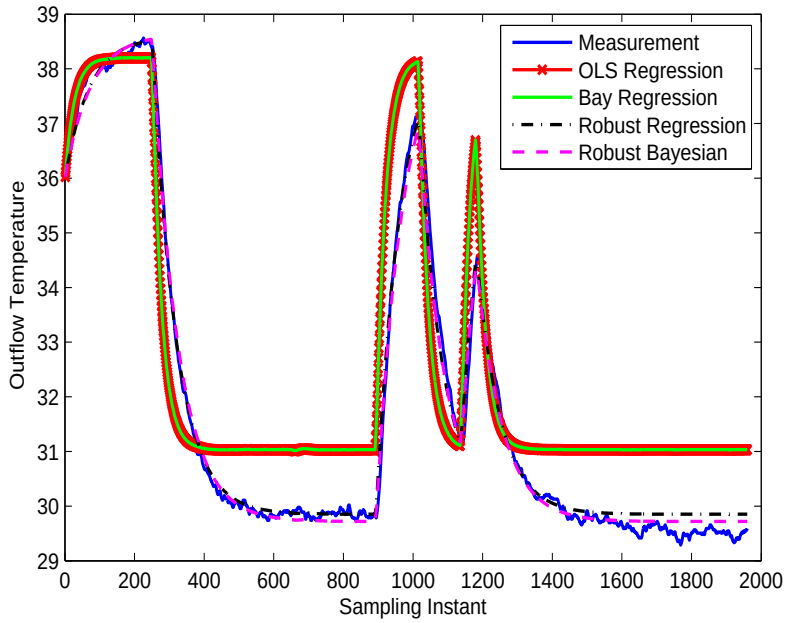


Figure 5.7: Prediction performance of the identified CSTH models; identification data-set is contaminated with the scale outliers.

can be observed that the robustness of the proposed Bayesian approach and Huber estimator significantly improve the predictive accuracy of the identified models. The prediction performance of the models identified using the robust Bayesian framework is better than that of the ones identified using Huber robust regression.

To summarize, this experimental study shows that the proposed robust Bayesian framework performs well under a wide variety of circumstances: with or without contamination, with scale or location outliers, and with symmetric or asymmetric contaminating distributions.

5.8 Concluding Remarks

Identification of ARX models in the presence of outliers was considered in this Chapter. To obtain a computationally feasible formulation, a set of indicator variables was introduced to denote the quality of each data point. Also, a contaminated Gaussian distribution

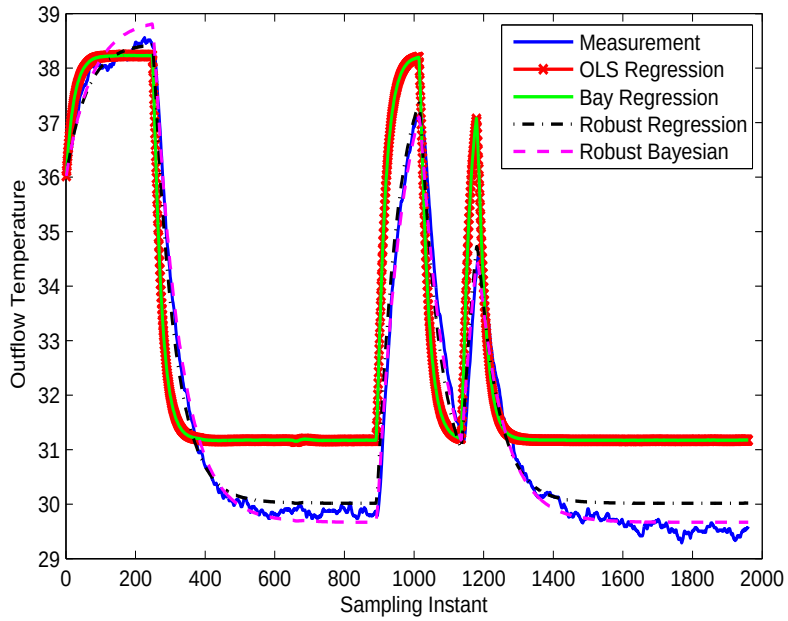


Figure 5.8: Prediction performance of the identified CSTH models; identification data-set is contaminated with the symmetric location outliers.

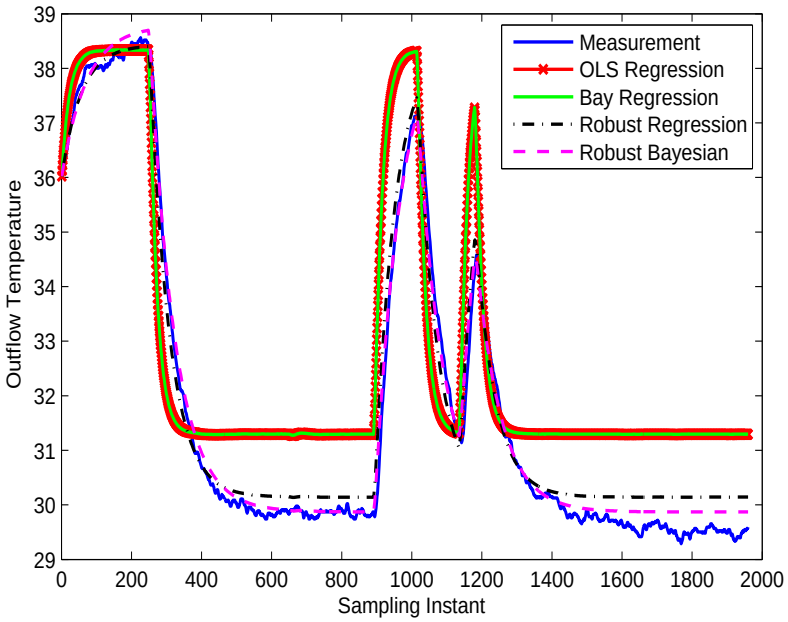


Figure 5.9: Prediction performance of the identified CSTH models; identification data-set is contaminated with the asymmetric location outliers.

was adopted to describe the observed data. The ARX identification problem was then formulated and solved under an iterative hierarchical Bayesian optimization framework. The layered optimization scheme allows us to obtain MAP estimates of model parameters with an automated mechanism for determining the hyperparameters and investigating the identity of each data point. The effectiveness of the developed framework for robust identification was demonstrated on the simulated and experimental data-sets. The layered optimization solution builds a unified framework that ensures that the model identification process is not significantly affected by outliers, which makes this method more applicable to the real world problems.

Bibliography

Agarwal, D. (2006). Detecting anomalies in cross-classified streams: a Bayesian approach. *Knowledge and Information Systems* **11**(1), 29–44.

Bishop, C. (1994). Novelty detection and neural network validation. *IEE Proceedings - Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks* **141**(4), 217–222.

Chandola, V., A. Banerjee and V. Kumar (2009). Anomaly detection : A survey. *ACM Computing Surveys* **41**(3), 124–129.

Chiang, L. H., R. J. Pell and M. B. Seasholtz (2003). Exploring process data with the use of robust outlier detection algorithms. *Journal of Process Control* **13**(5), 437–449.

Das, K. and J. Schneider (2007). Detecting anomalous records in categorical datasets. In: *Proceedings of the* 13^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press. San Jose, USA. pp. 220–229.

Fortuna, L., S. Graziani, A. Rizzo and M. G. Xibilia (2007). *Soft Sensors for Monitoring and Control of Industrial Processes*. first ed.. Springer-Verlag. London, UK.

Galatsanos, N. P., V. Z. Mesarović, R. Molina and A. K. Katsaggelos (2000). Hierarchical Bayesian image restoration from partially known blurs. *IEEE Transactions on Image Processing* **9**(10), 1784–1797.

Ghosh-Dastidar, B. and J. L. Schafer (2006). Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics* **22**(3), 487–506.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* **11**(1), 1–21.

Henson, M. A. and D. E. Seborg (1997). *Nonlinear Process Control*. first ed.. Prentice-Hall Inc.. Upper Saddle River, USA.

Hodge, V. J. and J. Austin (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review* **22**(2), 85–126.

Huber, P. J. (1981). *Robust Statistics*. first ed.. John Wiley & Sons. New York, USA.

Jin, X. (2010). Multiple arx model based identification for switching/nonlinear systems with em algorithm. Master's thesis. University of Alberta. Edmonton, Canada.

Kadlec, P., B. Gabrys and S. Strandt (2009). Data-driven soft sensors in the process industry. *Computers and Chemical Engineering* **33**(4), 795–814.

Kano, M. and Y. Nakagawa (2008). Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry. *Computers and Chemical Engineering* **32**(1-2), 12–24.

Khatibisepehr, S. and B. Huang (2008). Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Industrial and Engineering Chemistry Research* **47**(22), 8713–8723.

Khatibisepehr, S. and B. Huang (2012). Bayesian methods for process identification with outliers. In: *Proceedings of the 2012 American Control Conference (ACC)*. Montreal, Canada. pp. 3516–3521.

Khatibisepehr, S. and B. Huang (2013). A Bayesian approach to robust process identification with ARX models. *AIChE Journal* **59**(3), 845–859.

Kwok, J. T. (2000). The evidence framework applied to support vector machines. *IEEE Transactions on Neural Network* **11**(5), 1162–1173.

Lee, J., B. Kang and S. Kang (2011). Integrating independent component analysis and local outlier factor for plant-wide process monitoring. *Journal of Process Control* **21**(7), 1519–1528.

Liu, H., S. L. Shah and W. Jiang (2004). Online outlier detection and data cleaning. *Computers and Chemical Engineering* **28**(9), 1635–1647.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation* **4**(3), 415–447.

MacKay, D. J. C. (1995). Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* **6**(3), 469–505.

MacKay, D. J. C. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation* **11**(5), 1035–1068.

MacKay, D. J. C. (2002). *Information Theory, Inference, and Learning Algorithm*. first ed.. Cambridge University Press. New York, USA.

Molina, R., M. Vega, J. Mateos and A. K. Katsaggelos (2008). Variational posterior distribution approximation in Bayesian super resolution reconstruction of multispectral images. *Applied and Computational Harmonic Analysis* **24**(2), 251–267.

Muller, C. J., I. K. Craig and N. L. Ricker (2011). Modelling, validation, and control of an industrial fuel gas blending system. *Journal of Process Control* **21**(6), 852–860.

Prasad, V., M. Schley, L. P. Russo and B. Wayne Bequette (2002). Product property and production rate control of styrene polymerization. *Journal of Process Control* **12**(3), 353–372.

Raiffa, H. and R. Schlaifer (1961). *Applied Statistical Decision Theory*. first ed.. Harvard University. Boston, USA.

Ritter, G. and M. T. Gallegos (1997). Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters* **18**(6), 525–539.

Rousseeuw, P. and A. Leroy (1996). *Robust Regression and Outlier Detection*. second ed.. Wiley. New York, USA.

Sabbe, M. K., K. M. Van Geem, Marie Françoise Reyniers and G. B. Marin (2011). First principle-based simulation of ethane steam cracking. *AIChE Journal* **57**(2), 482–496.

Suykens, J. A. K., T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle (2002). *Least Squares Support Vector Machines*. first ed.. World Scientific. River Edge, USA.

Thornhill, N. F., S. C. Patwardhan and S. L. Shah (2008). A continuous stirred tank heater simulation model with applications. *Journal of Process Control* **18**(3-4), 347–360.

Varbanov, A. (1998). Bayesian approach to outlier detection in multivariate normal samples and linear models. *Communications in Statistics - Theory and Methods* **27**(3), 547–557.

Yu, J. (2012). A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. *Computers and Chemical Engineering* **41**(11), 134–144.

Zeng, J. and C. Gao (2009). Improvement of identification of blast furnace ironmaking

process by outlier detection and missing value imputation. *Journal of Process Control* **19**(9), 1519–1528.

Chapter 6

A Bayesian Approach to Design of Adaptive Multi-model Inferential Sensors with Application in Oil Sand Industry

6.1 Introduction

6.1.1 Practical Motivation

The design of inferential sensors finds its roots in process modeling. In general, inferential sensors produce valid results only for a particular region in which the underlying models have been identified. Hence, proper identification of a representative process model is key to the design of an efficacious inferential sensor. Having established the intended application of the sensor, selection of an optimal model structure that best captures the behavior of the system would be the first step in any model identification procedure. Depending on the level of *a priori* knowledge of the process, two different philosophies may guide the choice of model structure (Ljung, 1999): 1. First principles analysis and 2. Process data analysis. Regardless of the source of available information, inferential model structures can be further characterized as static and dynamic models; to develop a dynamic model the temporal dimension is added to the otherwise static model. Since

A version of this chapter has been published in Journal of Process Control, Volume 22 (Khatibisepehr and Huang, 2012).

many unit operations are naturally subject to temporal variations, industrial processes need to be modeled as dynamic systems in order to capture their time-varying characteristic. Nonetheless, some chemical processes experience discrete changes superimposed on their predominantly continuous dynamic behavior. The continuous-state dynamics is typically associated with physical phenomena involved, while the discrete-state dynamics may come from switching controllers, inherent non-linearities in the system, different operating conditions, or any other external discrete events influencing the process under investigation. In such applications, only **multi-model** inferential sensors can describe both the continuous dynamic behavior and the transitions between discrete modes (Murray-Smith and Johansen, 1997; Azimzadeh *et al.*, 1998).

The multi-model paradigm has increasingly attracted attention in process control community because of its many potential industrial applications (Paoletti *et al.*, 2007; Lauer, 2008). Hybrid models with multi-model structures have been adopted to represent time-varying dynamic behavior of industrial processes for prediction, estimation, or control purposes. The polymer industry is a typical application field where modeling of hybrid systems is of great interest. Suitable production policies drive a single polymer manufacturing plant switching among various operating conditions to produce many different grades; this system has multiple modes or regimes of behavior. Kim *et al.* (2005) have designed a clustering-based hybrid inferential sensor for an industrial Polypropylene process with grade changeover operation. Kadlec and Gabrys (2011) have designed an adaptive inferential sensor based on the local learning framework in order to predict the catalyst activation rate in a polymerization reactor. Angelov and Kordon (2010) have developed an adaptive multi-model inferential sensor based on the concept of evolving fuzzy models for product composition estimation in a distillation tower. Multi-model inferential sensors can be used to represent complex processes by concatenating multiple models with simple structures. For instance, Domlan *et al.* (2011) have developed

an inferential sensor with decoupled multiple model structure to approximate the non-linear dynamic behavior of a separation unit by switching among various piecewise linear models.s

The problem of multi-modal system identification has been considered widely and several approaches have been proposed such as algebraic procedure (Vidal *et al.*, 2003), the clustering-based procedure (Ferrari-Trecate *et al.*, 2003), the Bayesian procedure (Juloski, 2004; Juloski *et al.*, 2005), the EM-based procedure (Jin and Huang, 2010), and the bounded-error procedure (Bemporad *et al.*, 2005). The recursive identification procedure implemented in most of the aforementioned approaches comprises three steps. First, the operating space is partitioned into a finite number of non-overlapping regions. Next, identification data is attributed to relevant regions based on descriptive classification criteria; the identification data-set is divided into multiple exclusive sub-sets. Finally, the standard identification techniques are applied to develop sub-models that best describe the associated regions; the identified sub-models would be well supported by the corresponding identification data sub-sets.

Despite the increasing number of publications dealing with identification of multi-modal systems, yet several challenging issues remain open (Paoletti *et al.*, 2007; Lauer, 2008). Firstly, most of the existing methods focus mainly on switched linear system identification or piecewise affine function approximation. The identification of hybrid systems switching between non-linear continuous-state dynamics have not been extensively studied. Another issue to be considered is related to overlapping regions. Many of the identification procedures hinge on the assumption that an operating space can be partitioned into linearly separable regions. There are two main approaches available in the literature to deal with data points lying in the proximity of the intersection of multiple regions. In the first category of approaches, attributes with undecidable data points are discarded during the classification step (Bemporad *et al.*, 2003; Juloski, 2004). This could be a feasible

solution when there are only few undecidable data points; otherwise, excluding a large number of identification data from analysis may lead to a considerable loss of information and biased estimates. In the second category of approaches, each attribute is first classified to one and only one region. Next, a refinement step based on the certainly attributed closest neighbors is considered (Bemporad *et al.*, 2005; Jin and Huang, 2010). Although the size of the identification data-set is preserved in such methods, the misclassified data points could decrease the accuracy and generalization performance of inferential sub-models. Finally, available process knowledge cannot be easily incorporated in many of the current formalisms. Thus, including relevant prior information in the identification process is another important issue to be addressed.

6.1.2 Main Contributions

The main contribution of this work is to present a novel Bayesian procedure for the development and implementation of adaptive multi-model inferential sensors for industrial applications. The proposed approach provides a framework to accommodate the overlapping regions, facilitates the inclusion of prior knowledge about the operating conditions, and implements local adaptation mechanisms. An equally important contribution of this research is to demonstrate practicality and validity of the proposed approach through a successful application in the oil sand industry. Oil sands development is both a costly and technically complex business with potential environmental impacts due to land use, water consumption and air emissions. Therefore, it is of practical interest to further investigate techniques for design of inferential sensors in virtually all areas of this industry to improve process operations and control to reduce environmental footprints, improve recovery and lower the production costs of bitumen. Even a small incremental increase of less than 1% in Alberta oil sands production, for instance, can increase the total annual revenue of the producers by several millions of dollars in a typical oil sands complex (Dougan and McDowell, 1997).

6.1.3 Chapter Outline

The remainder of this Chapter is organized as follows. In Section 6.2, the identification problem of interest is formulated. Section 6.3 presents a Bayesian procedure for the design of multi-model inferential sensors. Section 6.4 presents a general procedure for the implementation of multi-model inferential sensors and discusses the importance of adaptation mechanisms for maintaining the on-line performance. In Section 6.5, the efficacy of the proposed procedures is demonstrated through a simulation case study. An adaptive multi-model inferential sensor is developed to predict the product concentration of a continuous stirred tank reactor which is a benchmark example of a process with non-linear dynamics. In Section 6.6, the effectiveness of the proposed Bayesian approach is further highlighted through an industrial case study. The objective was to design an inferential sensor for real-time monitoring of a key quality variable of an oil sands processing unit. Finally, Section 6.7 summarizes this Chapter with some concluding remarks.

6.2 Problem Statement

Consider an input-output representation of a multi-modal system expressed as

$$\begin{cases} y_t^{(m)} = f^{(m)}\big(\mathbf{r}_t; \Theta^{(m)}\big) \quad m = 1, \cdots, M \\ y_t = \sum_{m=1}^{M} \psi_t^{(m)}(s_t) y_t^{(m)} + \varepsilon_t \end{cases} \tag{6.1}$$

where $\mathbf{r}_t$ is the regressor vector constructed from the lagged outputs and inputs, M is the number of sub-models, and ε_t is the error term.

As suggested by Equation 6.1, the design of a multi-model inferential sensor comprises two steps. First, each sub-model, $m \in \{1 \cdots, M\}$, is represented by its functional form, $f^{(m)}$, and a set of corresponding parameters, $\Theta^{(m)}$. Next, a proper interpolation function is defined to assign an importance weight, $\psi^{(m)}$, to the output of each sub-model, $y^{(m)}$, in order to combine the information included in a set of local sub-models into a global predictive model. The interpolation function is often parameterized by a scheduling

variable, s_t, that effectively determines the discrete-state dynamics at time instant t. The choice of a suitable scheduling variable is problem specific.

A general introduction to the identification of multi-modal systems and a good review of several identification techniques are provided by Paoletti *et al.* (2007); Lauer (2008). The focus of this study is on Bayesian methods, which have the potential to overcome the aforementioned shortcomings of the other approaches in the design of inferential sensors.

6.3 Bayesian Approach for Design of Multi-model Inferential Sensors

A Bayesian identification procedure was proposed by Juloski *et al.* (2005) for piecewise autoregressive exogenous (PWARX) models and was extended by Juloski and Weiland (2006) for piecewise output error (PWOE) models. First, each attribute is classified to the mode with the highest probability by sequential processing of the identification data points. Next, Bayesian parameter estimation is performed to identify each sub-model from the corresponding data. The limitation of the described procedure is that the operating space is partitioned into a finite number of linearly separable regions *i.e.* at each time instant only one mode is active. In the context of process industries, however, the operating modes are often overlapped or have non-linear boundaries in continuous unit operations. Moreover, the classification rule only relies on evaluating the residuals obtained from each sub-model. Thus, the available information about the process operation cannot be fully incorporated in the identification procedure. Finally, if the identification data is not linearly separable or if the relevant residuals are comparable, the violating attributes are excluded from analysis. In this work, a Bayesian procedure is proposed in order to accommodate the overlapping regions and facilitate the inclusion of prior knowledge about the operating conditions.

The first step in the identification of industrial multi-model inferential sensors is to thoroughly investigate all the available source of information (*e.g.* first principles,

operational data, and etc.). The purpose of such preliminary studies is twofold. Firstly, it is desired to identify possible operating modes and, consequently, to determine the number of sub-models. Secondly, it is important to select a representative scheduling variable, through which the operating space can be properly partitioned. There are a few criteria that may guide the choice of a suitable scheduling variable:

1. A scheduling variable should reflect changes in plant dynamics as operating conditions vary.

2. A slowly-varying scheduling variable would guarantee smooth transitions between different sub-models.

3. The availability of real-time measurements of a scheduling variable is one of the key requirements for successful application of a multi-model inferential sensor.

From a Bayesian point of view, the interpolation function introduced in Equation 6.1 is defined as

$$\psi_t^{(m)} = p(i_t = m|s_t) \tag{6.2}$$

where i_t is the discrete-state dynamics at time instant t and $p(i_t = m|s_t)$ denotes the conditional probability of the m^{th} sub-model capturing the discrete-state dynamics at time instant t, given the scheduling variable s_t.

The above posterior probability can be evaluated using Bayes' theorem:

$$p(i_t = m|s_t) = \frac{p(s_t|i_t = m)p(i_t = m)}{\sum_{m=1}^{M} p(s_t|i_t = m)p(i_t = m)} \tag{6.3}$$

where $p(s_t|i_t = m)$ is the likelihood that s_t was generated by the m^{th} mode and $p(i_t = m)$ is the prior probability that the system operates in the m^{th} mode.

Given the identification data-set $\mathcal{D} = \{(s_t, \mathbf{r}_t, y_t)\}_{t=1}^{N} = \{Z_t\}_{t=1}^{N}$, the parameter and mode estimation problem is formulated under a Bayesian framework as follows

$$p(\{\Theta^{(m)}\}_{m=1}^{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\{\Theta^{(m)}\}_{m=1}^{M})p(\{\Theta^{(m)}\}_{m=1}^{M})}{p(\mathcal{D})} \tag{6.4}$$

where $\Theta^{(m)} = \left[\ \theta_1^{(m)}, \cdots, \theta_p^{(m)} \ \right]^T$.

The available knowledge regarding the model parameters is exploited by specifying informative prior distributions. In the absence of other relevant information, the prior probability of $\{\Theta^{(m)}\}_{m=1}^{M}$ can be defined as independent multivariate Gaussian distributions, such that

$$\begin{aligned} p(\{\Theta^{(m)}\}_{i=1}^{m}) &= \prod_{m=1}^{M} p(\Theta^{(m)}) \\ &\propto \prod_{m=1}^{M} \exp\Bigg(-\frac{1}{2}\Big(\Theta^{(m)} - \Theta_0^{(m)}\Big)^T \Sigma_{\Theta_0^{(m)}}^{-1} \Big(\Theta^{(m)} - \Theta_0^{(m)}\Big)\Bigg) \end{aligned} \tag{6.5}$$

where $\Theta_0^{(m)}$ denotes the explicitly specified expected values of $\Theta^{(m)}$ and

$$\Sigma_{\Theta_0^{(m)}} = \begin{bmatrix} \sigma^2_{\theta_1^{(m)}} & \sigma_{\theta_1^{(m)}}\sigma_{\theta_2^{(m)}} & \cdots & \sigma_{\theta_1^{(m)}}\sigma_{\theta_p^{(m)}} \\ \sigma_{\theta_1^{(m)}}\sigma_{\theta_2^{(m)}} & \sigma^2_{\theta_2^{(m)}} & \cdots & \sigma_{\theta_2^{(m)}}\sigma_{\theta_P^{(m)}} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{\theta_1^{(m)}}\sigma_{\theta_P^{(m)}} & \sigma_{\theta_2^{(m)}}\sigma_{\theta_P^{(m)}} & \cdots & \sigma^2_{\theta_P^{(m)}} \end{bmatrix} \tag{6.6}$$

represents the prior degree of belief over possible values of $\Theta^{(m)}$ that are centered around $\Theta_0^{(m)}$ in the parameter space. As mentioned previously, $\Theta_0^{(m)}$ is the parameter vector selected based on the available prior information. Thus, lack of prior knowledge about specific parameters would be quantitatively demonstrated by the relatively large values of the elements of covariance matrices, $\{\Sigma_{\Theta_0^{(m)}}\}_{m=1}^{M}$. For instance, if $\theta_1^{(1)}$ is specified based on the vague information, the high prior uncertainty is expressed through $0 \ll \sigma^2_{\theta_1^{(1)}}$. Hence, the importance of the non-informative or imprecise priors can be considerably reduced.

Applying the chain rule of probability theory, the likelihood can be expressed as

$$\begin{aligned} p(\mathcal{D}|\{\Theta^{(m)}\}_{m=1}^{M}) &= \prod_{t=1}^{N} p(Z_t|\{\Theta^{(m)}\}_{m=1}^{M}) \\ &\propto \prod_{t=1}^{N} \exp\Bigg(-\frac{1}{2\sigma_\varepsilon^2}\Big(y_t - \sum_{m=1}^{M} \psi_t^{(m)} \hat{y}_t^{(m)}\Big)^2\Bigg) \end{aligned} \tag{6.7}$$

where N is the size of identification data-set, y_t is the measured value, $\hat{y}_t^{(m)} = f^{(m)}\big(\mathbf{r}_t; \Theta^{(m)}\big)$ is the predicted value, and σ_ε^2 denotes the variance of the prediction error.

Substituting Equations 6.5 and 6.7 in Equation 6.4, the posterior probability of the model parameters becomes

$$p(\{\Theta^{(m)}\}_{m=1}^{M}|\mathcal{D}) \propto \prod_{t=1}^{N} \exp\Bigg(-\frac{1}{2\sigma_{\varepsilon}^{2}}\Big(y_t - \sum_{m=1}^{M} \psi_t^{(m)} \hat{y}_t^{(m)}\Big)^2\Bigg) \times \prod_{m=1}^{M} \exp\Bigg(-\frac{1}{2}\Big(\Theta^{(m)} - \Theta_0^{(m)}\Big)^T \Sigma_{\Theta_0^{(m)}}^{-1}\Big(\Theta^{(m)} - \Theta_0^{(m)}\Big)\Bigg) \tag{6.8}$$

The MAP estimates of the model parameters are obtained by maximizing the posterior probability, or equivalently, by minimizing the negative logarithm of Equation 6.8. The parameter estimation problem is then posed as

$$\min_{\{\Theta^{(m)}\}_{m=1}^{M}} \mathcal{J}_N\big(\{\Theta^{(m)}\}_{m=1}^{M}\big) = \frac{1}{2\sigma_{\varepsilon}^{2}} \sum_{t=1}^{N} \Big(y_t - \sum_{m=1}^{M} \psi_t^{(m)} \hat{y}_t^{(m)}\Big)^2 + \frac{1}{2} \sum_{m=1}^{M} \Big(\Theta^{(m)} - \Theta_0^{(m)}\Big)^T \Sigma_{\Theta_0^{(m)}}^{-1}\Big(\Theta^{(m)} - \Theta_0^{(m)}\Big) \tag{6.9}$$

The optimization problem of Equation 6.9 can be solved in a Bayesian Gauss-Newton framework (Mirikitani and Nikolaev, 2010). Approximating the cost function $\mathcal{J}_N\big(\{\Theta^{(m)}\}_{m=1}^{M}\big)$ by its second-order temporal Taylor expansion, we obtain

$$\begin{aligned}\mathcal{J}_N(\Theta) \approx\ & \mathcal{J}_N\big(\hat{\Theta}^{[k]}\big) \\ & + \nabla\mathcal{J}_N\big(\hat{\Theta}^{[k]}\big)\Big(\Theta - \hat{\Theta}^{[k]}\Big) \\ & + \frac{1}{2}\Big(\Theta - \hat{\Theta}^{[k]}\Big)^T \nabla^2\mathcal{J}_N(\hat{\Theta}^{[k]})\Big(\Theta - \hat{\Theta}^{[k]}\Big)\end{aligned} \tag{6.10}$$

where k denotes the iteration step and $\Theta = \Big[\ \theta_1^{(1)}, \cdots, \theta_P^{(1)}, \cdots, \theta_1^{(m)}, \cdots, \theta_P^{(m)}\ \Big]^T$ is introduced for notational convenience. The gradient of the approximated cost function is thus expressed as

$$\nabla\mathcal{J}_N(\Theta) = \nabla\mathcal{J}_N\big(\hat{\Theta}^{[k]}\big) + \nabla^2\mathcal{J}_N\big(\hat{\Theta}^{[k]}\big)\Big(\Theta - \hat{\Theta}^{[k]}\Big) \tag{6.11}$$

Putting the gradient equal to zero, the update equation for parameter estimation can be derived:

$$\hat{\Theta}^{[k+1]} = \hat{\Theta}^{[k]} - \Big(\nabla^2\mathcal{J}_N\big(\hat{\Theta}^{[k]}\big)\Big)^{-1}\nabla\mathcal{J}_N\big(\hat{\Theta}^{[k]}\big) \tag{6.12}$$

where,

$$\nabla \mathcal{J}_N(\Theta) = -\frac{1}{\sigma_\varepsilon^2}\sum_{t=1}^{N}\varphi_t(\Theta)\varepsilon_t(\Theta) + \sum_{m=1}^{M}\Sigma^{-1}_{\Theta_0^{(m)}}\left(\Theta^{(m)} - \Theta_0^{(m)}\right) \tag{6.13}$$

and

$$\nabla^2 \mathcal{J}_N(\Theta) = \frac{1}{\sigma_\varepsilon^2}\sum_{t=1}^{N}\varphi_t(\Theta)\varphi_t^T(\Theta) - \varphi_t^{'}(\Theta)\varepsilon_t(\Theta) + \sum_{m=1}^{M}\Sigma^{-1}_{\Theta_0^{(m)}} \tag{6.14}$$

with

$$\varphi_t(\Theta) = \left[\ \frac{\partial \hat{y}_t}{\partial \theta_1^{(1)}}, \cdots, \frac{\partial \hat{y}_t}{\partial \theta_P^{(1)}}, \cdots, \frac{\partial \hat{y}_t}{\partial \theta_1^{(m)}}, \cdots, \frac{\partial \hat{y}_t}{\partial \theta_P^{(m)}}\ \right]^T \tag{6.15}$$

The main advantage of the Bayesian parameter estimation is that the explicit prior knowledge on the parameters can be incorporated by specifying appropriate prior probabilities.

Based on the above theoretical derivations, the implementation procedure of the proposed Bayesian identification approach is outlined in Algorithm 6.1.

Algorithm 6.1. *Iterative Bayesian Procedure for Identification of Multi-model Inferential Sensors*

1. The following preparatory steps are completed to incorporate the relevant prior knowledge and include the available background information:

   1.1. Select a representative scheduling variable, s_t, that effectively determines the discrete-state dynamics at time instant t.

   1.2. Specify the number of sub-models, m, based on the prior knowledge of the nominal operating conditions or analysis of the historical operational data.

   1.3. Assign the prior probability of the system operating in the m^{th} mode, $p(i_t = m)$. Information about the typical operation schedule is obtained by interviewing the plant experts or analysis of the historical data.

   1.4. Determine the likelihood that s_t would be generated by the m^{th} mode, $p(s_t|i_t = m)$. Marginal and joint probability distributions of the scheduling variable and

the influential process variables could be investigated in order to specify such conditional probability density functions from historical data.

1.5. Characterize the prior distribution of model parameters, $p(\{\Theta^{(m)}\}_{m=1}^{M})$, based on the available background information.

2. Given the identification data-set $\mathcal{D} = \{(s_t, \mathbf{r}_t, y_t)\}_{t=1}^{N}$ evaluate the posterior probability of Equation 6.2 for all the identification data points to construct the interpolation matrix $\{\Psi^{(m)}\}_{m=1}^{M}$ with $\Psi^{(m)} = \{\psi_t^{(m)}\}_{t=1}^{N}$.

3. Choose a set of initial values for model parameters, $\hat{\Theta}^{[0]}$; the following steps will be repeated iteratively until no further improvement is gained:

3.1. Given the current estimate of the parameters $\hat{\Theta}^{[k]}$, calculate $\nabla \mathcal{J}_N(\hat{\Theta}^{[k]})$ and $\nabla^2 \mathcal{J}_N(\hat{\Theta}^{[k]})$ from Equations 6.13 and 6.14, respectively.

3.2. Update the MAP estimates of model parameters, $\hat{\Theta}^{[k+1]}$, using Equation 6.12.

6.4 Adaptation of Multi-model Inferential Sensors

The accuracy of an inferential sensor is usually guaranteed for only a particular region in which the model has been identified. However, most of the industrial processes exhibit a certain form of time-variant behavior due to fouling and/or abrasion in the process equipments, variation in the quality of feed, changes in the weather, and so on. In order to detect abrupt changes and gradual drifts in the process operations, process monitoring and on-line adaptation is often integrated in the implementation procedure. As a result of such precautions, the inferential model will be adjusted on-line to compensate for deviations from the off-line design conditions. Several on-line adaptation methods have been proposed in the literature on the basis of moving windows techniques, recursive adaptation techniques, and ensemble-based methods (Kadlec *et al.*, 2011; Khatibisepehr *et al.*, 2013).

The Bayesian decision-support scheme, presented in Section 6.3, inherently includes a global adaptation mechanism, within the envelope of previously identified operating conditions. The importance weights assigned to the sub-models are defined as time dependent to reflect the discrete-state dynamics at each time instant. The time-varying nature of the importance weights facilitate compensation of the expected slow and abrupt changes. Yet, there is a need to develop local adaptation mechanisms to detect and handle potential unknown drifts of process operating conditions which have not been captured in the model identification phase discussed previously. The local adaptation is accomplished through scaling and bias update; the latter is the common industrial practice. There are several approaches to dealing with unknown drifts caused by unidentified sources. The parameters of most of these adaptation techniques are often selected in a rather ad-hoc manner. This section extends the proposed Bayesian approach to provide a systematic procedure for estimation of adaptation parameters.

Consider an adaptive multi-model inferential sensor of the form

$$\begin{cases} y_t^{(m)} = \alpha_t^{(m)} f^{(m)}\big(\mathbf{r}_t; \Theta^{(m)}\big) + \beta_t^{(m)} \quad i = 1, \cdots, m \\ y_t = \sum_{m=1}^{M} \psi_t^{(m)}(s_t) y_t^{(m)} + \varepsilon_t \end{cases} \tag{6.16}$$

where $\alpha_t^{(m)}$ and $\beta_t^{(m)}$ respectively denote the scale factor and bias update term of the m^{th} sub-model at time instant t. There are a variety of update rules that can be specified to guide the adjustment of the scale factor and bias update term. In this work, the general form of an exponentially weighted moving average filter is employed to develop local adaptation mechanisms, such that

$$\alpha_{t+1}^{(m)} = \lambda^{(m)} \left(\frac{\psi_t^{(m)} y_t^{Ref} - \psi_{t-1}^{(m)} y_{t-1}^{Ref}}{\psi_t^{(m)} \Big[\alpha_t^{(m)} f^{(m)}(\mathbf{r}_t; \Theta^{(m)}) + \beta_t^{(m)} \Big] - \psi_{t-1}^{(m)} \Big[\alpha_t^{(m)} f^{(m)}(\mathbf{r}_{t-1}; \Theta^{(m)}) + \beta_t^{(m)} \Big]} \right) \alpha_t^{(m)}$$

$$+ (1 - \lambda^{(m)}) \alpha_t^{(m)} \tag{6.17}$$

$$\beta_{t+1}^{(m)} = \kappa^{(m)} \left(\psi_t^{(m)} \Big[y_t^{Ref} - \alpha_{t+1}^{(m)} f^{(m)}(\mathbf{r}_t; \Theta^{(m)}) - \beta_t^{(m)} \Big] + \beta_t^{(m)} \right)$$

$$+ \big(1 - \kappa^{(m)}\big) \beta_t^{(m)} \tag{6.18}$$

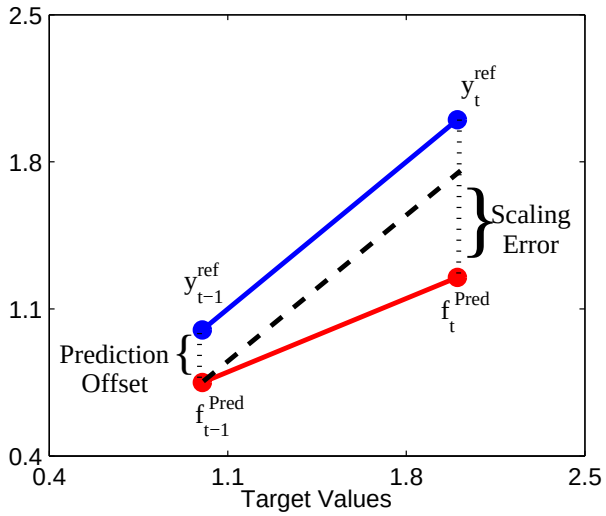


Figure 6.1: Inferential sensor calibration philosophy

where $\lambda^{(m)}$ and $\kappa^{(m)}$ are the smoothing parameters, also known as forgetting factors, that may be estimated off-line. y_t^{Ref} denotes the accurate reference value (*e.g.* laboratory measurement) at time instant t. Equations 6.17 and 6.18 can be best illustrated by Figure 6.1. As shown in this figure, the bias is updated to reduce the prediction offset; the bias error is corrected with reference to y_t^{Ref}, which is often sampled in a slow-rate. On the other hand, the scale factor is updated to adjust the slope of the imaginary line passing through predictions. The formulation adopted here allows for the straightforward implementation of on-line adaptation mechanisms for industrial inferential sensors, though other forms can also be considered.

The sub-model parameter vectors are augmented to include the local smoothing parameters, such that

$$\begin{aligned}\Theta &= \left[\ \Theta^{(1)}, \cdots, \Theta^{(m)}\ \right]^T \\ &= \left[\ \theta_1^{(1)}, \cdots, \theta_P^{(1)}, \lambda^{(1)}, \kappa^{(1)}, \cdots, \theta_1^{(m)}, \cdots, \theta_P^{(m)}, \lambda^{(m)}, \kappa^{(m)}\ \right]^T \end{aligned} \tag{6.19}$$

The augmented parameter vector can thus be used to directly extend the proposed Bayesian identification algorithm to estimate the local adaptation parameters and the sub-model

parameters simultaneously.

Once an adaptive multi-model inferential sensor has been identified, the Bayesian procedure presented in Algorithm 6.2 would be followed to obtain real-time predictions.

Algorithm 6.2. *Implementation Procedure of the Developed Adaptive Multi-model Inferential Sensors*

1. Based on the real-time measurement of the scheduling variable, evaluate the posterior probability of Equation 6.2 to construct the interpolation vector $\{\psi_t^{(m)}\}_{m=1}^{M}$.

2. Calculate the output of each sub-model form the identified local models.

3. Combine the calculations of Steps 1 and 2 in order to obtain a real-time prediction of the query variable (see Equation 6.16).

4. Upon the arrival of a laboratory measurement (or any other reliable off-line data), use Equations 6.17 and/or 6.18 to respectively update the scale factor and the bias term for the next interval until a new reference value is available.

6.5 CSTR Simulation Example

Continuous stirred tank reactors (CSTR) are commonly used in the process industries. An irreversible and exothermic reaction takes place inside the tank of a single perfectly mixed CSTR shown in Figure 6.2. The coolant water is continuously circulated through a cooling jacket surrounding the reactor to absorb the generated reaction heat. The governing equations of a CSTR are given by (Xu *et al.*, 2009)

$$\frac{dC_a(t)}{dt} = \frac{F_i}{V}\big(C_{ai} - C_a(t)\big) - k_0 C_a(t) \exp\left(\frac{-E}{RT}\right) \tag{6.20}$$

$$\begin{aligned}\frac{dT(t)}{dt} = {} & \frac{F_i}{V}\big(T_i - T(t)\big) - \frac{\Delta H k_0 C_a(t)}{\rho C_p} \exp\left(\frac{-E}{RT}\right) \\ & + \frac{\rho_c C_{pc}}{\rho C_p V} F_c(t) \left(1 - \exp\left(\frac{-hA}{F_c(t)\rho C_p}\right)\right)\big(T_{ci} - T(t)\big)\end{aligned} \tag{6.21}$$

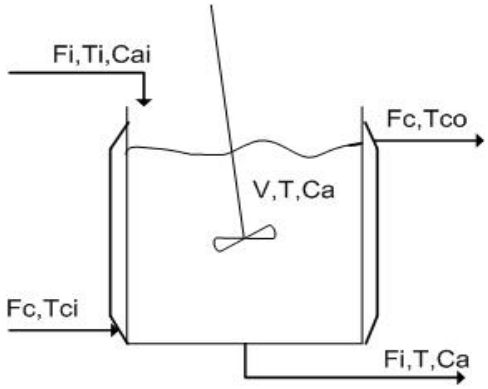


Figure 6.2: Schematic of a continuous stirred tank reactor

where C_a is the product concentration, T is the product temperature, and F_c is the coolant flow-rate. The CSTR model parameters and steady state operating conditions are listed in Tables 6.1 and 6.2, respectively (Xu *et al.*, 2009).

In this study, the product concentration (C_a) is monitored as the output variable and the coolant flow-rate (F_c) is selected as the manipulated variable, *i.e.* $y_t = C_a(t)$ and $r_t = F_c(t)$. As coolant flow-rate increases, the product temperature decreases and, consequently, the product concentration increases. The process is simulated using the non-linear dynamic model given in Equations 6.20 and 6.21. As listed in Table 6.2, five operating conditions are considered throughout the feasible range of the coolant flow-rate that is limited to the range of $95\ L/min$ to $111\ L/min$. To illustrate the non-linearity of the CSTR process over this operating range, the step responses from the coolant flow-rate to the product concentration are displayed in Figure 6.3.

The noise contaminating F_c and C_a are assumed to be zero mean Gaussian random variables with variance of 2.25×10^{-2} and 10^{-8}, respectively. Although not a common industrial practice, a random sequence of the operating conditions has been used for both identification and validation purposes. The random sequences were selected to illustrate that the application of the proposed identification and implementation procedures does not require any specific transition patterns. The validation values of the coolant flow-rate

Table 6.1: A summary of the CSTR model parameters

Parameter	Symbol	Value
Feed flow-rate	F_i	$100\ L/min$
Feed concentration	C_{ai}	$1\ mol/L$
Feed temperature	T_i	$350\ K$
Reactor volume	V	$100\ L$
Reactor rate constant	k_0	$7.2 \times 10^{10}\ min^{-1}$
Activation energy term	E/R	$1 \times 10^4\ K$
Heat of reaction	ΔH	$-2 \times 10^5\ cal/mol$
Reactant density	ρ	$1000\ g/L$
Reactant specific heat	C_p	$1\ cal/g/K$
Heat transfer term	hA	$7 \times 10^5\ cal/min/K$
Coolant inlet temperature	T_{ci}	$350\ K$
Coolant density	ρ_c	$1000\ g/L$
Coolant specific heat	C_{pc}	$1\ cal/g/K$

Table 6.2: CSTR steady state operating conditions

	Coolant Flow-rate L/min	Product Concentration mol/L	Product Temperature K
Mode 1	97	0.0795	443.4566
Mode 2	100	0.0885	441.1475
Mode 3	103	0.0989	438.7763
Mode 4	106	0.1110	436.3091
Mode 5	109	0.1254	433.6921

are different from the ones used in the identification phase. To switch between different operating points, the coolant flow-rate is changed by a fixed step size. The sampling interval is assumed to be $T_s = 0.1\ min$.

The non-linear dynamic behavior of the CSTR is modeled by a set of piecewise linear models interconnected through a Bayesian decision-support interpolation function. The coolant flow-rate is selected as the scheduling variable, *i.e.* $s_t = F_c(t)$, because it effectively determines the discrete-state process dynamics. As presented in Figure 6.4, the probability distribution of the coolant flow-rate can be approximated as a mixture of five

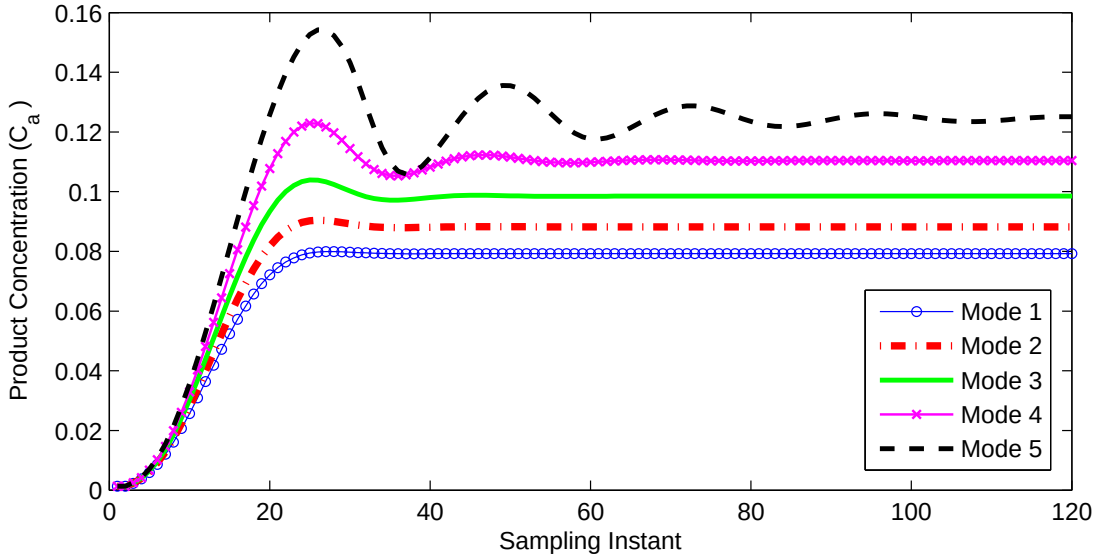


Figure 6.3: Step responses from the coolant flow-rate to the product concentration

Gaussian distributions. The m^{th} distribution reflects the likelihood that $F_c(t)$ was generated by the m^{th} operating condition. Also, the prior probability that the system operates around the m^{th} mode is assumed to be equal among all modes. Figure 6.5 illustrates the importance weights assigned to the sub-models within the feasible range of the scheduling variable.

Algorithm 6.1 is followed to identify sub-model parameters and Algorithm 6.2 is implemented to predict the product concentration given the coolant flow-rate. The predicted and reference values are compared in Figure 6.6.a and the importance weights are presented in Figure 6.6.b. A close agreement between the predictions and reference values are obtained on the identification data. The prediction performance of the identified hybrid model is further tested on the validation data. The predicted and reference values are compared in Figure 6.7.a and the importance weights are presented in Figure 6.7.b. It is observed that the identified system is capable of producing accurate predictions and tracking the significant changes in the reference data.

As discussed previously, the proposed Bayesian method accommodates the overlapping operating modes, which would increase the robustness of an inferential sensor. This

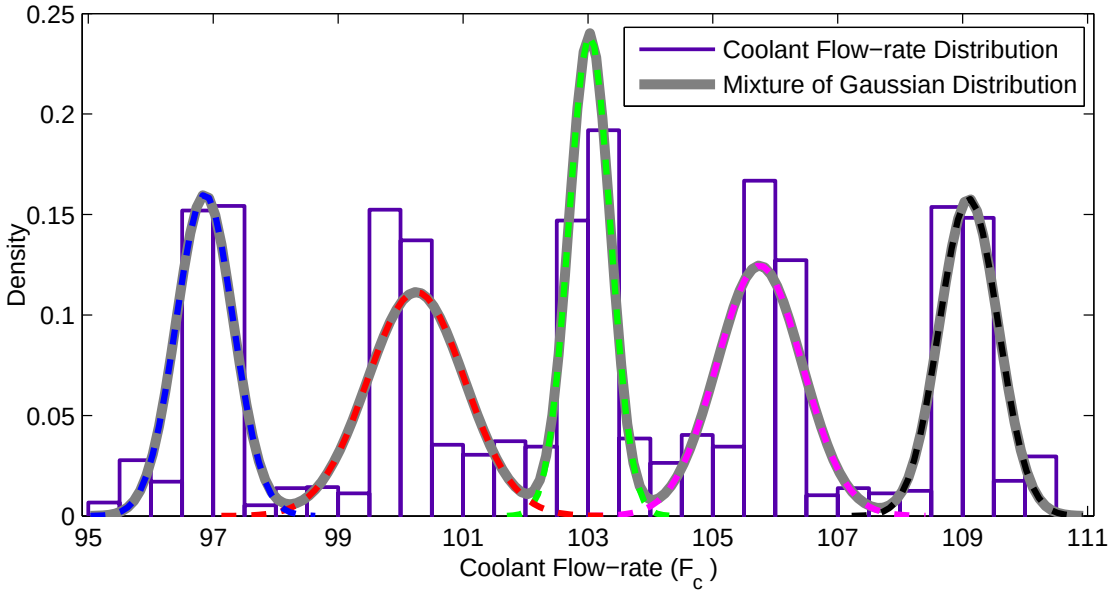


Figure 6.4: Probability distribution of coolant flow-rate

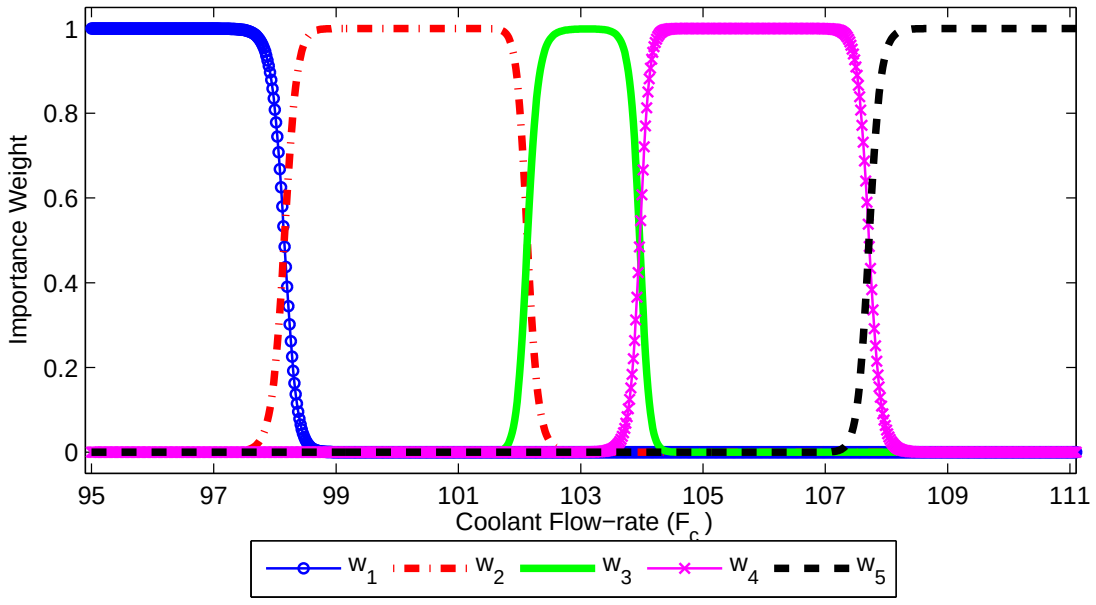
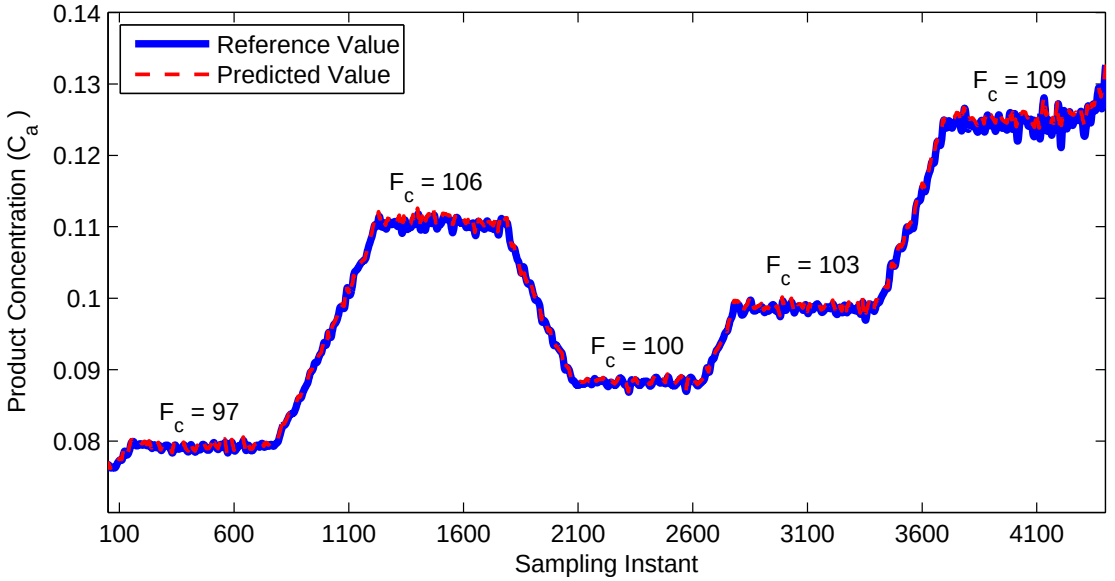
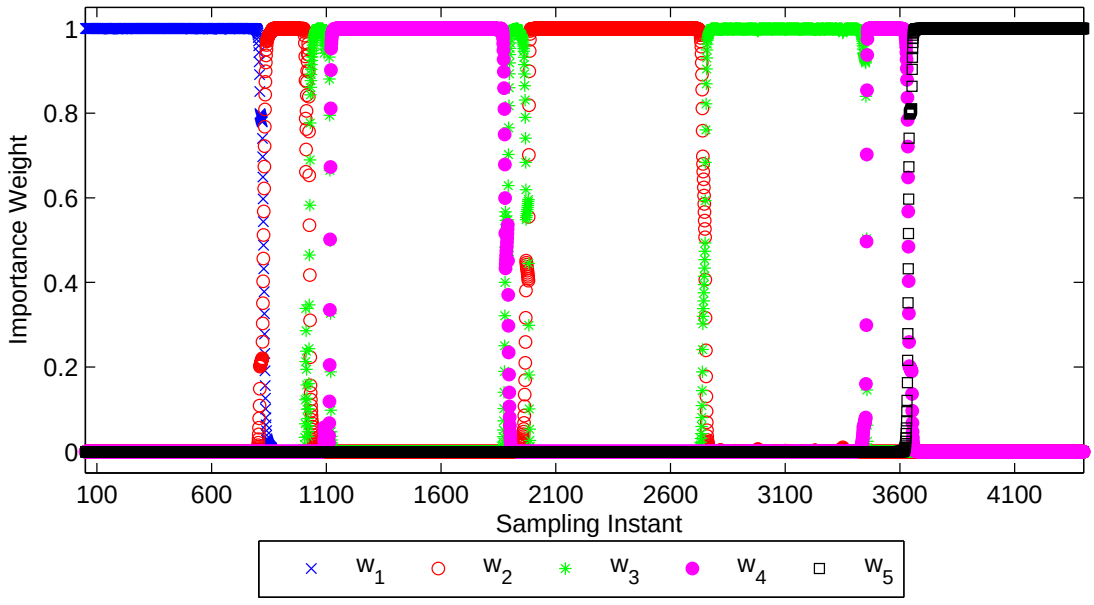


Figure 6.5: Importance weights assigned to the sub-models

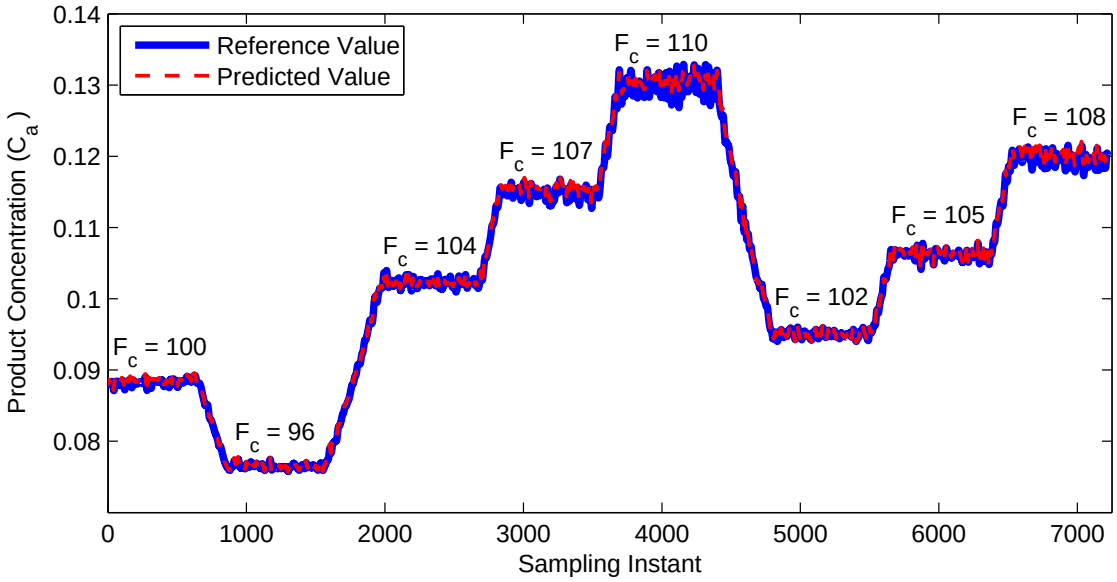


(a) Time trend comparison

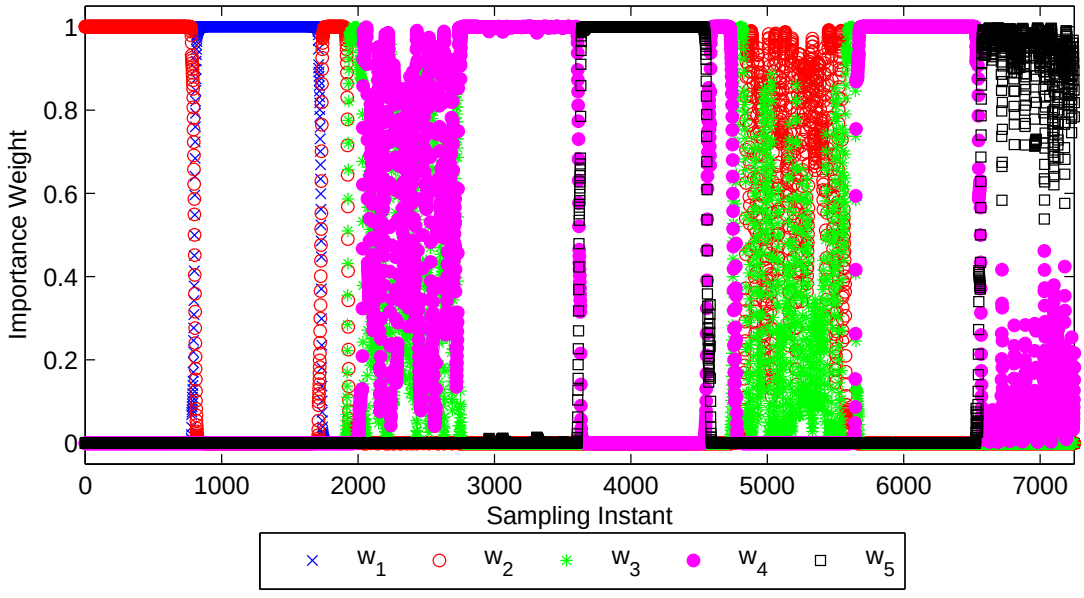


(b) Importance weights assigned to the sub-models

Figure 6.6: CSTR: Self-validation



(a) Time trend comparison



(b) Importance weights assigned to the sub-models

Figure 6.7: CSTR: Cross-validation

unique feature of our method is highlighted through demonstration of the smooth transition between operating modes. It is contrary to the abrupt changes observed in the predictions obtained from a conventional Bayesian method at transition periods, as demonstrated in Figure 6.8.

To illustrate the importance of local adaptation mechanisms, suppose that the coolant inlet temperature (T_{ci}) fluctuates randomly between $349.5\ K$ and $350\ K$ with switching probability of 0.01. Furthermore, it is assumed that F_c is contaminated by zero mean Gaussian noise with variance of 2.5×10^{-3}. Note that the coolant flow-rate measurements are considered to be available, while the random fluctuations in the coolant inlet temperature are treated as unmeasured.

The raw predictions, without bias compensation, and the measured values of product

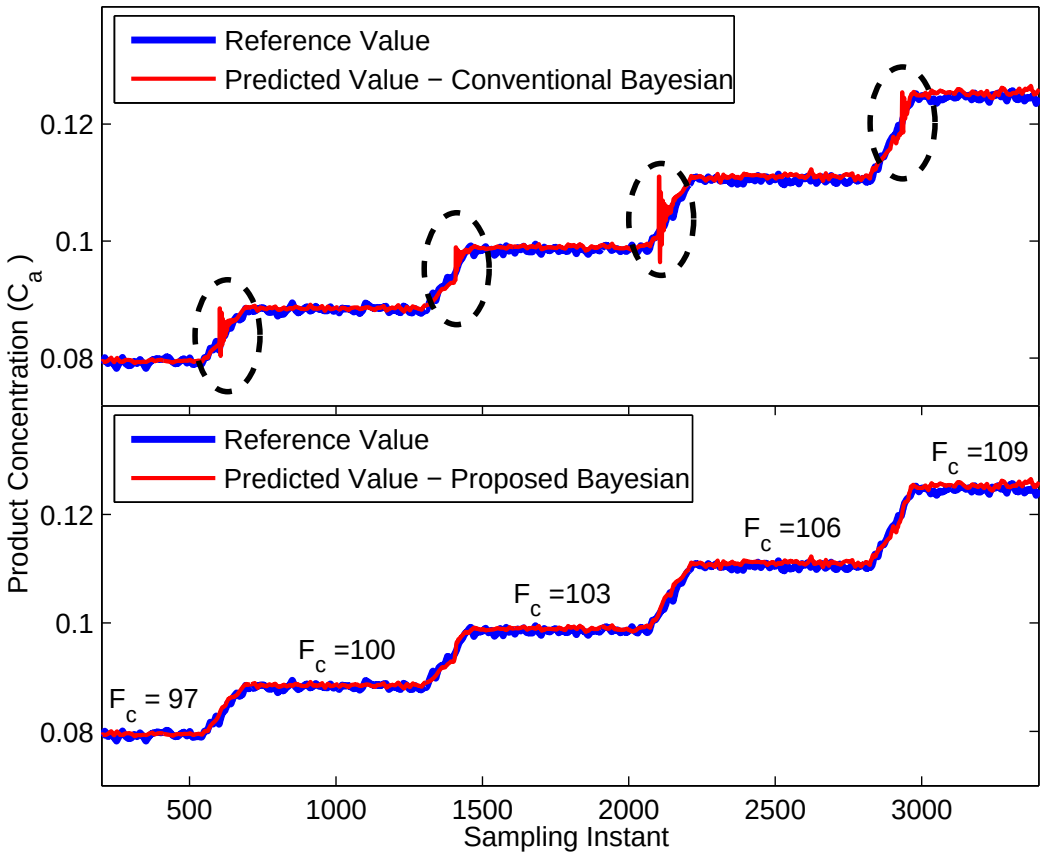


Figure 6.8: Comparison between the conventional and proposed Bayesian methods

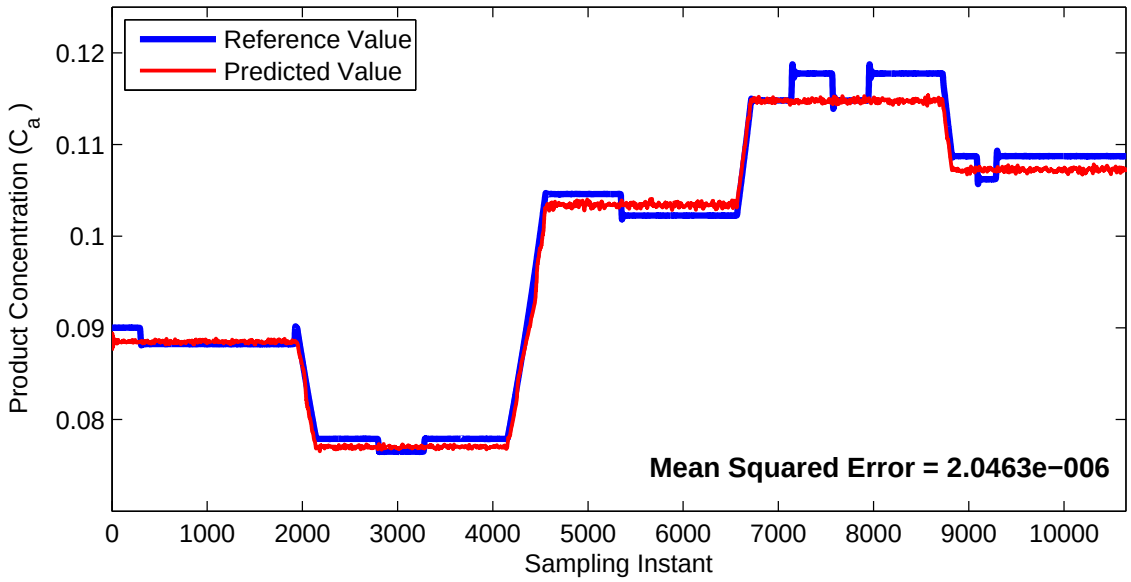
concentration are compared in Figure 6.9.a. Evidently, there is an offset between the predicted and reference values. In order to minimize the prediction errors, the bias is updated every 200 sampling instants. As shown in Figure 6.9.b, the bias compensation has certainly improved the prediction performance of the inferential sensor.

6.6 Industrial Case Study

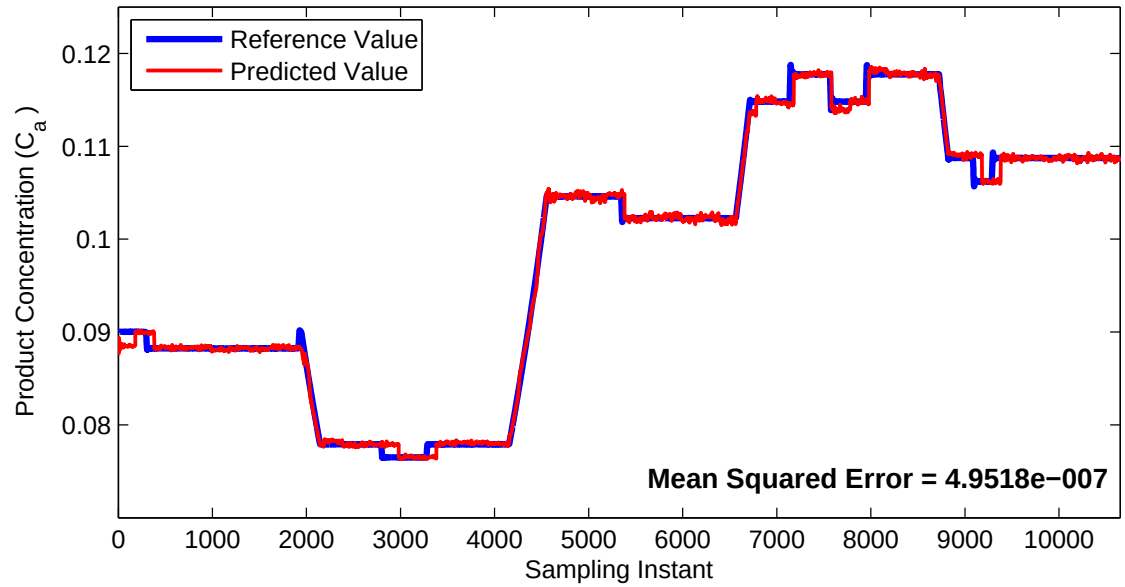
6.6.1 Process Description

The main objective of the oil sand extraction process is to separate bitumen from other components, which are mainly water and solids. The oil sand is first mixed with hot water and the resulting slurry is then fed into a primary separation vessel (PSV) to facilitate bitumen flotation and sand settling. The froth floats off the top of PSV and the deaerated froth is further treated in the froth treatment plant to remove residual water and fine solids. The froth is first mixed with diluent and some process aids such as demulsifier. The diluent is mixed with froth to produce lighter hydrocarbon phase and, consequently, to enhance the density difference between the various components, while the demulsifier is added to break water-oil emulsions. The diluted froth is fed into various separation units, most of which rely on gravity separation principles. The inclined plate settler (IPS) units are one of the key froth treatment processes. The IPS units allow for the space efficient gravity separation of diluted bitumen from the other components. The IPS overflow product stream mainly consists of the diluted bitumen floating to the top of the vessel. The other components of the diluted froth such as water and minerals settle down at the bottom of the vessel to be treated by the centrifuges (Domlan *et al.*, 2010; Shao *et al.*, 2011). A schematic diagram of an IPS unit is illustrated in Figure 6.10.

The Diluent to Bitumen (D:B) ratio in the IPS product stream is used to control the quality of diluted bitumen and, thus, serves as one of the key indicators of the separation process performance. It is very important to maintain the D:B ratio in both product and



(a) Comparison between the raw predictions (without bias compensation) and reference values



(b) Comparison between the updated predictions (with bias compensation) and reference values

Figure 6.9: CSTR: Time trend comparison between the predicted and actual values of product concentration contaminated with colored noise

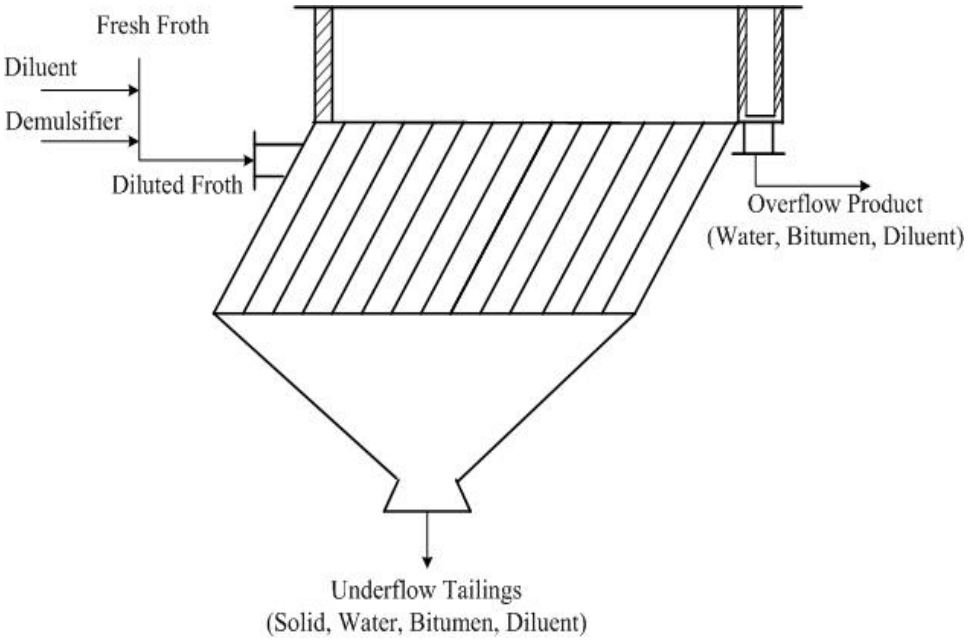


Figure 6.10: Schematic diagram of the inclined plates settler (IPS) operation

feed streams at certain levels so as to achieve optimal separation efficiency with effective cost. However, D:B measurements are normally determined by on-line analyzers or off-line analysis in laboratory, which are often not reliable. Furthermore, significant delay is incurred in laboratory testing such that the measured signal cannot be used as the feedback signal for control systems. Therefore, there is an economic necessity to develop inferential mechanisms in order to improve the accuracy and reliability of real-time D:B ratio estimates and, consequently, improve product quality. Specifically, the objective of this study is to design an inferential sensor for real-time monitoring of the D:B ratio in the IPS product stream, *i.e* $y_t = DB_t$. Two parallel IPS vessels are considered, namely **IPS A** and **IPS B**.

6.6.2 Process Data Analysis

A list of the influential process variables is presented in Table 6.3. These variables have been identified by exploiting the analytical knowledge as well as considering the availability of measuring devices. The real-time measurements are recorded every minute,

whereas the laboratory analysis of D:B is logged approximately every 2 hours. The operational and laboratory data was collected through an automated data historian. Since a large amount of historical data was available, two independent sub-sets were constructed. An identification data-set was constructed from the data recorded from September 1 to December 31, 2010; altogether 156140 fast-rate and 1286 slow-rate measurements are used. Also, a validation data-set was constructed from the data recorded from January 1 to April 30, 2011; altogether 168030 fast-rate and 1352 slow-rate measurements were used. The identification data-set was used for inferential model identification purposes, while the validation data-set was reserved for cross validating the performance of the designed inferential sensor. All industrial data presented here has been normalized in order to protect proprietary information.

The careful investigation of the collected historical data reveals that missing measurements and outlying observations are the main factors affecting the data quality. Yet, these factors did not show a considerable impact on the completeness and reliability of the selected identification data-set. Various pre-processing techniques were adopted to refine the quality of operational data.

Through Bayesian data analysis, the marginal and joint probability distributions of the influential process variables were investigated in order to extract hidden patterns (*e.g.* dependencies, operating ranges, and etc.) from historical data. Having incorporated the knowledge and experience of the plant experts, these patterns would be considered as a summary of the input data. As illustrated in Figures 6.11.a and 6.11.b, the probability distributions of historical diluent flow-rate and IPS feed flow-rate can be approximated as a mixture of three Gaussian distributions. This motivates the application of multi-model inferential sensors to cover different operating conditions, while preserving the accuracy of predictions in the normal operating region. The real-time feed flow-rate is selected as the scheduling variable, *i.e.* $s_t = F_{df,t}$. Consequently, the operating space of the IPS is

Table 6.3: A summary of the influential process variables

Process Variable	Symbol
Demulsifier flow-rate	F_{de}
Diluent flow-rate	F_{di}
IPS diluted feed flow-rate	F_{df}
IPS product flow-rate	F_{p}
IPS underflow flow-rate	F_{u}

partitioned into multiple operating modes, namely **low feed flow-rate**, **medium feed flow-rate**, and **high feed flow-rate**. It is noteworthy that the choice of the number of sub-models is a trade-off between accuracy of the distribution fitting and complexity of the inferential sensor.

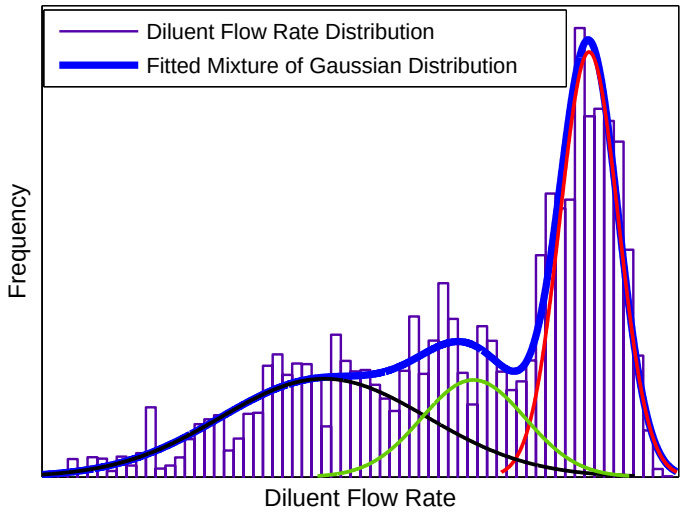
6.6.3 Model Identification

on-line estimation of D:B ratio on the basis of first principles analysis requires real-time density and composition measurements for different streams. Due to the lack of such measurements, it is not possible to develop a complete knowledge-driven model of the process. Instead, the available process knowledge is considered to search for an appropriate model structure, while historical operational data are used to reveal the parametric relationship between D:B ratio and on-line measurable process variables. Based on the insight obtained from first principles and process data analysis, the following model is found to provide reasonable real-time predictions of DB_t from $\mathbf{r}_t = [F_{de,t}, F_{di,t}, F_{df,t}, F_{p,t}, F_{u,t}]$:

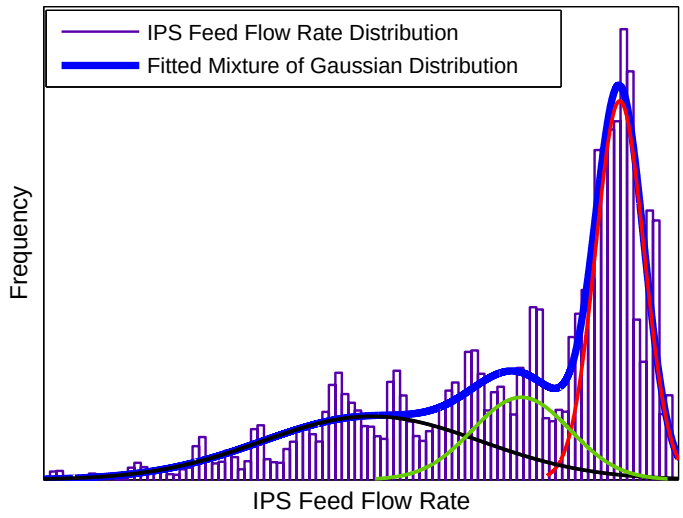
$$\begin{aligned}\widehat{DB}_t^{Fast} &= \sum_i \psi_t^{(m)} \widehat{DB}_t^{(m)} \\ &= \sum_i p(i_t = m|F_{df,t}) \widehat{DB}_t^{(m)} \qquad i \in \{Lo, Med, Hi\}\end{aligned} \tag{6.22}$$

with

$$p(i_t = m|F_{df,t}) = \frac{p(F_{df,t}|i_t = m)p(i_t = m)}{\Sigma_i p(F_{df,t}|i_t = m)p(i_t = m)} \tag{6.23}$$



(a) Probability distribution of diluent flow-rate



(b) Probability distribution of IPS feed flow-rate

Figure 6.11: Historical probability distributions of influential process variables

and

$$\widehat{DB}_t^{(m)} = \frac{1 - z^{-1}}{1 - \left(1 - \psi_t^{(m)}\kappa^{(m)}\right)z^{-1}} \times \left(\frac{F_{di,t}(\theta_1^{(m)}F_{df,t} + \theta_2^{(m)}F_{de,t} + \theta_3^{(m)}F_{di,t})^{-1} + \theta_4^{(m)}F_{u,t}F_{df,t}^{-1} + \theta_5^{(m)}}{1 + \theta_6^{(m)}F_{u,t}F_{df,t}^{-1}} \right) + \frac{\psi_t^{(m)}\kappa^{(m)}z^{-1}}{1 - \left(1 - \psi_t^{(m)}\kappa^{(m)}\right)z^{-1}} DB_T^{Lab} \tag{6.24}$$

where z^{-1} is the back-shift operator, T corresponds to the slower sampling rate (*e.g.* every 1 hour), and t corresponds to the faster sampling rate (*e.g.* every 1 minute). Therefore, $\widehat{DB}_t^{(m)}$ represents the real-time fast-rate estimate obtained from the m^{th} sub-model, while $z^{-1}DB_T^{Lab} = DB_{T-1}^{Lab}$ denotes the off-line slow-rate laboratory measurement with $T - 1 < t \leq T$. For the m^{th} sub-model, an adaptive bias term with smoothing parameter $\kappa^{(m)}$ is included to keep track of drifts in the process data through the adjustment of on-line predictions to the newly arrived laboratory data. By manipulating such adaptive structure, the second term on the right-hand side of Equation 6.24 would appear as a function of the lagged slow-rate D:B measurements (laboratory data). Moreover, $p(i_t = m|F_{df,t})$ is the likelihood that $F_{df,t}$ was generated by the m^{th} mode and $p(i_t = m)$ is the prior probability that the system operates in the m^{th} mode. The prior probabilities and the likelihood functions are determined on the basis of the available operational information as well as the historical process data. Finally, the sub-model parameters are estimated according to the procedure outlined in Algorithm 6.1. It is noteworthy that the developed inferential sensor is only parameter-varying with respect to the scheduling variable, *i.e.* the functional form of the sub-models obtained from mass balance equations remains the same. Figure 6.12 illustrates the relative variation of the sub-model parameters, *i.e.* $\Theta^{(m)} = \{\theta_1^{(m)}, \cdots, \theta_6^{(m)}, \kappa^{(m)}\}$, with respect to the operating modes.

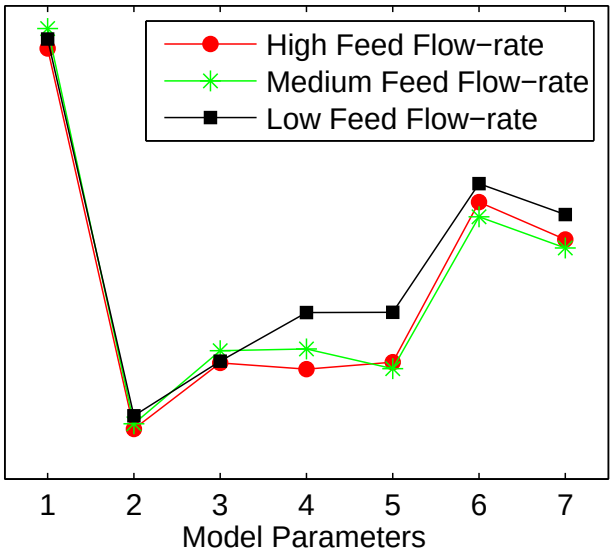


Figure 6.12: Sub-model parameter estimates

6.6.4 Model Evaluation

6.6.4.1 Performance Evaluation Criteria

Generally, the major purpose of model validation is to evaluate the accuracy and reliability of the developed inferential sensor. Accuracy is the level of agreement between the predicted and target values, while reliability is the degree to which the prediction errors vary. Evaluating the performance of an inferential sensor thus amounts to analyzing the characteristics of prediction errors, which are also referred to as residuals. The graphical techniques used in analysis of residuals are listed below:

- **Scatter plot of predicted values versus target values**: The ideal case would be for all the data points to lie on the identity line ($y = x$), indicating perfect agreement between the predicted and target values.

- **Run-sequence plot of predicted and target values**: The time trend of the predicted and target values are plotted together to visually assess the accuracy and reliability of the inferential model.

To provide an arithmetical basis for evaluating the performance of the developed inferential sensor, mean absolute error (MAE), standard deviation of errors (StdE), and mean squared error (MSE) are assessed.

The MAE is a measure of accuracy defined as the average of absolute prediction errors:

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}|\varepsilon_n| \tag{6.25}$$

where N is the number of observations and ε_n is the prediction error for the n^{th} observation.

The StdE is a measure of reliability expressed through the variation of prediction errors:

$$\text{StdE} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\varepsilon_n - \bar{\varepsilon})^2} \tag{6.26}$$

where $\bar{\varepsilon}$ is the mean of error distribution.

Finally, the MSE indicates the overall prediction performance in terms of both accuracy and reliability:

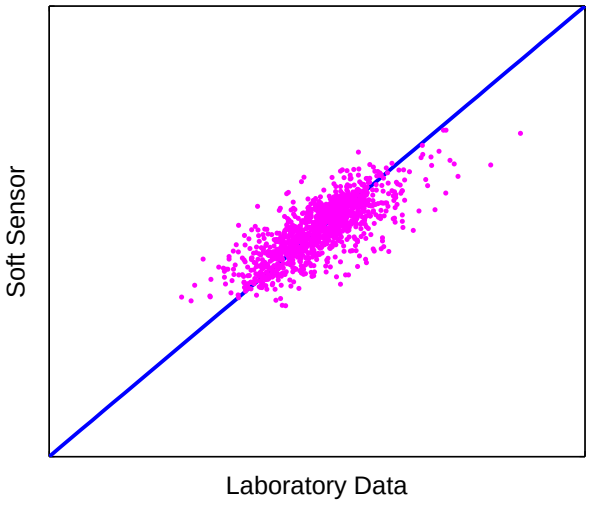
$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\varepsilon_n^2} \tag{6.27}$$

Readers are referred to Section 2.4.1 for more details.

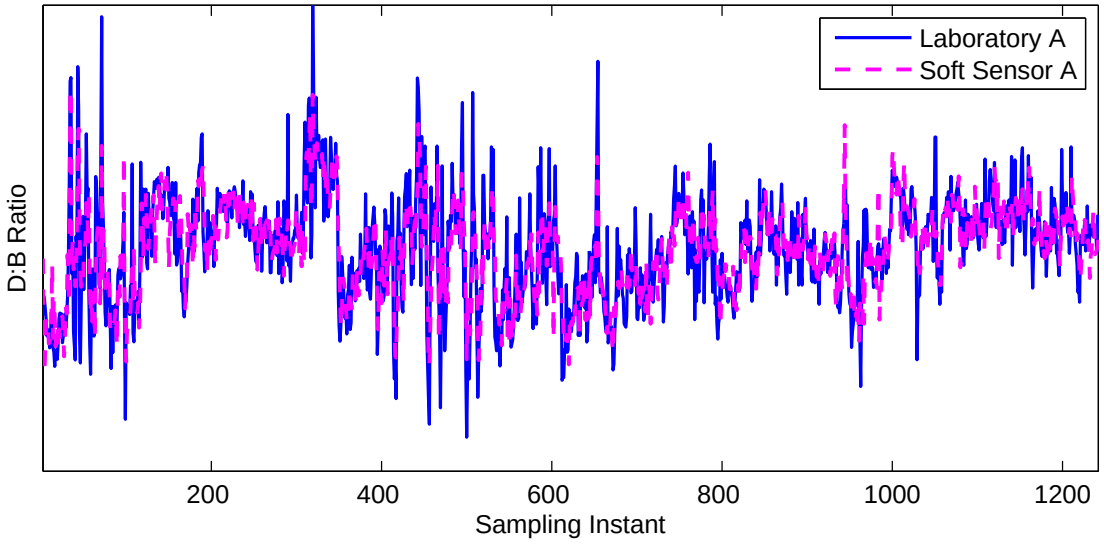
6.6.4.2 Off-line Performance Evaluation

The off-line performance evaluation comprises two steps, namely, **self-validation** and **cross-validation**. Self-validation determines the adequacy of fit by evaluating the prediction performance of the developed inferential model on the identification data. Cross-validation assesses the generalization capability by evaluating the prediction performance of the identified inferential model on the validation data that has not been involved in the identification procedure.

The performance of the developed inferential sensor is first verified on the identification data-set. Figures 6.13.a and 6.14.a show the scatter plots of the D:B predictions versus laboratory measurements for IPS A and IPS B, respectively. Also, Figures 6.13.b and

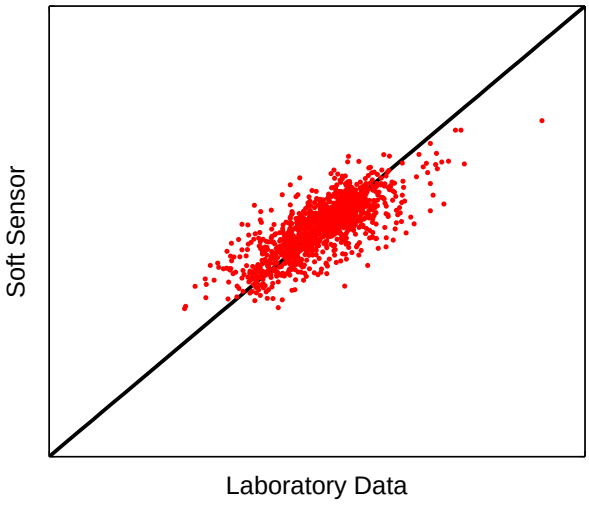


(a) Scatter plot comparison

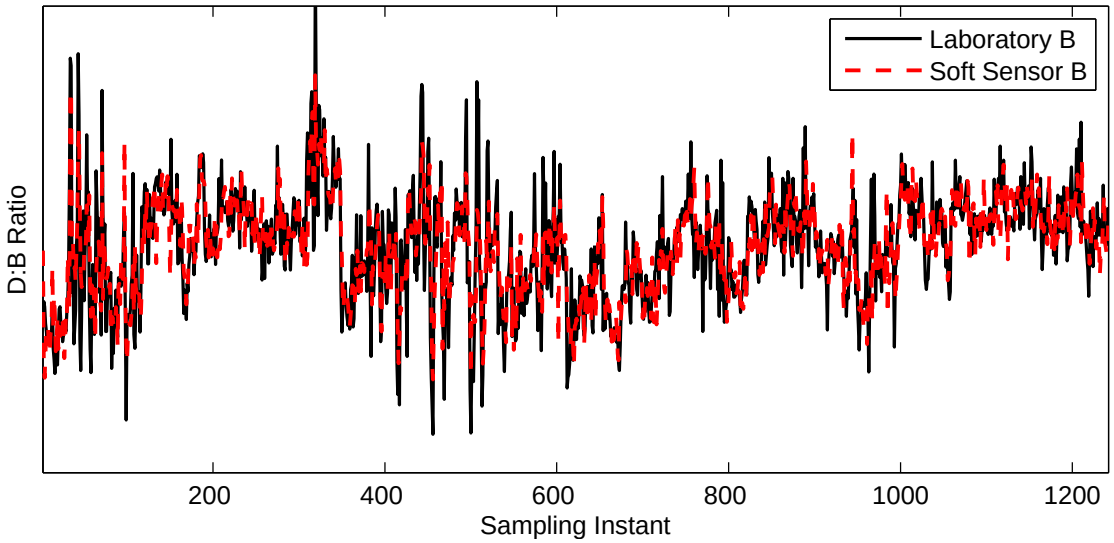


(b) Time trend comparison

Figure 6.13: IPS A: Self-validation



(a) Scatter plot comparison



(b) Time trend comparison

Figure 6.14: IPS B: Self-validation

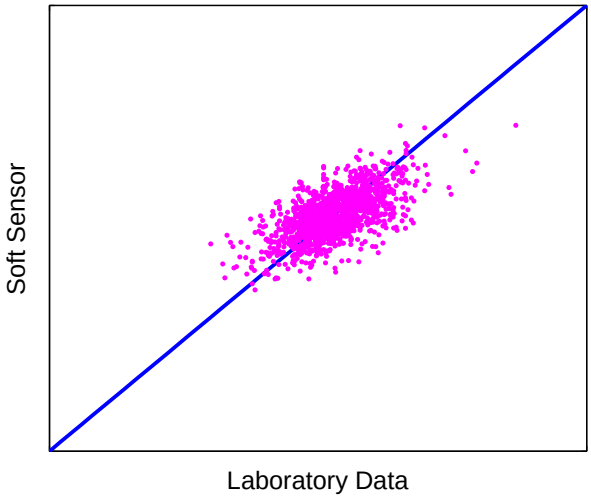
6.14.b display the run-sequence plots of predicted and target values for IPS A and IPS B, respectively. It is observed that the D:B predictions for both IPS A and IPS B are able to accurately fit the laboratory data. However, the adequacy of model fit does not reliably ascertain the prediction performance *i.e.* satisfactory prediction capability on the identification data does not guarantee generalization to other data-sets. Thus, the prediction performance of the developed inferential sensor is next evaluated on the validation data-set. Figures 6.15.a and 6.16.a show the scatter plots of the D:B predictions versus laboratory measurements for IPS A and IPS B, respectively. Also, Figures 6.15.b and 6.16.b display the run-sequence plots of predicted and target values for IPS A and IPS B, respectively. It is observed that the developed inferential sensor is capable of producing fairly accurate D:B predictions and tracking the significant changes in laboratory data.

6.6.4.3 On-line Performance

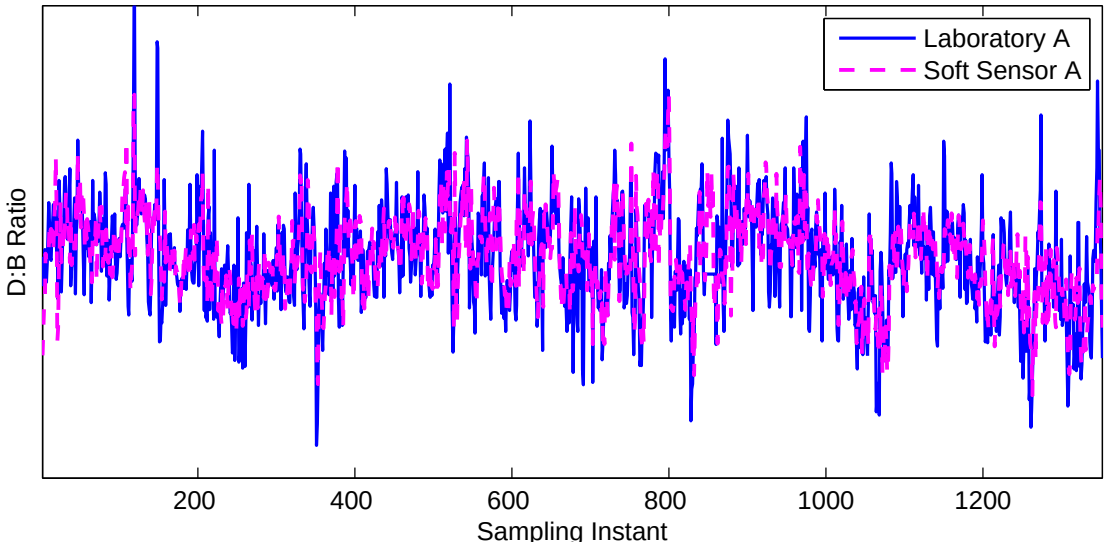
Since the off-line performance of the developed inferential sensor was satisfactory, the sensor was further tested on-line in the IPS unit of an oil sands processing plant. The model parameters were estimated off-line before the on-line implementation. In order to perform an on-line implementation of the soft sensor, an object linking and embedding for process control (OPC) in MATLAB has been used as the communication channel between the inferential sensor and a tag created on the distributed control system (DCS). All the necessary computations for the inferential sensor are performed in MATLAB and the predicted values as well as performance indices are sent back to the DCS through the OPC connection.

Implementation of inferential sensors entail many challenges that may arise due to the varying quality of industrial data. In order to enhance the robustness of the designed inferential sensors, various univariate and multivariate pre-processing procedures were developed to assess the availability and reliability of the input measurements.

The developed inferential sensor has been running on-line reliably and successfully

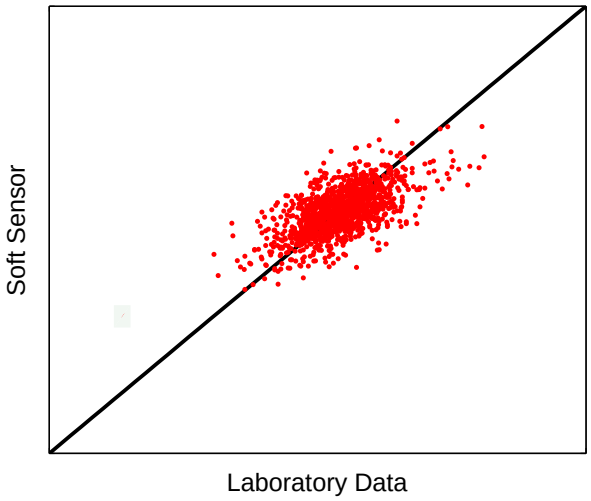


(a) Scatter plot comparison

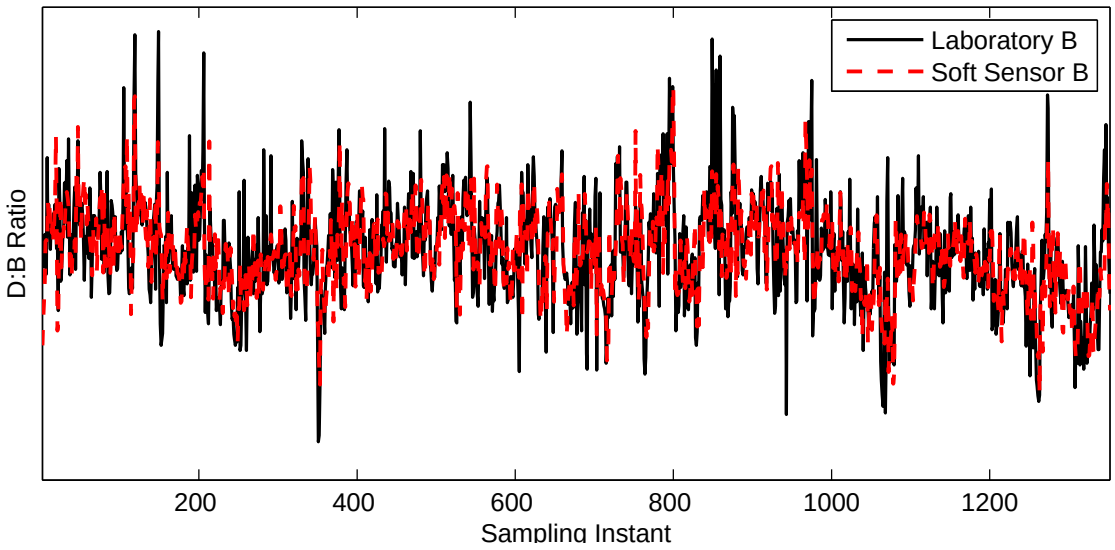


(b) Time trend comparison

Figure 6.15: IPS A: Cross-validation

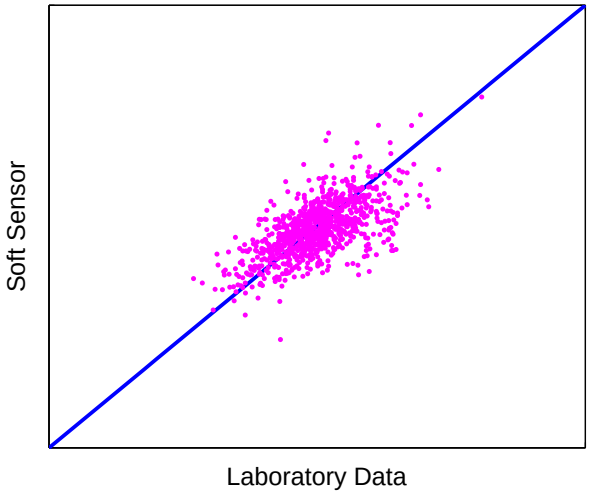


(a) Scatter plot comparison

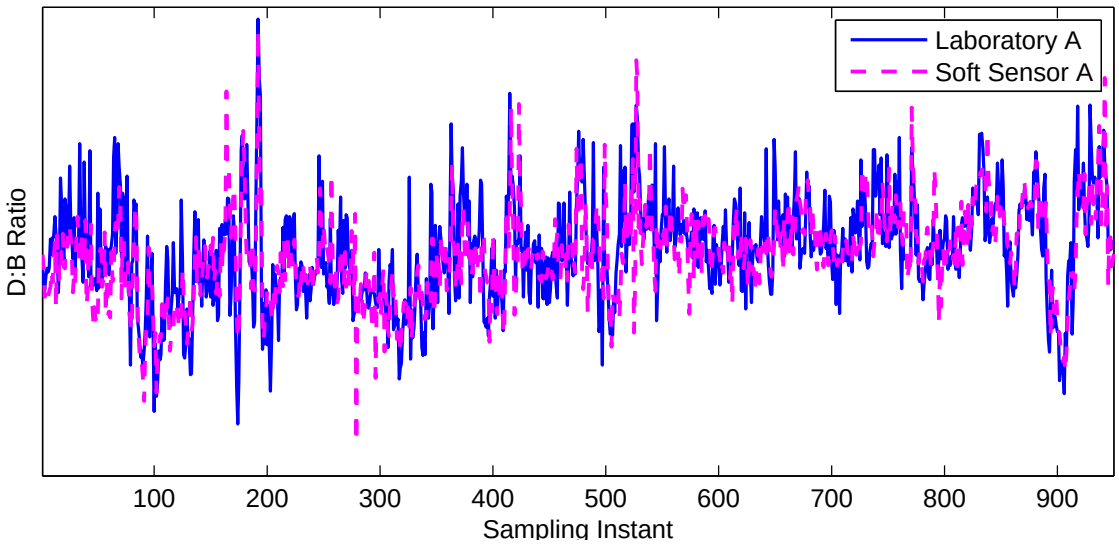


(b) Time trend comparison

Figure 6.16: IPS B: Cross-validation

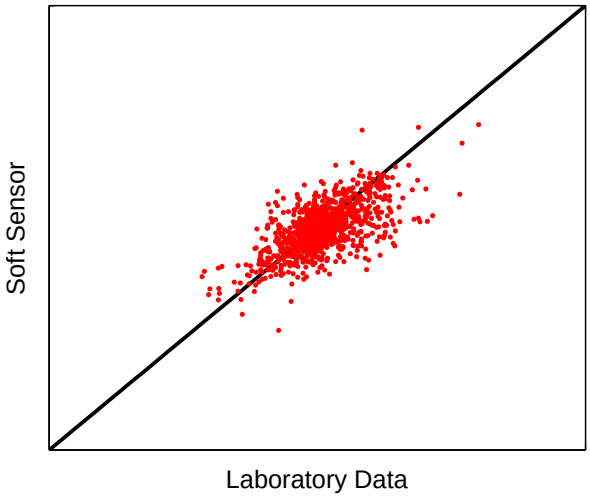


(a) Scatter plot comparison

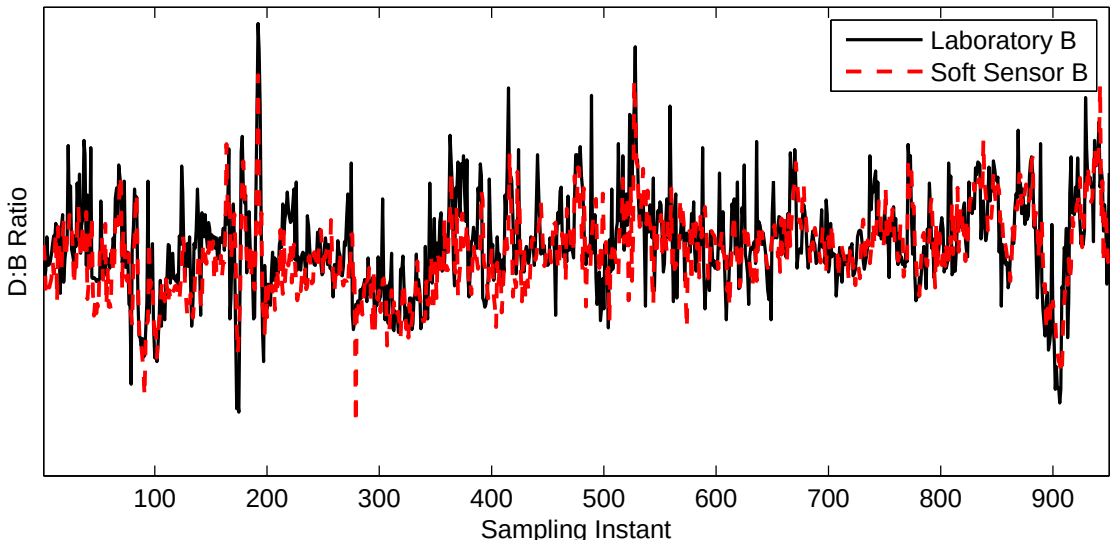


(b) Time trend comparison

Figure 6.17: IPS A: On-line testing



(a) Scatter plot comparison



(b) Time trend comparison

Figure 6.18: IPS A: On-line testing

Table 6.4: A summary of the performance measures

	MAE	STD	MSE
Self-validation			
IPS A	0.0180	0.0242	5.8613e-004
IPS B	0.0181	0.0243	5.9165e-004
Cross-validation			
IPS A	0.0195	0.0254	6.4431e-004
IPS B	0.0195	0.0255	6.4892e-004
On-line Testing			
IPS A	0.0205	0.0276	7.7303e-004
IPS B	0.0186	0.0249	6.5491e-004

since July 20, 2011. Figures 6.17.a and 6.18.a show a snapshot of the scatter plots of the D:B predictions versus laboratory measurements for IPS A and IPS B, respectively. Also, Figures 6.17.b and 6.18.b display the run-sequence plots of predicted and target values for IPS A and IPS B, respectively. It is evident that the developed inferential sensor provides fairly accurate D:B predictions and tracks the significant changes in laboratory data.

Finally, MAE, StdE, and MSE for the real-time D:B predictions are reported in Table 6.4; the results are obtained through the comparison of the soft sensor predictions and the laboratory measurements. The designed inferential sensor is considered by the engineers to be both accurate (with small MAE) and reliable (with smaller StdE). Moreover, relatively small values of MSE implies the overall good performance.

6.7 Conclusion

In this Chapter, a Bayesian framework for the development and implementation of adaptive multi-model inferential sensors was proposed. The presented Bayesian procedure for model identification allows for accommodating the overlapping operating modes and facilitating the inclusion of the prior knowledge about the process operation. Also, the presented implementation procedure inherently includes a global Bayesian adaptation

mechanism, within the envelope of previously identified operating conditions. Moreover, local adaptation mechanisms were included in order to detect and handle potential unknown drifts. The effectiveness of the identification and implementation procedures were demonstrated through simulation and industrial applications. In the simulation case study, a multi-model inferential sensor was developed to capture the non-linear dynamic behavior of the CSTR by concatenating multiple linear models through a Bayesian decision-support system. In the industrial application, two inferential sensors were successfully designed for predicting the quality of the product of a separation unit in oil sands processing. The variable of interest was the Diluent to Bitumen (D:B) ratio in the IPS product streams. The off-line validation and on-line implementation results showcase the prediction performance of the designed adaptive multi-model inferential sensors.

Bibliography

Angelov, P. and A. Kordon (2010). Adaptive inferential sensors based on evolving fuzzy models. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **40**(2), 529–539.

Azimzadeh, F., H. A. Palizban and J. A. Romagnoli (1998). Online optimal control of a batch fermentation process using multiple model approach. In: *Proceedings of the* 37^{th} *IEEE Conference on Decision and Control*. number 1. IEEE. Tampa, USA. pp. 455–460.

Bemporad, A., A. Garulli, S. Paoletti and A. Vicino (2005). A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control* **50**(10), 1567–1580.

Bemporad, A., A. Garulli, S. Paoletti and A. Vicinoy (2003). A greedy approach to identification of piecewise affine models. In: *Hybrid Systems: Computation and Control* (O. Maler and A. Pnueli, Eds.). Vol. 2623 of *Lecture Notes in Computer Sciences*. pp. 97–112. Springer-Verlag. New York, USA.

Dey, S. and J. A. Stori (2005). A Bayesian network approach to root cause diagnosis of process variations. *International Journal of Machine Tools and Manufacture* **45**(1), 75–91.

Domlan, E., B. Huang, F. Xu and A. Espejo (2010). Multiple model approach for inferential instruments design. In: *Proceedings of the* 9^{th} *International Symposium on Dynamics*

and Control of Process Systems (DYCOPS) (M. Kothare, M. Tade, A. Vande Wouwer and I. Smets, Eds.). IFAC. Leuven, Belgium.

Domlan, E., B. Huang, F. Xu and A. Espejo (2011). A decoupled multiple model approach for soft sensors design. *Control Engineering Practice* **11**(2), 126–134.

Dougan, P. and K. McDowell (1997). Sensor development in oil sand processing. In: *Proceedings of the Dynamic Modeling Control Applications for Industry Workshop* (M. Kothare, M. Tade, A. Vande Wouwer and I. Smets, Eds.). IEEE Industry Applications Society. Vancouver, Canada. pp. 68–73.

Ferrari-Trecate, G., M. Muselli, D. Liberati and M. Morari (2003). A clustering technique for the identification of piecewise affine systems. *Automatica* **39**(2), 205–217.

Jin, X. and B. Huang (2010). Robust identification of piecewise/switching Autoregressive eXogenous process. *AIChE Journal* **56**(7), 1829–1844.

Juloski, A. (2004). Observer Design And Identification Methods For Hybrid Systems. PhD thesis. Eindhoven University of Technology. Eindhoven, Netherlands.

Juloski, A. and S. Weiland (2006). A Bayesian approach to the identification of piecewise linear output error models. In: *Proceedings of the* 14^{th} *IFAC Symposium on System Identification* (B. Ninness and H. Hjalmarsson, Eds.). number 14. IFAC. Newcastle, Australia. pp. 374–379.

Juloski, A. L., S. Weiland and W. P. M. H. Heemels (2005). A Bayesian approach to identification of hybrid systems. *IEEE Transactions on Automatic Control* **50**(10), 1520–1533.

Kadlec, P. and B. Gabrys (2011). Local learning-based adaptive soft sensor for catalyst activation prediction. *AIChE Journal* **57**(5), 1288–1301.

Kadlec, P., R. Grbić and B. Gabrys (2011). Review of adaptation mechanisms for data-driven soft sensors. *Computers and Chemical Engineering* **35**(1), 1–24.

Khatibisepehr, S. and B. Huang (2008). Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Industrial and Engineering Chemistry Research* **47**(22), 8713–8723.

Khatibisepehr, S. and B. Huang (2012). A Bayesian approach to design of adaptive multi-model inferential sensors with application in oil sand industry. *Journal of Process Control* **22**(10), 1913–1929.

Khatibisepehr, S., B. Huang and S. Khare (2013). Design of inferential sensors in the process industry: A review of bayesian methods. *Journal of Process Control* p. in press.

Kim, M., Y. H. Lee, I. S. Han and C. Han (2005). Clustering-based hybrid soft sensor for an industrial polypropylene process with grade changeover operation. *Industrial and Engineering Chemistry Research* **44**(2), 334–342.

Korb, K. B. and A. E. Nicholson (2004). *Bayesian Artificial Intelligence*. first ed.. Chapman & Hall/CRC. London, UK.

Lauer, F. (2008). From Support Vector Machines to Hybrid System Identification. PhD thesis. Nancy University. Lorraine, France.

Ljung, L. (1999). *System Identification - Theory For the User*. second ed.. Prentice Hall. Upper Saddle River, USA.

Mirikitani, D. T. and N. Nikolaev (2010). Recursive Bayesian recurrent neural networks for time-series modeling. *IEEE Transactions on Neural Network* **21**(2), 262–274.

Murray-Smith, R. and T. R. Johansen (1997). *Multiple Model Approaches to Modelling and Control*. first ed.. Taylor & Francis. London, UK.

Paoletti, S., A. L. Juloski, G. Ferrari-Trecate and R. Vidal (2007). Identification of hybrid systems: A tutorial. *European Journal of Control* **13**(2-3), 242–260.

Qi, F. and B. Huang (2011). Bayesian methods for control loop diagnosis in presence of temporal dependent evidences. *Automatica* **47**(7), 1349–356.

Shao, X., B. Huang, J. M. Lee, F. Xu and A. Espejo (2011). Bayesian method for multirate data synthesis and model calibration. *AIChE Journal* **57**(6), 1514–1525.

Vidal, R., S. Soatto, Y. Ma and S. Sastry (2003). An algebraic geometric approach to the identification of a class of linear hybrid systems. In: *Proceedings of the* 42^{nd} *IEEE Conference on Decision and Control*. number 1. IEEE. Maui, USA. pp. 167–172.

Xu, Z., J. Zhao and J. Qian (2009). Nonlinear MPC using an identified LPV model. *Industrial and Engineering Chemistry Research* **48**(6), 3043–3051.

Chapter 7

A Probabilistic Framework for Real-time Performance Assessment of Inferential Sensors

7.1 Introduction

7.1.1 Practical Motivation

In order to maintain the reliability of an inferential sensor, it is important to assess the accuracy of its on-line predictions. Model uncertainty (plausible alternative model structures and/or parameters) is one of the major sources of prediction uncertainty (McKay *et al.*, 1999). In the context of process industries, deviations from design operating conditions are the main factors resulting in the model uncertainty and thus deterioration in performance of inferential sensors. In most of the classical identification methods, the objective is to minimize prediction errors pertaining to the identification data-set. Therefore, the generalization performance of the resulting inferential sensors is not guaranteed. In such cases, significant changes in the operating space in which the model has been identified would contribute to the model uncertainty.

Therefore, the conditional dependence of the reliability of inferential sensor predictions

A version of this chapter has been submitted to Control Engineering Practice (Khatibisepehr *et al.*, 2013*b*). An abbreviated version of this chapter will be presented at the 12^{th} IFAC International Symposium on Dynamics and Control of Process Systems, December 18-20, 2013, Mumble, India (Khatibisepehr *et al.*, 2013*a*).

on characteristics of the input space and reliability of the empirical process model should be thoroughly assessed in order to develop an on-line performance measure. From the application point of view, a desired performance measure has two essential characteristics. First, it should effectively estimate any significant deterioration in the prediction performance when process operates outside the valid inferential region. Second, implementation and interpretation of a performance metric should be simple enough for practitioners to use. Therefore, designing a proper performance index is not straightforward. Although inferential sensors have been widely used in process industries, there are only a few publications providing a methodology to assess their on-line performance. In Nomikos and MacGregor (1995); Vries and Braak (1995), approximate confidence intervals have been developed to assess the accuracy of PLS predictions based on the traditional statistical properties. The principal limitation of these approaches is that the internal empty regions within the identification data (*i.e.* the internal regions that do not contain any identification data points) cannot be diagnosed (Soto *et al.*, 2011). Kaneko *et al.* (2010) proposed a distance-based method to quantify the relationship between applicability domains and accuracy of inferential sensor predictions. The authors discussed that a larger Euclidean distance of an observation to the center of identification data and to its nearest neighbors would indicate a lower prediction accuracy. This method suffers from two major drawbacks. First, variability of the input variables is not taken into account when determining the Euclidean distance from the center. Second, the different effects of input variables on the prediction uncertainty are ignored by correlating the prediction accuracy with a general distance measure. Yang *et al.* (2009) applied an ensemble method to evaluate the uncertainty of inferential sensor predictions. The basic idea is to repeatedly generate bootstrap samples of the identification data-set to re-estimate inferential model parameters. With this multitude of models, the model variation and the average model bias can be estimated. Depending on the identification procedure used, however, this method

could be computationally intensive and would not be suited for on-line applications. Kaneko and Funatsu (2011) proposed to develop a multi-model inferential sensor based on the time difference of input variables in order to combine the information included in a set of local sub-models into a global predictive model. Furthermore, the accuracy of global predictions has been estimated using empirical models describing the relationship between standard deviation of local predictions and standard deviation of prediction errors. The major problem of this method is that small variation in local predictions does not necessarily imply a small prediction error. The proposed metric only reflects the degree of similarities between the prediction performance of different models and does not contain any information about the reliability of each individual model.

7.1.2 Main Contributions

To address the aforementioned issues, this Chapter provides a data-driven Bayesian framework for real-time performance assessment of inferential sensors. Such Bayesian frameworks utilizing discrete probability distributions have proven to be useful for a variety of fault diagnosis problems such as diesel engine fault diagnosis (Pernestål, 2007) and control loop performance diagnosis (Qi *et al.*, 2010). The major contribution of the present work is to formulate and solve the problem of inferential sensor performance assessment under a Bayesian framework utilizing both discrete and continuous probability distributions. The main focus is to characterize the effect of the operating space on the prediction accuracy in the absence of target measurements. The proposed method has the following attractive features:

1. *A priori* knowledge of process operation and underlying mechanisms can be easily incorporated in a Bayesian scheme so as to identify the criteria that might affect on-line performance of the designed inferential sensor.

2. Since probability density functions would reflect the actual data distribution, empty

regions within the identification data-set can be identified.

3. Correlations between input variables are taken into account.

4. Contribution of each input variable to prediction uncertainty is studied.

5. Its application does not depend on the identification techniques employed for inferential model development.

6. Its real-time implementation is computationally efficient.

7.1.3 Chapter Outline

The remainder of this Chapter is organized as follows. The problem of real-time performance assessment of inferential sensors is explained in Section 7.2. In Section 7.3, the problem of reliability analysis of real-time predictions is rigorously formulated under a Bayesian framework. The details of the Bayesian solution are presented for discrete operating statuses. The details of the Bayesian solution for continuous operating statuses are given in Section 7.4. The real-time performance assessment of multi-model inferential sensors is discussed in Section 7.5. A simulated continuous fermentation reactor is used as a working example to outline the ideas throughout Sections 7.3, 7.4, and 7.5. In Section 7.6, the effectiveness of the proposed Bayesian approach is demonstrated through industrial case studies; the methodology is applied for performance assessment of two industrial inferential sensors. Section 7.7 summarizes this Chapter with concluding remarks.

7.2 Problem Statement

Consider a class of inferential models given by

$$\hat{y}_t = g(\mathbf{u}_t; \Theta) \tag{7.1}$$

where $\hat{y}_t$ denotes the predicted value of query variable inferred from the real-time measurements of influential process variables, $\mathbf{u}_t = \{u_{k,t}\}_{k=1}^{K}$.

Evaluating the performance of an inferential sensor often amounts to analyzing the characteristics of prediction errors. Prediction error, also known as residual, is defined as the difference between the actual and predicted values of query variable:

$$e_t = y_t - \hat{y}_t \tag{7.2}$$

where y_t denotes the actual value of the query variable.

The absolute value of the prediction errors can be used to identify the events that would affect the reliability of the inferential model. Suppose that the performance of the inferential sensor at each time instant, r_t, can take R_e discrete reliability statuses, *i.e.* $r_t \in \{r^1, ..., r^{R_e}\}$. For instance, when $R_e = 3$, different degrees of reliability can be assigned to the inferential sensor predictions as follows:

$$r^j = \begin{cases} \text{Reliable} & 0 < |e_t| \leq 2\sigma_e \\ \text{Moderately reliable} & 2\sigma_e < |e_t| \leq 3\sigma_e \\ \text{Unreliable} & \text{Otherwise} \end{cases} \tag{7.3}$$

where the thresholds are considered as design parameters reflecting the tolerable amount of prediction error, and need to be adjusted based on the requirements of each application.

If y_t is observed, calculation of the performance index is straightforward. During on-line implementation of an inferential sensor, however, such real-time measurements are often not available frequently and regularly. Therefore, the main challenge is to assess the reliability of the inferential sensor predictions in the absence of actual values. Mathematically, the objective is to evaluate the conditional probability mass function $f\Big(e_t|\mathbf{u}_t, \hat{y}_t; \sigma_e\Big)$.

7.3 Real-time Performance Assessment from Discrete Operating Statuses

In this section, the problem of reliability analysis of real-time predictions is rigorously formulated under a Bayesian framework utilizing discrete operating statuses.

Given the identification (training) data-set $\mathcal{D} = \{(\mathbf{u}_t, y_t)\}_{t=1}^{N}$, an inferential sensor provides a real-time prediction, $\hat{y}_t$, on the basis of real-time measurements of k input variables, $\mathbf{u}_t = \{u_{k,t}\}_{k=1}^{K}$. It is noteworthy that the identification data-set contains both input and output measurements that are typically available from plant tests and/or historical plant operations often at lower sampling frequency. Therefore, the model prediction errors within the identification data-set, $\{e_t\}_{t=1}^{N}$, can be directly calculated from $\{(\mathbf{u}_t, y_t)\}_{t=1}^{N}$.

A set of indicator variables, $\{\mathbf{q}_t\}_{t=1}^{N} = \{q_t^{u_1}, \ldots, q_t^{u_K}\}_{t=1}^{N} \in \mathbb{R}^{K \times N}$, is introduced to partition the operating space into multiple modes. Suppose that each real-time input measurement, $u_{k,t}$, can take O_{u_k} discrete operating statuses. Prior knowledge of process operation (*e.g.* normal or unusual operating conditions) can be incorporated to properly partition the operating range of each process variable as well as the operating space of a set of process variables. In the absence of *a priori* knowledge, statistical analysis of operational and laboratory data may guide the choice of partitions. For instance, if it can be assumed that the input variables are Gaussian distributed random variables such that $u_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$, then different operating statuses may be assigned to the input measurements as follows:

$$q_t^{u_k} = \begin{cases} \text{Normal Oper. 1} & 0 < u_{k,t} - \mu_k \leq 2\sigma_k \\ \text{Normal Oper. 2} & -2\sigma_k \leq u_{k,t} - \mu_k \leq 0 \\ \text{Risky} & 2\sigma_k < u_{k,t} - \mu_k \leq 3\sigma_k \\ \text{Abnormal Oper 1} & 3\sigma_k < u_{k,t} - \mu_k \\ \text{Abnormal Oper. 2} & u_{k,t} - \mu_k < -2\sigma_k \end{cases} \tag{7.4}$$

where the thresholds are considered as design parameters chosen to provide adequate coverage of the operating space. Moreover, the generalization performance of the inferential sensor may guide the assignment of operating statuses. We would like to emphasize that our proposed method does not require any assumption about the probability density function (PDF) of input variables.

The on-line performance assessment of the inferential sensor amounts to evaluating the

posterior probability distribution of r_t given the operating status of the current measured inputs with reference to the historical data. The maximum *a posteriori* (MAP) estimate of reliability status is thus obtained from the following expression:

$$\widehat{r}_t = \underset{r_t}{\operatorname{argmax}}\, p(r_t|\mathbf{q}_t, \mathcal{D}) \tag{7.5}$$

Following the approach of Pernestål (2007); Qi *et al.* (2010), the desired posterior $p(r_t|\mathbf{q}_t, D)$ is derived next. Applying Bayes' rule, the posterior probability of r given reliability status of the current measured inputs and output can be written as

$$\begin{aligned} p(r_t|\mathbf{q}_t, \mathcal{D}) &= \frac{p(\mathbf{q}_t|r_t, \mathcal{D})p(r_t)}{p(\mathbf{q}_t|\mathcal{D})} \\ &= \gamma p(\mathbf{q}_t|r_t, \mathcal{D})p(r_t) \end{aligned} \tag{7.6}$$

where $\gamma^{-1} = p(\mathbf{q}_t|\mathcal{D}) = \sum p(\mathbf{q}_t|r_t, \mathcal{D})p(r_t)$ is a normalizing constant.

The random variable r_t is a categorical variable and can be modeled by

$$\begin{aligned} p(r_t) &= \prod_{j=1}^{R_e} p(r_t = r^j)^{[r_t = r^j]} \\ &= \prod_{j=1}^{R_e} \left(\varpi_j^e\right)^{[r_t = r^j]} \end{aligned} \tag{7.7}$$

where the operation $[r_t = r^j]$ evaluates to 1 if $r_t = r^j$ and evaluates to 0 otherwise. Note that $\{\varpi_j^e\}_{j=1}^{R_e}$ can be determined based on the prior knowledge of the inferential sensor prediction performance as well as the misclassification cost associated with each reliability status. Depending on the intended application of the inferential sensor, it might be desired to consider unequal misclassification costs. For instance, if the predicted values are automatically used for set-point adjustment, the unreliable predictions are the most expensive ones to misclassify. Such economic, operational, or environmental requirements can be considered within the proposed Bayesian framework by specifying appropriate prior distributions over the possible reliability statuses. If such knowledge is not available or relevant, a uniform prior distribution can be considered. The impact of the

prior distributions on the characteristics of the performance assessment framework will be illustrated through an industrial case study in Section 7.6.1.

Each input indicator variable is a random categorical variable. As a result, the vector of indicator variables $\mathbf{q}_t$ is an assembly of K random categorical variables. Given the reliability status of the inferential sensor, $\mathbf{q}_t$ can thus be modeled by a joint multinomial distribution with $S = \prod_{k=1}^{K} O_{u_k}$ points in its sample space (*i.e.* $\mathbf{q}_t \in \{Q^1, ..., Q^S\}$):

$$\begin{aligned} p(\mathbf{q}_t|\varpi_j^Q, r_t = r^j, \mathcal{D}) &= \prod_{s=1}^{S} p(\mathbf{q}_t = Q^s|r_t = r^j, \mathcal{D})^{[\mathbf{q}_t = Q^s]} \\ &= \prod_{s=1}^{S} \left(\varpi_{s|j}\right)^{[\mathbf{q}_t = Q^s]} \end{aligned} \tag{7.8}$$

where $\varpi_j^Q = \{\varpi_{s|j}\}_{s=1}^{S}$ is a set of hyperparameters characterizing the likelihood function in Equation 7.6.

Since the hyperparameters are typically not known *a priori*, the likelihood function is evaluated by integrating over the hyperparameters' space:

$$p(\mathbf{q}_t|r_t = r^j, \mathcal{D}) = \int p(\mathbf{q}_t|\varpi_j^Q, r_t = r^j, \mathcal{D}) p(\varpi_j^Q|r_t = r^j, \mathcal{D}) d\varpi_j^Q \tag{7.9}$$

The first term in the above integral is given by Equation 7.8. Besides, Bayes' rule can be applied to derive an explicit expression for the second term. Therefore, the posterior probability distribution of the hyperparameters given the identification data $\mathcal{D} = \{(\mathbf{q}_t, e_t)\}_{t=1}^{N}$ can be written as

$$\begin{aligned} p(\varpi_j^Q|r_t = r^j, \mathcal{D}) &= \frac{p(\mathcal{D}|\varpi_j^Q, r_t = r^j) p(\varpi_j^Q|r_t = r^j)}{p(r_t = r^j|\mathcal{D})} \\ &= \xi p(\mathcal{D}|\varpi_j^Q, r_t = r^j) p(\varpi_j^Q|r_t = r^j) \end{aligned} \tag{7.10}$$

where $\xi^{-1} = p(r_t = r^j|\mathcal{D}) = \int p(\mathcal{D}|\varpi_j^Q, r_t = r^j) p(\varpi_j^Q|r_t = r^j) d\varpi_j^Q$ is a normalizing constant.

The chain rule of probability theory is used to factorize the likelihood function in

Equation 7.10:

$$\begin{aligned} p(\mathcal{D}|\varpi_j^Q, r_t = r^j) &= \prod_{t=1}^{N_j} p(\mathbf{q}_t|\mathbf{q}_1, \ldots, \mathbf{q}_{t-1}, \varpi_j^Q, r_t = r^j) \\ &= \prod_{t=1}^{N_j} p(\mathbf{q}_t|\varpi_j^Q, r_t = r^j) \\ &= \prod_{s=1}^{S} \left(\varpi_{s|j}\right)^{\nu_{s|j}} \end{aligned} \tag{7.11}$$

where $N_j = \sum_{s=1}^{S} \nu_{s|j}$ denotes the number of samples in the identification data-set for which the reliability status of inferential sensor predictions was r^j. Equation 7.11 holds true only if it is reasonable to assume that the indicator variables are time-wise statistically independent.

Furthermore, the following Dirichlet distribution is considered as the hyperprior in Equation 7.10 to assure generality:

$$p(\varpi_j^Q|r_t = r^j) = \frac{1}{\mathrm{B}(A_j)} \prod_{s=1}^{S} \left(\varpi_{s|j}\right)^{\alpha_{s|j}-1} \tag{7.12}$$

where $\{\alpha_{s|j}\}_{s=1}^{S}$ are the Dirichlet parameters specified such that $A_j = \sum_{s=1}^{S} \alpha_{s|j}$ denotes the number of prior samples for which the reliability status of inferential sensor predictions was r^j. Also, $B(A_j)$ is the normalizing constant expressed in terms of the gamma function:

$$\mathrm{B}(A_j) = \frac{\prod_{s=1}^{S} \Gamma(\alpha_{s|j})}{\Gamma\left(\sum_{s=1}^{S} \alpha_{s|j}\right)} \tag{7.13}$$

where $\Gamma(x) = (x-1)!$ for all positive integers x. The fact that the Dirichlet distribution is the conjugate prior to the multinomial distributions justifies the choice of the Dirichlet hyperprior.

Combining Equations 7.10, 7.11 and 7.12, the posterior probability of the hyperparameters then becomes (DeGroot, 1970)

$$\begin{aligned} p(\varpi|r_t = r^j, \mathcal{D}) &= \frac{\prod_{s=1}^{S} \left(\varpi_{s|j}\right)^{\nu_{s|j}+\alpha_{s|j}-1}}{\int \prod_{s=1}^{S} \left(\varpi_{s|j}\right)^{\nu_{s|j}+\alpha_{s|j}-1} d\varpi} \\ &= \frac{\Gamma(A_j + N_j)}{\prod_{s=1}^{S} \Gamma(\alpha_{s|j} + \nu_{s|j})} \prod_{s=1}^{S} \left(\varpi_{s|j}\right)^{\nu_{s|j}+\alpha_{s|j}-1} \end{aligned} \tag{7.14}$$

Substituting Equations 7.8 and 7.14 into Equation 7.9, the posterior predictive distribution can be further expressed as

$$p(\mathbf{q}_t|r_t = r^j, \mathcal{D}) = \frac{\Gamma(A_j + N_j)}{\prod_{s=1}^{S}\Gamma(\alpha_{s|j} + \nu_{s|j})} \int \prod_{s=1}^{S} \left(\varpi_{s|j}\right)^{[\mathbf{q}_t = Q^s]} \left(\varpi_{s|j}\right)^{\nu_{s|j}+\alpha_{s|j}-1} d\varpi_j^Q \quad (7.15)$$

Hence,

$$\begin{aligned} p(\mathbf{q}_t = Q^d|r_t = r^j, \mathcal{D}) &= \frac{\Gamma(A_j + N_j)}{\Gamma(A_j + N_j + 1)} \times \frac{\Gamma(\alpha_{d|j} + \nu_{d|j} + 1)\prod_{s\neq d}^{S}\Gamma(\alpha_{s|j} + \nu_{s|j})}{\prod_{s=1}^{S}\Gamma(\alpha_{s|j} + \nu_{s|j})} \\ &= \frac{\alpha_{d|j} + \nu_{d|j}}{A_j + N_j} \end{aligned} \quad (7.16)$$

Finally, Equations 7.7 and 7.16 can be combined to obtain an explicit expression for the posterior probability distribution of Equation 7.6:

$$p(r_t = r^j|\mathbf{q}_t = Q^d, \mathcal{D}) = \gamma\varpi_j^e \frac{\alpha_{d|j} + \nu_{d|j}}{A_j + N_j} \quad (7.17)$$

The above posterior probability distribution can be evaluated to obtain the MAP estimates of the qualitative reliability status (*e.g.* highly reliable, moderately reliable, uncertain, etc.) of the inferential sensor (see Equation 7.5). As illustrated in Figure 7.1, $r_t = r^j$ implies that $b^{j-1} < |y_t - \hat{y}_t| \leq b^j$, where b^{j-1} and b^j are the lower and upper boundaries of $|e_t|$, respectively. Note that an expression similar to Equation 7.17 has also been derived by Qi *et al.* (2010) for fault isolation assuming that each discrete random variable can only take two values (*e.g.* faulty and normal). In this work, however, Equation 7.17 applies to multiple values of the discrete random variables.

In order to quantify the real-time performance of inferential sensors, it is proposed to associate a numerical value to each reliability status in the light of the historical probability distribution of prediction errors. Suppose that $F(\tilde{e}_t|\mathbf{u}_t, y_t)$ denotes the cumulative distribution function (CDF) of the absolute value of prediction error, $\tilde{e}_t = |y_t - \hat{y}_t|$. As the final stage of the training process, a quantifiable measure of reliability may be defined

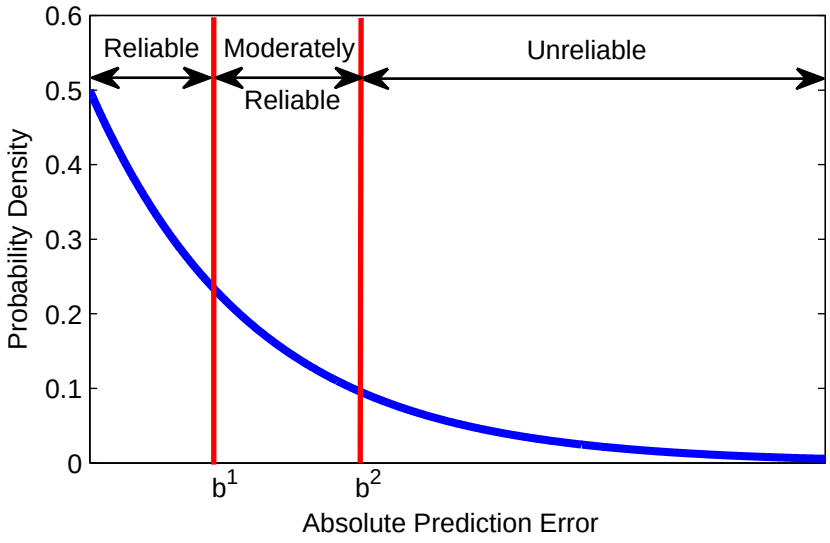


Figure 7.1: Probability density function of absolute value of prediction errors

based on the CDF of the random variable $\tilde{e}_t$ as follows:

$$\begin{aligned} r^j &\triangleq \frac{p(\tilde{e}_t > b^j)}{p(\tilde{e}_t > b^1)} \\ &= \frac{1 - p(\tilde{e}_t < b^j)}{1 - p(\tilde{e}_t < b^1)} \\ &= \frac{1 - F(b^j)}{1 - F(b^1)} \qquad \text{for} \quad j = 1 \cdots R_e \end{aligned} \tag{7.18}$$

where $p(\tilde{e}_t > b^1)$ is a normalizing constant.

Moreover, $F(b^j) = p(\tilde{e}_t < b^j)$ is the historical probability of the inferential model resulting in a prediction error smaller than b^j. Alternatively, $r^j \propto p(\tilde{e}_t > b^j)$ is the historical probability of the inferential model resulting in a prediction error greater than b^j. The values of r^j satisfy the following conditions:

$$r^1 = 1 \qquad \text{and} \qquad r^{R_e} \to 0 \ \text{ as } \ b^{R_e} \to \infty \tag{7.19}$$

where r^1 and r^{R_e} corresponds to the highest and lowest performance of the inferential sensor, respectively. To illustrate, the following reliability statuses can be specified, when r^j can only take three values, with reference to the cumulative distribution function shown

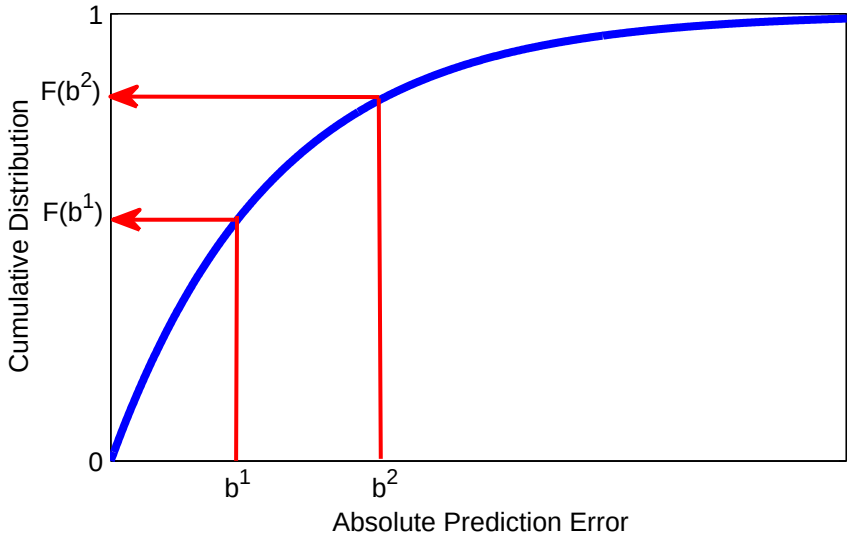


Figure 7.2: Cumulative distribution function of absolute value of prediction errors

in Figure 7.2:

$$r^j = \begin{cases} 1 & 0 < \tilde{e}_t \leq b^1 \\ \dfrac{1 - F(b^2)}{1 - F(b^1)} & b^1 < \tilde{e}_t \leq b^2 \\ 0 & b^2 < \tilde{e}_t \end{cases} \tag{7.20}$$

Note that the random variable $\tilde{e}_t$ can take any non-negative real value.

Finally, a reliability index (RI) can be assigned to each real-time prediction such that,

$$\begin{aligned} RI_t &\triangleq \mathbb{E}[r_t] \\ &= \sum_j p(r_t = r^j | \mathbf{q}_t = Q^d, \mathcal{D}) r^j \\ &= \frac{1}{1 - F(b^1)} \sum_j p(r_t = r^j | \mathbf{q}_t = Q^d, \mathcal{D})(1 - F(b^j)) \end{aligned} \tag{7.21}$$

where $RI_t \in [0, 1]$ is the expected value of the reliability status based on the estimated posterior probability of r_t.

If the absolute value of prediction errors follows a half-normal distribution, the reliability index is therefore expressed as

$$RI_t = \left(1 - \text{erf}\left(\frac{b^1}{\sqrt{2}\sigma_e}\right)\right)^{-1} \sum_j p(r_t = r^j | \mathbf{q}_t = Q^d, \mathcal{D}) \left(1 - \text{erf}\left(\frac{b^j}{\sqrt{2}\sigma_e}\right)\right) \tag{7.22}$$

7.3.1 Design Procedure

To summarize our discussion thus far, the procedure followed to design a Bayesian performance assessment framework is outlined in Algorithm 7.1.

Algorithm 7.1. *Real-time Performance Assessment of Inferential Sensors from Discrete Operating Statuses*

1. Include the prior knowledge of process operation to properly partition the operating range of each process variable as well as the operating space of a set of process variables. In the absence of relevant prior information, the operating range of the k^{th} input variable may be partitioned as follows:

   1.1. Approximate the CDF of the k^{th} input, $F_k(.)$, based on the identification data.

   1.2. Specify the operating range of each input variable as $F_k^{-1}(b) - F_k^{-1}(a)$, where $a, b \in [0, 1]$ and $b > a$. Note that a and b are design parameters chosen based on the quality of identification data. For instance, $a = 0.05$ and $b = 0.95$ can be selected to reduce the effect of outlying observations.

   1.3. Decide on the number of operating statuses, O_{u_k}, to be considered.

   1.4. Partition the operating range of u_k into equal-width intervals, *i.e.* the width of each interval would be equal to $(F_k^{-1}(b) - F_k^{-1}(a))/(O_{u_k} - 2)$.

   It should be noted that any other data-driven approach can be used to partition the operating space (Equation 7.4).

2. Specify a set of indicator variables to denote the operating status of each input variable.

3. Calculate the model prediction errors within the identification data-set.

4. Specify possible reliability statuses of inferential sensor predictions by analyzing the PDF of the absolute value of prediction error (Equation 7.3).

5. Assign a numeric value to each reliability status based on the CDF of the absolute value of prediction error (Equation 7.18).

6. Determine the prior distribution of reliability statuses, $\{\varpi_j^e\}_{j=1}^{R_e}$, based on the known prior information about the inferential sensor behavior. In the absence of such prior information, the prior distribution of reliability statuses can be specified based on the expected prediction performance of the inferential sensor as well as the misclassification costs involved in inaccurately predicting the reliability of predictions.

7. Specify appropriate Dirichlet parameters, $\{\alpha_{s|j}\}_{s=1}^{S}$, such that $A_j = \sum_{s=1}^{S} \alpha_{s|j}$ denotes the number of prior samples for which the reliability status of inferential sensor predictions was r^j (Equation 7.12).

8. Determine the likelihood parameters, $\{\nu_{s|j}\}_{s=1}^{S}$, such that $N_j = \sum_{s=1}^{S} \nu_{s|j}$ denotes the number of samples in the identification data-set for which the reliability status of inferential sensor predictions was r^j (Equation 7.11).

9. Characterize the posterior probability distribution of each reliability status, $p(r_t = r^j|\mathbf{q}_t, \mathcal{D})$ (Equation 7.17).

7.3.2 Continuous Fermentation Reactor Simulation

The governing equations of a continuous fermentation reactor (CFR) are given by (Henson and Seborg, 1997)

$$\dot{X} = -DX + \mu X \tag{7.23}$$

$$\dot{S} = D(S_f - S) - \frac{1}{Y_{X/S}}\mu X \tag{7.24}$$

$$\dot{P} = -DP + (\alpha\mu + \beta)X \tag{7.25}$$

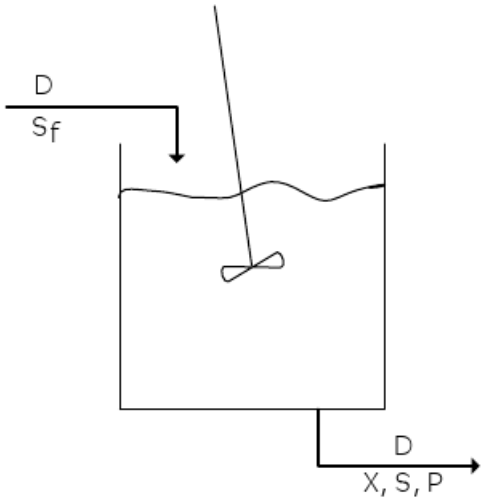


Figure 7.3: A simplified schematic of the CFR

where specific growth rate (μ) is defined as

$$\mu = \frac{\mu_m\left(1 - \frac{P}{P_m}\right)S}{K_m + S + \frac{S^2}{K_i}} \tag{7.26}$$

As shown in Figure 7.3, biomass concentration (X), substrate concentration (S) and product concentration (P) are state variables of the system. Dilution rate (D) and feed substrate concentration (S_f) are considered as system inputs. Moreover, cell-mass yield ($Y_{X/S}$), yield parameters (α, β), maximum specific growth rate (μ_m), product saturation constant (P_m), substrate saturation constant (K_m) and substrate inhibition constant (K_i) are model parameters.

The identification data was simulated using the variable settings presented in Table 7.1 as well as the non-linear dynamic model given by Equations 7.23-7.26. Data were collected at a relatively slow sampling rate so that data can be considered at the steady-state. An empirical linear model has been identified to describe the steady-state relationship between the input variables, dilution rate and feed substrate concentration, and the output quality variable, biomass concentration. Linear models are often used for development of inferential sensors in practical applications. In this case study, however, the identified linear

Table 7.1: Summary of simulated variables of CFR

Description	Distribution	Unit
Dilution rate, u_1	$\mathcal{N}(0.165, 4.5 \times 10^{-4})$	hr^{-1}
Substrate concentration, u_2	$\mathcal{N}(25, 14.15)$	kg/m^3
Noise added to biomass concentration, w	$\mathcal{N}(0, 0.022)$	-

Table 7.2: Parameter settings for performance assessment of the CFR inferential model

Property	Parameter Setting
Number of operating statuses for u_1	10
Number of operating statuses for u_2	10
Number of reliability statuses	3
Reliability statuses of inferential sensor	Reliable iff $0 < \|e_t\| \leq 1.379$
	Moderately reliable iff $1.379 < \|e_t\| \leq 2.758$
	Unreliable iff $2.758 < \|e_t\|$
Prior probability of a reliable prediction	$\varpi^e = \{0.40, 0.48, 0.12\}$
Total number of prior samples	$A = 22$
Total number of identification samples	$N = 2000$

model may not sufficiently represent the non-linear behavior of the fermentation process over such a wide operating space. Due to the inherent structural limitations of the identified model, the inferential sensor is thus expected to exhibit a degraded prediction performance in operating regions with low densities of identification data. Therefore, it is desirable to estimate the real-time prediction performance of the inferential sensor as well.

To determine the real-time prediction performance of this inferential sensor, a set of binary indicator variables is introduced as $\{\mathbf{q}_t\}_{t=1}^N = \{(q_t^{u_1}, q_t^{u_2})\}_{t=1}^N \in \mathbb{R}^{2\times N}$. Given the reliability status of the identified inferential model, the vector of quality variables $\mathbf{q}_t$ has $S = 10^2$ points in its sample space. The proposed Bayesian approach is used to assess the reliability status of the predictions. The parameter settings required to design a Bayesian reliability index are presented in Table 7.2. The boundaries of each reliability status have been selected based on the PDF of the historical absolute prediction error shown in Figure

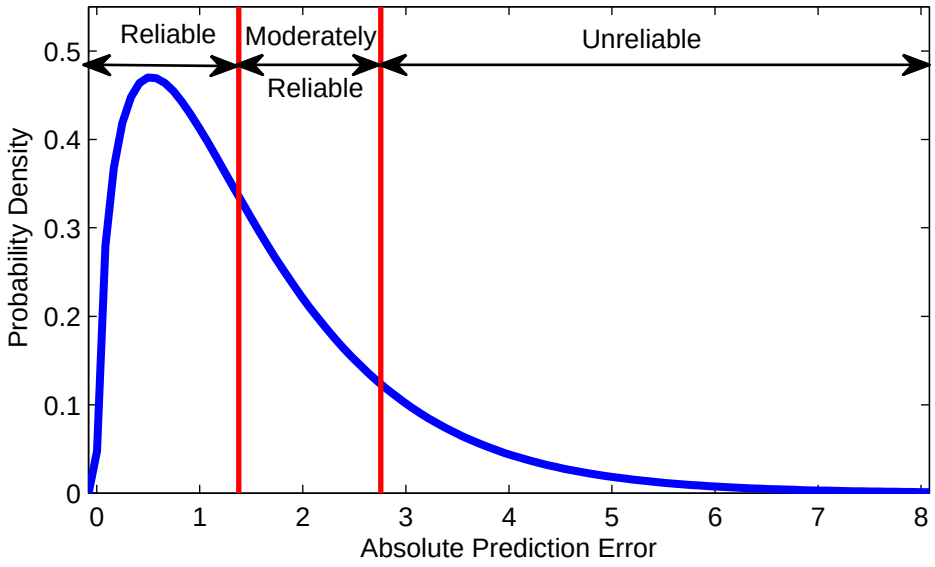


Figure 7.4: Probability density function of the absolute prediction error obtained from the CFR inferential model

7.4. Moreover, the data-driven approach recommended in Section 7.3.1 was applied to partition the operating range of each input variable.

Table 7.3 shows the confusion matrix obtained based on the reliability analysis results for $N = 1000$ test samples. The diagonal and cross-diagonal elements of the confusion matrix shown in Table 7.3 represent the number of predictions with correctly and incorrectly identified reliability status, respectively. The low number of incorrectly identified instances indicates that the method could effectively determine the reliability of inferential model predictions.

Let $n_{j|i}$ denote the number of instances with reliability status i (*i.e.* $r_t = r^i$) that are predicted to have reliability status j (*i.e.* $\hat{r}_t = r^j$). The entries of the confusion matrix can be used to quantify the performance of the proposed method in terms of the following metrics (Yang, 1999; Sebastiani, 2002; Sokolova and Lapalme, 2009):

1. **False positive or type I error**: The sum of cross-diagonal elements along the j^{th} column is the number of instances that are incorrectly assessed to have the j^{th}

Table 7.3: Confusion matrix for the Bayesian reliability analysis of the CFR inferential model using discrete operating statuses

		Predicted Status		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	432	68	0
	Moderately Reliable	20	371	15
	Unreliable	0	5	89

reliability status:

$$FP_j = \sum_{i \neq j} n_{j|i} \tag{7.27}$$

2. **False negative or type II error**: The sum of cross-diagonal elements along the j^{th} row is the number of instances with reliability status j that are incorrectly assessed to have the other reliability statuses:

$$FN_j = \sum_{j \neq i} n_{i|j} \tag{7.28}$$

3. **Sensitivity**: The number of instances correctly assessed to have the j^{th} reliability status, among all the instances with the j^{th} reliability status, determines the sensitivity to detecting the j^{th} reliability status:

$$Sen_j = \frac{n_{j|j}}{n_{j|j} + FN_j} \tag{7.29}$$

Moreover, the overall micro-averaged sensitivity can be defined as follows:

$$Sen^o = \frac{\sum_j n_{j|j}}{\sum_j (n_{j|j} + FN_j)} \tag{7.30}$$

4. **Precision**: The number of instances correctly assessed to have the j^{th} reliability status, among all the instances correctly and incorrectly assessed to have the j^{th} reliability status, determines the relevance of the instances classified to the j^{th} reliability status:

$$Pre_j = \frac{n_{j|j}}{n_{j|j} + FP_j} \tag{7.31}$$

   Moreover, the overall micro-averaged precision can be defined as follows:

$$Pre^o = \frac{\sum_j n_{j|j}}{\sum_j (n_{j|j} + FP_j)} \tag{7.32}$$

5. **Accuracy**: The fraction of instances correctly assessed whether or not they have the j^{th} reliability status determines the accuracy of detecting the j^{th} reliability status:

$$Acc_j = \frac{n_{j|j} + n_{\neg j|\neg j}}{\sum_j \sum_i n_{j|i}} \tag{7.33}$$

   Moreover, the overall micro-averaged accuracy can be defined as the fraction of all instances with correctly identified reliability status:

$$Acc^o = \frac{\sum_j n_{j|j} + n_{\neg j|\neg j}}{\sum_i \sum_j n_{i|j}} \tag{7.34}$$

A summary of the metrics quantifying the performance of the Bayesian reliability analysis of the CFR inferential model is reported in Table 7.4. The large values of the sensitivity, precision and accuracy are indicative of the effectiveness of the proposed method.

Regardless of the distribution of prediction error, a quantifiable measure of reliability can be defined solely based on the CDF of the absolute prediction error. From the CDF shown in Figure 7.5, it is evident that the prediction error does not follow a Gaussian distribution in this example. Figure 7.6 shows the reliability indices assigned to the inferential sensor predictions obtained for the test data. It can be observed that smaller reliability indices are assigned to larger prediction errors.

Table 7.4: Performance metrics for the Bayesian reliability analysis of the CFR inferential model using discrete operating statuses

Reliability Class	Type I Err.	Type II Err.	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	20	68	86.4	95.6	91.2
Moderately Reliable	73	35	91.4	83.6	89.0
Unreliable	15	5	94.7	85.6	98.0
Total	108	108	89.2	89.2	89.2

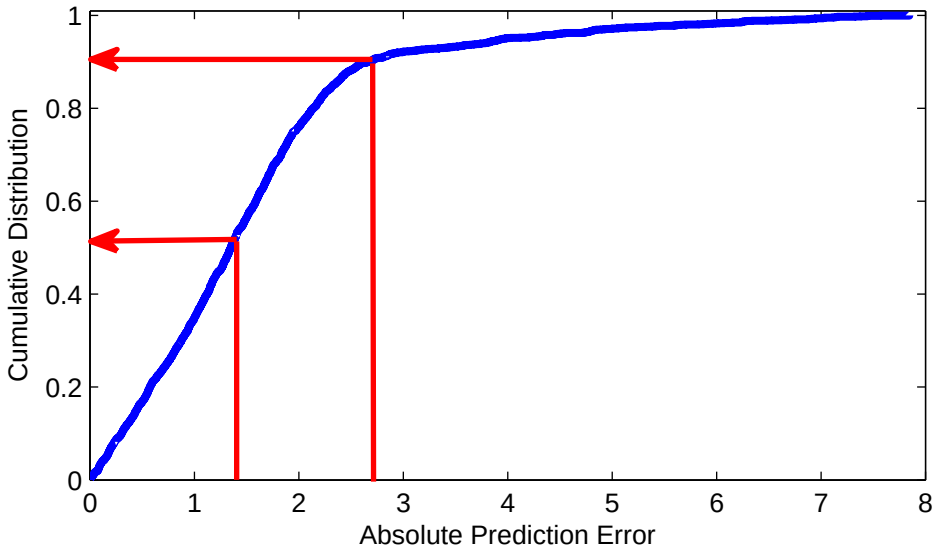


Figure 7.5: Cumulative distribution function of the absolute prediction error obtained from the CFR inferential model

7.4 Real-time Performance Assessment from Continuous Operating Statuses

As the number of operating statuses of input variables increases, discretization of the operating space becomes computationally intensive. In addition, in some applications, it may not be feasible to partition the input space. In such cases, the operating statuses can

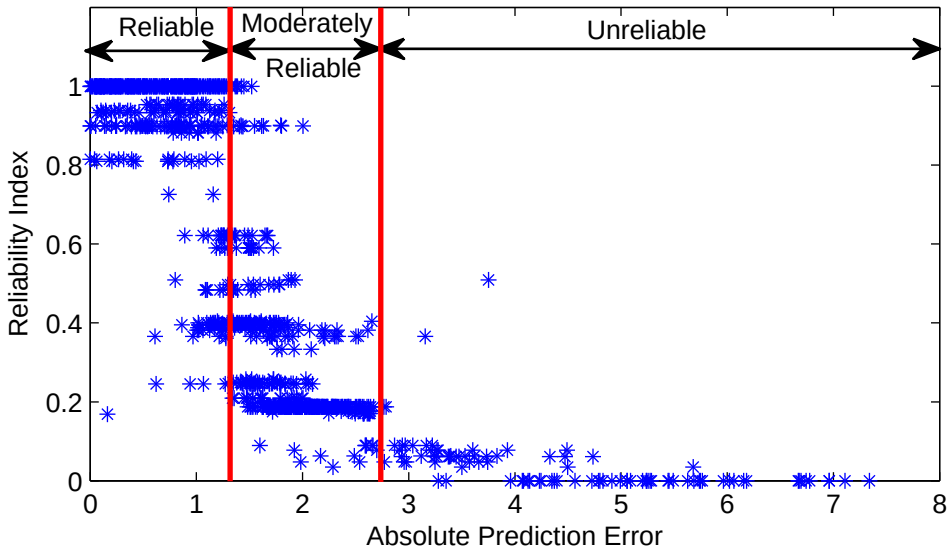


Figure 7.6: Performance assessment of the CFR inferential model using discrete operating statuses

be treated as continuous variables.

Given the identification data-set $\mathcal{D} = \{(\mathbf{u}_t, y_t)\}_{t=1}^{N} = \{Z_t\}_{t=1}^{N}$ with $\mathbf{u}_t = \{u_{k,t}\}_{k=1}^{K}$, the MAP estimate of reliability status is obtained from the following expression:

$$\widehat{r}_t = \underset{r_t}{\operatorname{argmax}}\, p(r_t|\mathbf{u}_t, \mathcal{D}) \tag{7.35}$$

Applying Bayes' rule, the posterior probability of r given input measurements can be written as

$$\begin{aligned} p(r_t|\mathbf{u}_t, \mathcal{D}) &= \frac{p(\mathbf{u}_t|r_t, \mathcal{D})p(r_t)}{p(\mathbf{u}_t|\mathcal{D})} \\ &= \phi p(\mathbf{u}_t|r_t, \mathcal{D})p(r_t) \end{aligned} \tag{7.36}$$

where $\phi^{-1} = p(\mathbf{q}_t|\mathcal{D}) = \sum p(\mathbf{u}_t|r_t, \mathcal{D})p(r_t)$ is a normalizing constant.

Given the reliability status of the inferential sensor, the likelihood of inferential sensor inputs in Equation 7.36 may be approximated by a multivariate Gaussian distribution such that

$$p(\mathbf{u}_t|\Sigma_j, \mu_j, r_t = r^j, \mathcal{D}) = \frac{1}{\sqrt{(2\pi)^K|\Sigma_j|}} \exp\Big(-\frac{1}{2}(\mathbf{u}_t - \mu_j)^T\Sigma_j^{-1}(\mathbf{u}_t - \mu_j)\Big) \tag{7.37}$$

where Σ_j denotes the covariance matrix and $\mu_j = [\mu_{1|j}, ..., \mu_{K|j}]$ denotes the vector of conditional mean values.

Σ_j and μ_j are the hyperparameters characterizing the likelihood function in Equation 7.37. If these hyperparameters are not known *a priori*, the likelihood function is evaluated by integrating over the hyperparameters' space:

$$p(\mathbf{u}_t|r_t = r^j, \mathcal{D}) = \int p(\mathbf{u}_t|\Sigma_j, \mu_j, r_t = r^j, \mathcal{D}) p(\Sigma_j, \mu_j|r_t = r^j, \mathcal{D}) d\Sigma_j d\mu_j \tag{7.38}$$

The first term in the above integral is given by Equation 7.37. Besides, Bayes' rule can be applied to derive an explicit expression for the second term. The posterior probability distribution of the hyperparameters given the identification data can be expressed as

$$\begin{aligned} p(\Sigma_j, \mu_j|r_t = r^j, \mathcal{D}) &= \frac{p(\mathcal{D}|\Sigma_j, \mu_j, r_t = r^j) p(\Sigma_j, \mu_j|r_t = r^j)}{p(\mathcal{D}|r_t = r^j)} \\ &= \xi p(\mathcal{D}|\Sigma_j, \mu_j, r_t = r^j) p(\mu_j|\Sigma_j, r_t = r^j) p(\Sigma_j|r_t = r^j) \end{aligned} \tag{7.39}$$

where $\xi^{-1} = p(\mathcal{D}|r_t = r^j) = \int p(\mathcal{D}|\Sigma_j, \mu_j, r_t = r^j) p(\Sigma_j, \mu_j|r_t = r^j)$ is a normalizing constant.

The chain rule of probability theory is used to factorize the likelihood function in Equation 7.39:

$$\begin{aligned} p(\mathcal{D}|\Sigma_j, \mu_j, r_t = r^j) &= \prod_{t=1}^{N_j} p(\mathbf{u}_t|u_1, \ldots, u_{t-1}, \Sigma_j, \mu_j, r_t = r^j) \\ &= \prod_{t=1}^{N_j} p(\mathbf{u}_t|\Sigma_j, \mu_j, r_t = r^j) \\ &= \prod_{t=1}^{N_j} \frac{1}{\sqrt{(2\pi)^K |\Sigma_j|}} \exp\left(-\frac{1}{2} (\mathbf{u}_t - \mu_j)^T \Sigma_j^{-1} (\mathbf{u}_t - \mu_j) \right) \\ &= \frac{1}{\sqrt{(2\pi)^{KN_j} |\Sigma_j|^{N_j}}} \exp\left(-\frac{1}{2} \sum_{t=1}^{N_j} (\mathbf{u}_t - \mu_j)^T \Sigma_j^{-1} (\mathbf{u}_t - \mu_j) \right) \\ &= \frac{1}{\sqrt{(2\pi)^{KN_j} |\Sigma_j|^{N_j}}} \exp\left(-\frac{1}{2} \mathrm{tr}(\Omega_j \Sigma_j^{-1}) \right) \end{aligned} \tag{7.40}$$

where $\Omega_j = \sum_{t=1}^{N_j} (\mathbf{u}_t - \mu_j)(\mathbf{u}_t - \mu_j)^T$.

It should be noted that $u_1, \ldots, u_t$ are assumed to be independent in the derivation of Equation 7.40. This assumption is reasonable as we can consider the steady-state data of sufficient sampling intervals.

The prior knowledge of the mean values may be incorporated through a Gaussian hyperprior:

$$p(\mu_j|\Sigma_j, r_t = r^j) = \sqrt{\frac{B_j^K}{(2\pi)^K|\Sigma_j|}} \exp\Big(-\frac{B_j}{2}(\mu_j - \mu_j^0)^T\Sigma_j^{-1}(\mu_j - \mu_j^0)\Big) \tag{7.41}$$

where μ_j^0 is the prior mean and B_j is the number of prior samples on the Σ_j scale, *i.e.* $\mu_j \sim \mathcal{N}(\mu_j^0, \Sigma_j/B_j)$.

Also, a hyperprior for the covariance matrix can be represented by the inverse Wishart distribution, which is the multivariate analogue of the inverse chi-squared distribution:

$$p(\Sigma_j|r_t = r^j) = \frac{h(A_j, \Psi_j)}{\sqrt{|\Sigma_j|^{A_j+K+1}}} \exp\Big(-\frac{1}{2}\text{tr}(\Psi_j\Sigma_j^{-1})\Big) \tag{7.42}$$

where A_j and Ψ_j respectively denote the degrees of freedom and the scale matrix for the inverse-Wishart distribution on Σ_j. Also, h is the normalizing constant given by

$$h(A_j, \Psi_j) = \Gamma_K^{-1}\Big(\frac{A_j}{2}\Big)\sqrt{\frac{|\Psi_j|^{A_j}}{2^{KA_j}}} \tag{7.43}$$

where Γ_{K+1} is the multivariate Gamma function,

$$\Gamma_K\Big(\frac{A_j}{2}\Big) = \pi^{K(K-1)/4}\prod_{k=1}^{K}\Gamma\Big(\frac{A_j + 1 - k}{2}\Big) \tag{7.44}$$

Therefore, putting together Equations 7.41 and 7.42 the normal inverse Wishart distribution defines a joint prior probability distribution over the hyperparameters:

$$\begin{aligned} p(\Sigma_j, \mu_j|r_t = r^j) &\triangleq \mathcal{NIW}(\mu_j^0, B_j, A_j, \Psi_j) \\ &= h(A_j, \Psi_j)\sqrt{\frac{B_j^K}{(2\pi)^K|\Sigma_j|^{A_j+K+2}}} \\ &\times \exp\Big(-\frac{1}{2}\text{tr}(\Psi_j\Sigma_j^{-1}) - \frac{B_j}{2}(\mu_j - \mu_j^0)^T\Sigma_j^{-1}(\mu_j - \mu_j^0)\Big) \end{aligned} \tag{7.45}$$

The fact that the normal inverse Wishart distribution is the conjugate prior to the multivariate Gaussian distribution justifies the choice of the hyperprior given by Equation 7.45. As pointed out by Gelman *et al.* (2003), as $A_j \to -1$, $B_j \to 0$ and $|\Psi_j| \to 0$ the joint prior probability distribution over the hyperparameters can be expressed as

$$p(\Sigma_j, \mu_j | r_t = r^j) \propto \frac{1}{\sqrt{|\Sigma_j|^{K+1}}} \tag{7.46}$$

which is the multivariate Jeffrey's prior distribution viewed as the limit of the normal inverse Wishart distribution.

Substituting Equations 7.45 and 7.40 into Equation 7.39, the joint posterior probability distribution is expressed as (Murphy, 2007)

$$\begin{aligned} p(\Sigma_j, \mu_j | r_t = r^j, \mathcal{D}) &= \mathcal{NIW}(\Sigma_j, \mu_j | \mu_j^*, B_j^*, A_j^*, \Psi_j^*) \\ &= h(A_j^*, \Psi_j^*) \sqrt{\frac{(B_j^*)^K}{(2\pi)^K |\Sigma_j|^{A_j^*+K+2}}} \\ &\times \exp\Big(-\frac{1}{2}\text{tr}(\Psi_j^* \Sigma_j^{-1}) - \frac{B_j^*}{2}(\mu_j - \mu_j^*)^T \Sigma_j^{-1}(\mu_j - \mu_j^*) \Big) \end{aligned} \tag{7.47}$$

where,

$$A_j^* = A_j + N_j \tag{7.48}$$

$$B_j^* = B_j + N_j \tag{7.49}$$

$$\bar{u}_j = N_j^{-1} \sum_{t=1}^{N_j} \mathbf{u}_t \tag{7.50}$$

$$\mu_j^* = \frac{B_j}{B_j + N_j} \mu_j^0 + \frac{N_j}{B_j + N_j} \bar{u}_j \tag{7.51}$$

$$\Psi_j^* = \Psi_j + \Omega_j + \frac{B_j N_j}{B_j + N_j} (\bar{u}_j - \mu_j^0)(\bar{u}_j - \mu_j^0)^T \tag{7.52}$$

Substituting Equations 7.47 and 7.37 into Equation 7.38, the predictive posterior distribution can be expressed as (Gelman *et al.*, 2003; Jackman, 2009)

$$p(\mathbf{u}_t | r_t = r^j, \mathcal{D}) = t_{A_j^* - K + 1}\big(\mathbf{u}_t | \mu_j^*, \Upsilon_j\big) \tag{7.53}$$

where

$$\Upsilon_j = \frac{\Psi_j^*(B_j^*+1)}{B_j^*(A_j^*-K+1)} \tag{7.54}$$

Finally, Equations 7.7 and 7.53 can be combined to obtain an explicit expression for the posterior probability distribution of Equation 7.36:

$$\begin{aligned} p(r_t = r^j|\mathbf{u}_t, \mathcal{D}) &= \frac{\phi\varpi_j^e}{\sqrt{|\Upsilon_j|(A_j^*-K+1)^K\pi^K}}\Gamma\left(\frac{A_j^*+1}{2}\right)\Gamma^{-1}\left(\frac{A_j^*-K+1}{2}\right) \\ &\times \left(1+\frac{1}{A_j^*-K+1}(\mathbf{u}_t-\mu_j^*)^T\Upsilon_j^{-1}(\mathbf{u}_t-\mu_j^*)\right)^{-(A_j^*+1)/2} \\ &= \frac{\phi\varpi_j^e}{\sqrt{|\Upsilon_j|(A_j^*-K+1)^K\pi^K}}\prod_{l=0}^{K/2-1}\left(\frac{A_j^*+1-K}{2}+l\right) \\ &\times \left(1+\frac{1}{A_j^*-K+1}(\mathbf{u}_t-\mu_j^*)^T\Upsilon_j^{-1}(\mathbf{u}_t-\mu_j^*)\right)^{-(A_j^*+1)/2} \end{aligned} \tag{7.55}$$

As discussed previously, the reliability index defined in Equation 7.21 can be used to obtain a quantifiable measure of reliability from $\widehat{r}_t = r^j$.

7.4.1 Design Procedure

To summarize our discussion on the performance assessment under a Bayesian framework utilizing continuous operating statuses, the design procedure is outlined in Algorithm 7.2.

Algorithm 7.2. *Real-time Performance Assessment of Inferential Sensors from Continuous Operating Statuses*

1. Calculate the model prediction errors within the identification data-set.
2. Specify possible reliability statuses of inferential sensor predictions by analyzing the PDF of the absolute value of prediction error (Equation 7.3).
3. Assign a numeric value to each reliability status based on the CDF of the absolute value of prediction error (Equation 7.18).

4. Determine the prior distribution of reliability statuses, $\{\varpi_j^e\}_{j=1}^{R_e}$, based on the expected prediction performance of the inferential sensor and/or the misclassification costs involved in inaccurately predicting the reliability of predictions.

5. Determine the prior distribution of hyperparameters given the reliability status, $p(\varpi_j^Q|r_t = r^j)$, based on the explicit prior knowledge. In the case of continuous operating statuses, the prior information over hyperparameters can be generally well-represented by normal inverse Wishart distribution (Equation 7.45). In the absence of relevant prior information, a non-informative prior distribution such as the multivariate Jeffrey's prior distribution can be assumed (Equation 7.46).

6. Characterize the posterior probability distribution of hyperparameters given the reliability status, $p(\varpi_j^Q|r_t = r^j, \mathcal{D})$ (Equation 7.47).

7. Characterize the likelihood of input variables for each reliability status, $p(\mathbf{u}_t|r_t = r^j, \mathcal{D})$, by integrating over the hyperparameters' space (Equation 7.53).

8. Characterize the posterior probability distribution of each reliability status, $p(r_t = r^j|\mathbf{u}_t, \mathcal{D})$ (Equation 7.55).

7.4.2 Continuous Fermentation Reactor Simulation

Let us once again consider the CFR simulation example presented in Section 7.3.2. The operating statuses of dilution rate and feed substrate concentration can be treated as continuous variables. The Bayesian approach for continuous operating statuses outlined in Section 7.4.1 is used to assess the reliability status of inferential sensor predictions. However, only a sub-set of identification data with N = 1000 is used to train the performance assessment framework. Table 7.5 shows the confusion matrix obtained based on the reliability analysis results for the same test data considered before. The low number of incorrectly identified instances indicates that the method could more effectively

determine the reliability of inferential model predictions. A summary of the metrics quantifying the performance of the Bayesian reliability analysis of the CFR inferential model is reported in Table 7.6. The large values of the sensitivity, precision and accuracy are indicative of the effectiveness of the proposed method. Comparing the performance metrics reported in Tables 7.4 and 7.6, it can be observed that the Bayesian framework with continuous operating statuses shows a better performance in estimating the reliability of inferential sensor predictions. It is noteworthy that the size of the identification data-set used in the continuous case is only half of the one used in the discrete case. Figure 7.7 shows the reliability indices assigned to inferential sensor predictions obtained for the test data. Like the discrete case, smaller reliability indices are assigned to larger prediction errors using the CDF in Figure 7.5. In comparison with the reliability indices shown in Figure 7.6, higher resolutions can be obtained using the continuous operating statuses.

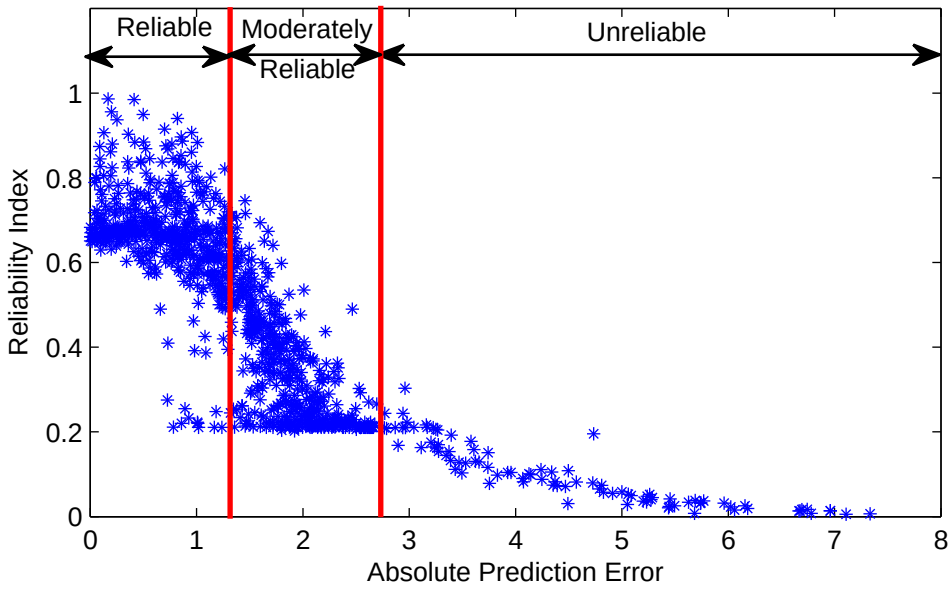


Figure 7.7: Performance assessment of the CFR inferential model using continuous operating statuses

Table 7.5: Confusion matrix for the Bayesian reliability analysis of the CFR inferential model using continuous operating statuses

		Predicted Status		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	446	54	0
	Moderately Reliable	28	368	10
	Unreliable	0	8	86

Table 7.6: Performance metrics for the Bayesian reliability analysis of the CFR inferential model using continuous operating statuses

Reliability Class	Type I Err.	Type II Err.	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	28	54	89.2	94.1	91.8
Moderately Reliable	62	38	90.6	85.6	90.0
Unreliable	10	8	91.5	89.6	98.2
Total	100	100	90.0	90.0	90.0

7.4.3 Comparison Between Discrete and Continuous Operating Statuses

In Sections 7.3 and 7.4, the Bayesian performance assessment frameworks have been developed for both discrete and continuous operating statuses. Each of these cases may have advantages and disadvantages depending on the application. The main features of the discrete and continuous operating statuses are compared below:

1. **Gaussian assumption.** The proposed Bayesian method does not make any

assumption about the PDF of the input variables in the case of discrete operating statuses. In the case of continuous operating statuses, however, it is assumed that the joint PDF of the input variables can be approximated by a multivariate Gaussian distribution or a mixture of multivariate Gaussian distributions. Although many practical data may be approximated by a Gaussian distribution, the approximation may cause loss of performance in some cases with strong non-Gaussian distributions.

2. **Information loss.** Discretization of the operating space may incur an information loss, since the input data falling in the same region become indistinguishable. In addition, small variations in values close to the border of neighboring partitions may cause unjustifiable effects on the posterior probability distribution of reliability status on inferential predictions. Therefore, the area and border of each partition should be chosen carefully.

3. **Computational load.** As the number of operating statuses of input variables increases, discretization of the operating space becomes computationally intensive. Moreover, a larger number of identification data is required to train the likelihood function. In such cases, the application of the continuous operating statuses is recommended. Regardless of the type of operating statuses chosen, the real-time implementation of the proposed Bayesian framework is computationally efficient and does not involve any practical difficulties.

7.5 Real-time Performance Assessment of Multi-model Inferential Sensors

In industrial applications to handle varying operating conditions, multi-model inferential sensors have also been used to represent multi-modal behavior of complex processes by concatenating multiple sub-models through an appropriate interpolation function. Therefore, the sub-models are reliable only for a particular operating region. The

interpolation function is parameterized by a set of representative process variables, through which the real-time operating mode can be properly identified.

Consider an input-output representation of a multi-model inferential sensor expressed as

$$\begin{cases} \hat{y}_{t|m} = f^{(m)}(\mathbf{u}_t; \Theta^{(m)}) \quad m = 1, \cdots, M \\ \hat{y}_t = \sum_{m=1}^{M} \psi_t^{(m)} \hat{y}_{t|m} \end{cases} \tag{7.56}$$

where M is the number of sub-models or, equivalently, the number of identified operating modes. Each sub-model, $m \in \{1, \cdots, M\}$, is characterized by its functional form, $f^{(m)}$, and a set of corresponding parameters, $\Theta^{(m)}$. The output of the m^{th} sub-model, $\hat{y}_{t|m}$, is inferred from the real-time measurements of influential process variables, $\mathbf{u}_t = \{u_t^k\}_{k=1}^K$. A proper interpolation function is defined to assign an importance weight, $\psi_t^{(m)}$, to the output of each sub-model, $\hat{y}_{t|m}$, in order to obtain a global prediction, $\hat{y}_t$.

The importance weight assigned to the output of the m^{th} sub-model can be viewed as the conditional probability of the m^{th} sub-model capturing the process behavior and can be evaluated using Bayes' theorem (Khatibisepehr and Huang, 2012):

$$p(m|s_t, \mathcal{D}) = \frac{p(s_t|m, \mathcal{D})p(m)}{\Sigma_{m=1}^{M} p(s_t|m, \mathcal{D})p(m)} \tag{7.57}$$

where s_t is a scheduling variable parameterized the interpolation function.

Alternatively, the conditional probability of the m^{th} sub-model capturing the process behavior can be evaluated as follows:

$$p(m|\{RI_t^{(m)}\}_{m=1}^M, \mathcal{D}) = \frac{p(\{RI_t^{(m)}\}_{m=1}^M|m, \mathcal{D})p(m)}{\Sigma_{m=1}^{M} p(\{RI_t^{(m)}\}_{m=1}^M|m, \mathcal{D})p(m)} \tag{7.58}$$

where $RI^{(m)}$ is the reliability index assigned to the m^{th} sub-model. In this way, the proposed Bayesian framework provides an automated mechanism for evaluating the representativeness of each sub-model, thus validating the adequacy of the interpolation function.

The overall reliability status of multi-model inferential sensors can be evaluated through

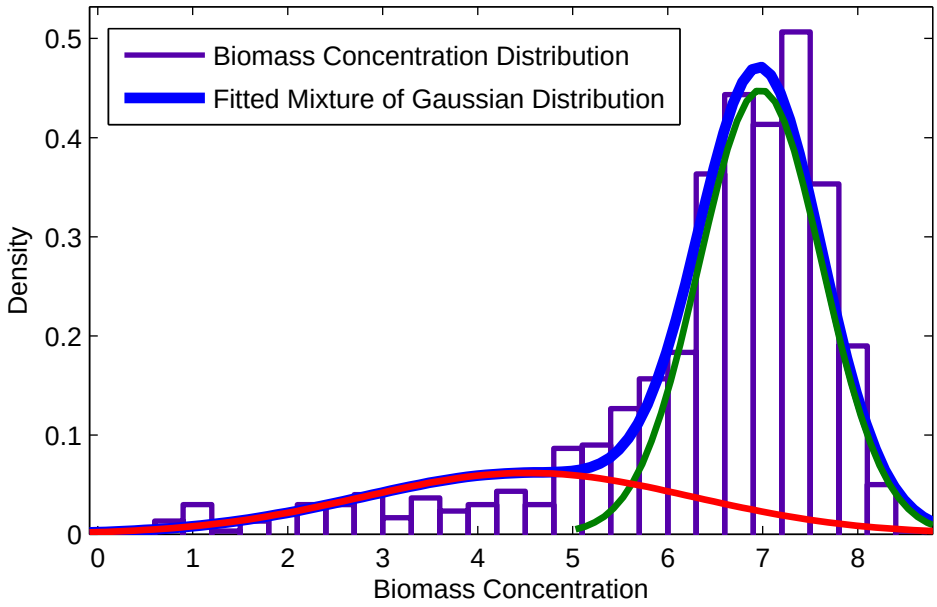


Figure 7.8: Probability distribution of biomass concentration

marginalization over plausible alternative sub-models, such that

$$p(r_t|\mathbf{u}_t, s_t, \mathcal{D}) = \sum_{m=1}^{M} p(r_t|m, \mathbf{u}_t, \mathcal{D}) p(m|s_t, \mathcal{D}) \tag{7.59}$$

or

$$p(r_t|\mathbf{u}_t, \{RI_t^{(m)}\}_{m=1}^{M}, \mathcal{D}) = \sum_{m=1}^{M} p(r_t|m, \mathbf{u}_t, \mathcal{D}) p(m|\{RI_t^{(m)}\}_{m=1}^{M}, \mathcal{D}) \tag{7.60}$$

It might be desirable to evaluate the overall reliability status of multi-model inferential sensors independently, without investigating the internal operations (*e.g.* the operating mode or performance of each sub-model). In this case, $p(r_t|\mathbf{u}_t, \mathcal{D})$ can be directly evaluated applying the Bayesian method outlined in Sections 7.3 and 7.4.

7.5.1 Continuous Fermentation Reactor Simulation

As illustrated in Figure 7.8, the probability distribution of biomass concentration can be approximated as a mixture of two Gaussian distributions. This motivates the application of multi-model inferential sensors to approximate the non-linear dynamic behavior of the CFR

Table 7.7: Confusion matrix for the Bayesian reliability analysis of the CFR multi-model inferential sensor using continuous operating statuses

		Predicted Status		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	578	73	0
	Moderately Reliable	61	248	4
	Unreliable	0	5	31

by switching between two piece-wise linear models. The Bayesian procedure proposed in Khatibisepehr and Huang (2012) is employed to design a multi-model inferential sensor and partition the operating space. In this way, the mean squared error of prediction is reduced to $MSE = 1.2$ from $MSE = 3.8$ obtained from the single model identified in Section 7.4.2. Therefore, tighter degrees of reliability are assigned to the multi-model inferential sensor predictions such that

$$r_t = \begin{cases} \text{Reliable} & 0 < |e_t| \leq 1.1 \\ \text{Moderately reliable} & 1.1 < |e_t| \leq 2.2 \\ \text{Unreliable} & 2.2 < |e_t| \end{cases} \tag{7.61}$$

The reliability status of the overall predictions are evaluated using Equation 7.59. Table 7.7 shows the confusion matrix obtained based on the reliability analysis results for $N = 1000$ test samples. Also, Table 7.8 presents the metrics quantifying the performance of the Bayesian performance assessment of the CFR inferential model. The large values of the sensitivity, precision and accuracy are indicative of the effectiveness of the proposed method. Figure 7.9 shows the reliability indices assigned to the overall inferential sensor predictions obtained for the test data. As expected, smaller reliability indices are assigned to larger prediction errors.

Table 7.8: Performance metrics for the Bayesian reliability analysis of the CFR multi-model inferential sensor using continuous operating statuses

Reliability Class	Type I Err.	Type II Err.	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	61	73	88.8	90.5	86.6
Moderately Reliable	78	65	79.2	76.1	85.7
Unreliable	4	5	86.1	88.9	99.1
Total	143	143	85.7	85.7	85.7

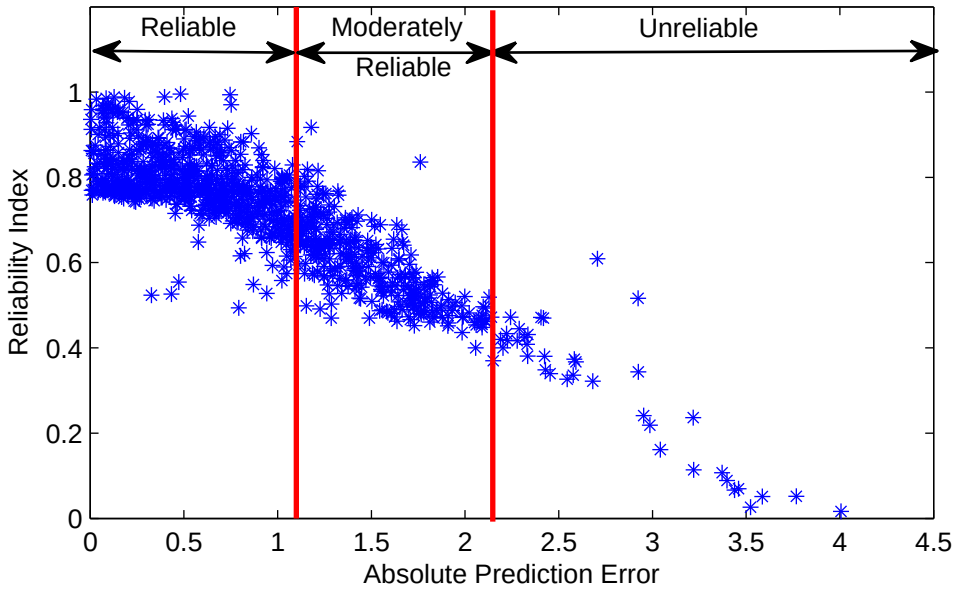


Figure 7.9: Performance assessment of the CFR multi-model inferential sensor using continuous operating statuses

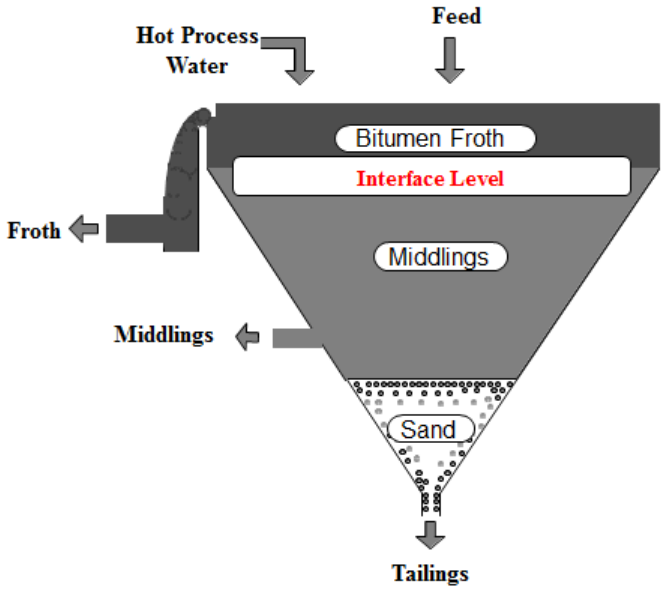


Figure 7.10: Schematic of the primary separation vessel

7.6 Industrial Case Studies

To demonstrate the effectiveness of the proposed method, the real-time performance assessment of inferential sensors developed from industrial data-sets is considered in this section.

7.6.1 Oil Sands Primary Extraction Plant

In the Bitumen primary extraction process, the conditioned oil sands slurry is fed into primary separation vessels (PSVs) to facilitate bitumen flotation and sand settling as illustrated in Figure 7.10. As a result of the gravity separation and frothing process, three layers are formed inside the PSV. The rocks and sand settle to the bottom of the vessel, forming a dense sand slurry tailings layer. The majority of the bitumen rises to the froth phase formed at the top of the vessel. The hard-to-separate clay particles accumulate between the froth and tailings layers, forming a middle layer inside the vessel called middlings.

The interface level between the froth and middlings layers is one of the key quality

variables that directly affect bitumen recovery. The camera-based real-time measurements are automatically used for monitoring and regulatory control purposes. However, the camera fails to provide any visible image of the interface level lying outside of the sight glass area. Therefore, a data-driven inferential sensor has been developed to complement/substitute the camera-based sensor. The influential process variables were chosen based on the simplified first principles analysis as well as the availability of hardware instruments providing real-time measurements of secondary process variables. Since the interface level measurements inferred from the camera signals are considered as the most trustful information source, they are selected as the reference values. To evaluate the reliability of the inferential sensor predictions, a Bayesian performance assessment framework was developed following the procedure presented in Section 7.3.1 for discrete operating statuses. As discussed in Section 7.3, the choice of the prior distribution of the reliability statuses would impact the characteristics of the presented performance assessment framework. Figure 7.11 and 7.12 show the impact of the prior distribution on the sensitivity and precision of the designed framework in detecting each reliability status. It can be observed that varying the prior probabilities would enable us to adjust the decision boundaries to some extent. For instance, one can increase the sensitivity in detecting the unreliable predictions through sacrificing the corresponding precision. In fact, the prior probabilities of the reliability statuses can be viewed as the importance weights assigned to satisfy the operational requirements.

In order to effectively evaluate the impact of the prior distribution of the reliability statuses, an appropriate measure of classification performance has to be selected. Due to the underlying assumptions of equal misclassification costs and relatively uniform class distribution, the performance measures derived from the confusion matrix may not be always suitable for comparison purposes. To address this issue, a cost matrix can be defined to weight the entries of the confusion matrix. In this way, the total misclassification cost

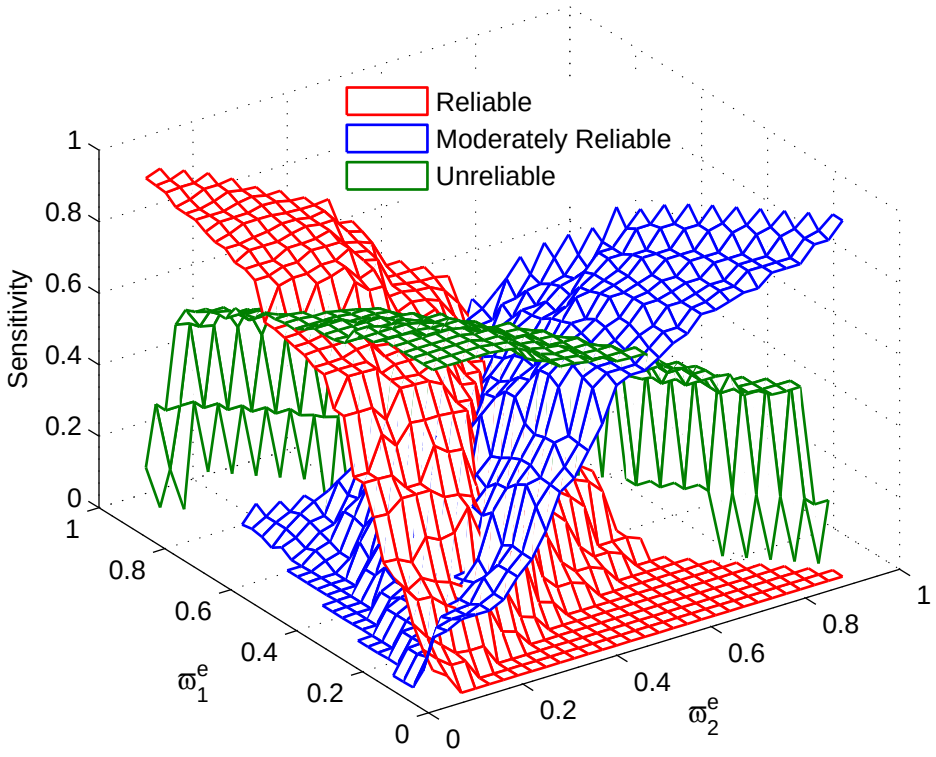


Figure 7.11: The impact of the prior distribution on the sensitivity of the designed framework in detecting each reliability status

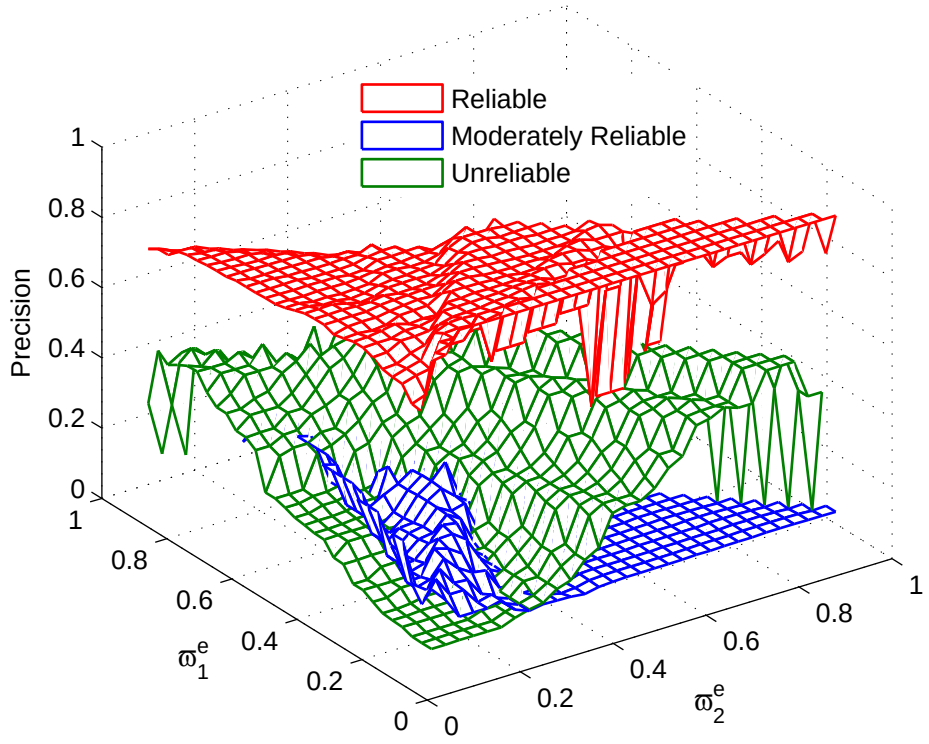


Figure 7.12: The impact of the prior distribution on the precision of the designed framework in detecting each reliability status

becomes a more effective metric for evaluating whether or not the operational requirements are satisfied within the developed performance assessment framework. Given confusion and cost matrices, the total misclassification cost (TMC) is defined as follows (Gorunescu, 2011):

$$MC = \sum_{i} \sum_{j} n_{i|j} c_{i|j} \tag{7.62}$$

where $n_{i|j}$ denote the number of instances with the j^{th} reliability status that are estimated to have the i^{th} reliability status and $c_{i|j}$ is the corresponding misclassification cost. Most often the cost of correct classification is zero, *i.e.* $c_{j|j} = 0$, because the right decision has been made. If the misclassification costs are all equal and the cost of correct classification is zero, it is straightforward to show that

$$MC = 1 - Acc^{o} \sum_{i} \sum_{j} n_{i|j} \tag{7.63}$$

Different sets of cost matrices can be used to illustrate the effect of the cost entries on the TMC. Table 7.9 shows the three cost matrices considered in this case study. Cost matrix I is defined to implement equal misclassification costs. Cost matrix II is defined to balance the confusion matrix such that each reliability status is represented by approximately equal proportions. In other words, the cost associated with incorrectly estimating each reliability status is lowered proportionally to their relative frequency. Cost matrix III is defined to not only balance the confusion matrix, but also assign higher costs to misclassifying the unreliable predictions.

Figure 7.13 shows the impact of the prior distribution of the reliability statuses on the TMC given different cost matrices. As expected, the TMC gradually varies across the entire range of possible prior probabilities. It is evident that the optimal prior distribution resulting in the minimum TMC depends on the cost matrix. It is of interest to investigate the effect of the optimal distributions on the sensitivity and precision of the designed framework in detecting each reliability status. The optimal prior distributions are used

Table 7.9: Cost matrices for the Bayesian performance assessment of the interface level inferential sensor

	Cost Matrix I	**Predicted Status**		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	0	1	1
	Moderately Reliable	1	0	1
	Unreliable	1	1	0

	Cost Matrix II	**Predicted Status**		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	0	$\frac{1}{\sum_i n_{i\|1}}$	$\frac{1}{\sum_i n_{i\|1}}$
	Moderately Reliable	$\frac{1}{\sum_i n_{i\|2}}$	0	$\frac{1}{\sum_i n_{i\|2}}$
	Unreliable	$\frac{1}{\sum_i n_{i\|3}}$	$\frac{1}{\sum_i n_{i\|3}}$	0

	Cost Matrix III	**Predicted Status**		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	0	$\frac{1}{\sum_i n_{i\|1}}$	$\frac{1}{\sum_i n_{i\|1}}$
	Moderately Reliable	$\frac{1}{\sum_i n_{i\|2}}$	0	$\frac{1}{\sum_i n_{i\|2}}$
	Unreliable	$\frac{2}{\sum_i n_{i\|3}}$	$\frac{1}{\sum_i n_{i\|3}}$	0

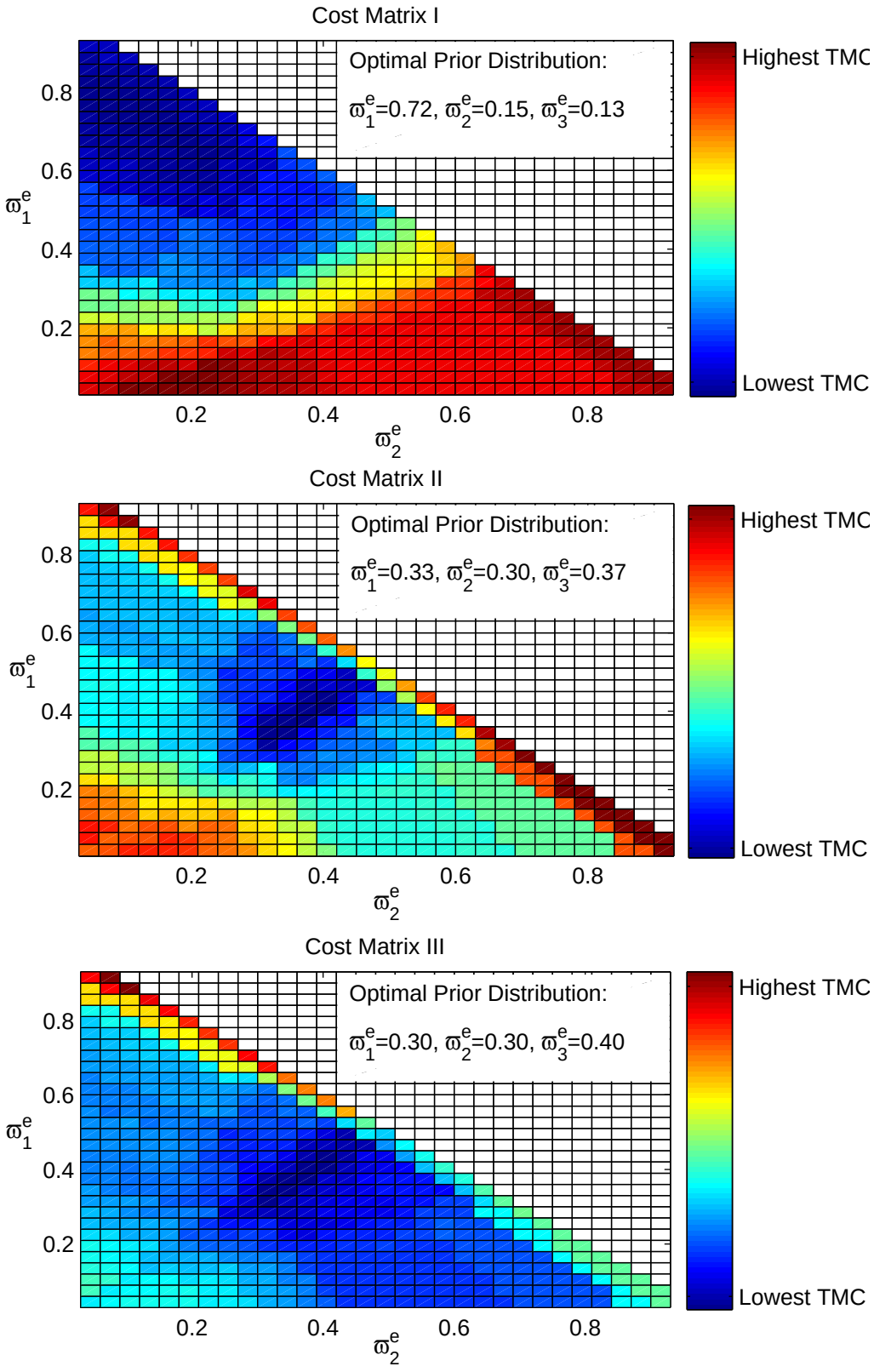


Figure 7.13: The impact of the prior distribution on the total misclassification cost

Table 7.10: Unbalanced confusion matrices for the Bayesian performance assessment of the interface level inferential sensor

	Confusion Matrix I	**Predicted Status**		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	2407	23	113
	Moderately Reliable	447	27	54
	Unreliable	141	7	217
	Confusion Matrix II	**Predicted Status**		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	1565	636	318
	Moderately Reliable	115	324	110
	Unreliable	26	41	301
	Confusion Matrix III	**Predicted Status**		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	1498	637	385
	Moderately Reliable	108	312	128
	Unreliable	19	41	308

to train three different frameworks for evaluating the reliability of the inferential sensor predictions. Table 7.10 presents the original (unbalanced) confusion matrices obtained from the reliability analysis results on the validation data-set. Cost matrix I is the default cost matrix in which all reliability statuses are treated equally. As a result, the misclassification of the majority reliability status (*i.e.* reliable predictions) contributes the most to the TMC. Therefore, in the resulting optimal prior distribution a large value is assigned to the majority reliability status, *i.e.* $\varpi_1^e = 0.72, \varpi_2^e = 0.15, \varpi_3^e = 0.13$. This would be equivalent to specifying the prior distribution by calculating the proportion of training samples attributed to each reliability status. As shown by confusion matrix I, the corresponding framework tends to favor the most frequent reliability status. Although the portion of correctly identified reliable predictions is high, the sensitivity to detecting moderately reliable and unreliable predictions is relatively low. By implementing cost matrix II, the cost associated with incorrectly estimating each reliability status is lowered proportionally to their relative frequency. In this way, all reliability statuses would have equal impact on the TMC. Therefore, the resulting optimal prior distribution is fairly uniform, *i.e.* $\varpi_1^e = 0.33, \varpi_2^e = 0.30, \varpi_3^e = 0.37$. Confusion matrix II shows that the corresponding framework demonstrates equally good sensitivity and precision with respect to all reliability statuses. In cost matrix III, the cost of misclassifying an unreliable prediction as reliable is higher than that of cost matrix II. As a result, the minority reliability status (*i.e.* unreliable predictions) has a greater impact on the TMC. Therefore, in the resulting optimal prior distribution a larger value is assigned to the minority reliability status, *i.e.* $\varpi_1^e = 0.30, \varpi_2^e = 0.30, \varpi_3^e = 0.40$. As can be seen from confusion matrix III, the corresponding framework tends to classify more instances as unreliable. It should be noted that the number of predictions incorrectly estimated to be unreliable increases with attempts to detect higher percentages of unreliable predictions.

Due to the intended application of the interface level inferential sensor, cost matrix II has

Table 7.11: Performance metrics for the Bayesian reliability analysis of the interface level predictions

Reliability Class	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	62.1	68.9	78.1
Moderately Reliable	59.0	61.9	74.2
Unreliable	81.8	71.5	83.1

been used in the design of the Bayesian performance assessment framework. A summary of the metrics quantifying the performance of the designed framework is reported in Table 7.11. These performance metrics have been derived from the balanced confusion matrix. The large values of the sensitivity, precision and accuracy are indicative of the effectiveness of the proposed method.

7.6.2 Oil Sands Secondary Extraction Plant

Following the approach of Khatibisepehr and Huang (2012), an adaptive multi-model inferential sensor was designed for real-time monitoring of Diluent to Bitumen (D:B) ratio in the product stream of an inclined plate settler (IPS). In the froth treatment plant, the bitumen froth is first mixed with diluent to enhance the density difference between the various components. The diluted bitumen froth is fed into various separation units including IPS and centrifuges. The IPS unit is one of the key froth treatment processes which allows for the space efficient gravity separation of diluted bitumen froth from the other components as illustrated in Figure 7.14. The IPS overflow product stream mainly consists of the diluted bitumen floating to the top of the vessel. The other components of the diluted froth such as water and minerals settle down at the bottom of the vessel. In the IPS unit, D:B ratio is monitored to control the quality of diluted bitumen and, thus, serves

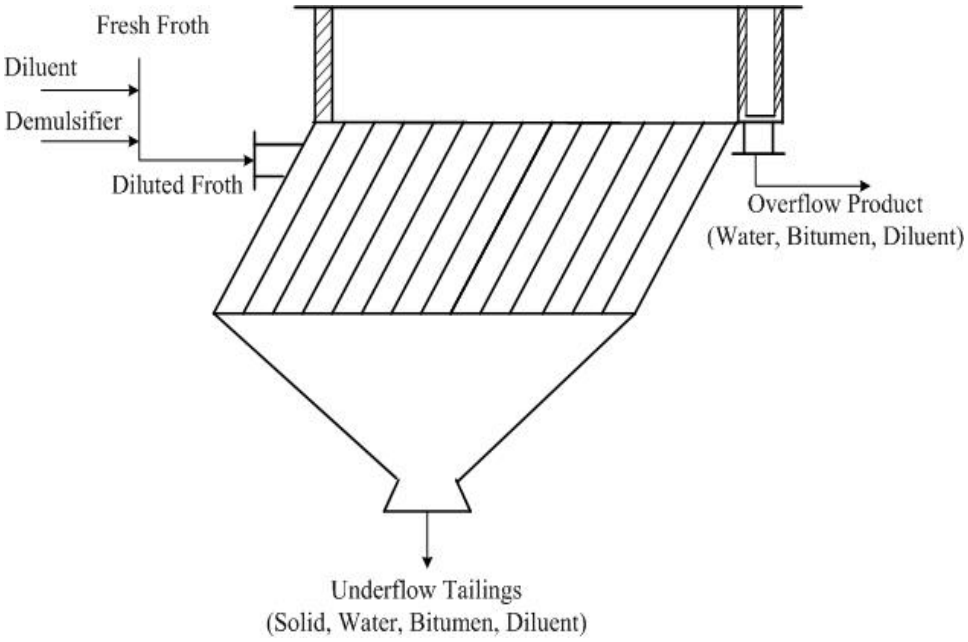


Figure 7.14: Schematic diagram of the inclined plates settler (IPS) operation

as one of the key indicators of the separation process performance. Based on the insight obtained from first principles and process data analysis, the following model was proposed to provide real-time predictions of DB_t:

$$\widehat{DB}_t = \sum_{m=1}^{M} p(m|F_{df,t}, \mathcal{D})\widehat{DB}_t^{(m)} \tag{7.64}$$

where $M = 2$ is the number of sub-models, $\widehat{DB}_t^{(m)}$ is the prediction obtained from the m^{th} sub-model, $\widehat{DB}_t$ is the global prediction, and $F_{df,t}$ is IPS diluted feed flow-rate that has been selected as the scheduling variable. Readers are referred to Khatibisepehr and Huang (2012) for more detail.

Real-time performance assessment of this inferential sensor was one of the practical issues that arose in the implementation stage. Therefore, it was required to develop a procedure for evaluating the accuracy of the predicted values in order to enhance the reliability of the designed inferential sensor. For direct evaluation of the reliability of the overall D:B predictions, $p(r_t|\mathbf{u}_t, \mathcal{D})$, a Bayesian performance assessment framework was developed following the procedure presented in Section 7.4.1 for continuous operating

statuses. An identification data-set was constructed from the data recorded from June 1 to December 30, 2012. Also, a validation data-set was constructed from the data recorded from January 1 to March 15, 2013. The identification data-set was used to train the performance assessment framework, while the validation data-set was reserved for cross validating the performance of the developed framework. All industrial data presented here has been normalized in order to protect proprietary information. Different degrees of reliability were assigned to the overall D:B predictions as follows:

$$\textbf{Case I:} \quad r^j = \begin{cases} \text{Reliable} & 0 < |e_t| \leq 2.5\sigma_e \\ \text{Unreliable} & \text{Otherwise} \end{cases} \tag{7.65}$$

$$\textbf{Case II:} \quad r^j = \begin{cases} \text{Reliable} & 0 < |e_t| \leq 0.75\sigma_e \\ \text{Moderately reliable} & 0.75\sigma_e < |e_t| \leq 2.5\sigma_e \\ \text{Unreliable} & \text{Otherwise} \end{cases} \tag{7.66}$$

where $e_t = DB_t^{Lab} - \widehat{DB}_t$ denotes the overall prediction error at time instant t.

Table 7.12 shows the confusion matrices obtained based on the reliability analysis results on the validation data-set for both Case I and Case II. Table 7.13 presents a summary of the metrics quantifying the performance of the designed frameworks. These performance metrics have been derived from the balanced confusion matrix. The large values of the sensitivity, precision and accuracy for Case I show the effectiveness of the designed performance assessment framework. The framework designed for Case II still performs well in identifying the unreliable D:B predictions. However, it remains difficult to distinguish between reliable and moderately reliable predictions due to the following reasons: (1) Since the structure of the sub-models have been selected based on mass balance equations, the model uncertainty is only reflected in the model parameters. (2) A large operating space has already been accounted for in design of this multi-model inferential sensor. Yet, the large values of accuracy for all reliability statuses imply the overall good performance in Case II.

Table 7.12: Confusion matrices for the Bayesian performance assessment of the D:B multi-model inferential sensors using continuous operating statuses

	Case I	**Predicted Status**		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	649	_	54
	Moderately Reliable	_	_	_
	Unreliable	1	_	14
	Case II	**Predicted Status**		
		Reliable	Moderately Reliable	Unreliable
Actual Status	Reliable	214	164	16
	Moderately Reliable	97	171	41
	Unreliable	0	1	14

Table 7.13: Performance metrics for the Bayesian performance assessment of the D:B multi-model inferential sensors using continuous operating statuses

Case I Reliability Class	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	92.2	93.3	92.8
Unreliable	93.3	92.4	92.8
Case II Reliability Class	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	54.3	63.4	74.3
Moderately Reliable	55.3	53.4	69.0
Unreliable	93.3	84.4	92.0

As discussed in Section 7.5, the importance weight assigned to the output of the m^{th} sub-model can be obtained by evaluating either $p(m|s_t, \mathcal{D})$ or $p(m|\{RI_t^{(m)}\}_{m=1}^{M}, \mathcal{D})$. In Khatibisepehr and Huang (2012), the importance weights were obtained based on $p(m|F_{df,t}, \mathcal{D})$ given the real-time feed flow-rate measurements (Method 1). In the present work, the importance weights were obtained based on $p(m|\{RI_t^{(m)}\}_{m=1}^{2}, \mathcal{D})$ given the estimated reliability indices assigned to the sub-models (Method 2). That is, the following model is used to provide real-time predictions of DB_t:

$$\widehat{DB}_t = \sum_{m=1}^{2} p(m|\{RI_t^{(m)}\}_{m=1}^{2}, \mathcal{D})\widehat{DB}_t^{(m)} \tag{7.67}$$

In order to evaluate $p(m|\{RI_t^{(m)}\}_{m=1}^{2}, \mathcal{D})$, the following steps were taken:

1. The Bayesian approach outlined in Section 7.4.1 was followed to design a performance assessment framework for each sub-model. The metrics quantifying the performance of the designed frameworks are reported in Tables 7.14 and 7.15.

Table 7.14: Performance metrics for the Bayesian performance assessment of the first sub-model using continuous operating statuses

Case I Reliability Class	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	90.5	91.0	90.8
Unreliable	91.0	90.6	90.8
Case II Reliability Class	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	63.5	63.5	75.6
Moderately Reliable	54.4	53.2	68.9
Unreliable	86.6	88.5	91.8

2. Numeric values were assigned to the reliability status of $\widehat{DB}_t^{(m)}$ based on the historical CDF of the absolute value of the prediction errors resulted from the m^{th} sub-model, *i.e.* $|e_t^{(m)}| = |DB_t^{Lab} - \widehat{DB}_t^{(m)}|$.

Mean absolute error (MAE), standard deviation of errors (StdE), and mean squared error (MSE) for the real-time D:B predictions from Equations 7.64 and 7.67 have been compared. The results are obtained through the comparison of the inferential sensor predictions and the laboratory measurements. It was observed that the values of MAE, StdE and MSE using Method II-Case I have been reduced by 6.2%, 4.6% and 9.6%, respectively. Similarly, the values of MAE, StdE and MSE using Method II-Case II have been reduced by 4.6%, 4.5% and 9.3%, respectively. Since the real-time performance of the sub-models can be successfully assessed within the designed framework, the estimated reliability indices can be confidently used to assign smaller weights to less reliable sub-models. As a result, the importance weights assigned in Method II improve the overall prediction performance of

Table 7.15: Performance metrics for the Bayesian performance assessment of the second sub-model using continuous operating statuses

Case I Reliability Class	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	87.1	100	93.5
Unreliable	100	88.5	93.5
Case II Reliability Class	Sensitivity (%)	Precision (%)	Accuracy (%)
Reliable	56.1	60.8	72.8
Moderately Reliable	52.3	51.2	66.8
Unreliable	92.9	85.8	92.5

the multi-model inferential sensors.

7.7 Concluding Remarks

In this Chapter, a data-driven Bayesian framework for real-time performance assessment of inferential sensors was proposed. The main focus was to characterize the effect of the operating space on the prediction reliability in the absence of target measurements. The details of the design procedures for both discrete and continuous operating statuses were presented. Moreover, the real-time performance assessment of multi-model inferential sensors was discussed. It was shown that the application of the proposed Bayesian solution does not depend on the identification techniques employed for inferential model development. Furthermore, its real-time implementation is computationally efficient and simple for practitioners to use. The effectiveness of the proposed method was demonstrated through simulation and industrial case studies.

Bibliography

DeGroot, M. (1970). *Optimal Statistical Decisions*. first ed.. McGraw-Hill. New York, USA.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (2003). *Bayesian Data Analysis*. second ed.. Chapman & Hall/CRC. Boca Raton, USA.

Gorunescu, F. (2011). *Data Mining: Concepts, Models And Techniques*. first ed.. Springer-Verlag. Berlin, Germany.

Henson, M. A. and D. E. Seborg (1997). *Nonlinear Process Control*. first ed.. Prentice-Hall Inc.. Upper Saddle River, USA.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. first ed.. John Wiley & Sons. Chichester, UK.

Kaneko, H. and K. Funatsu (2011). Improvement and estimation of prediction accuracy of soft sensor models based on time difference. In: *Modern Approaches in Applied Intelligence* (K. G. Mehrotra, C. K. Mohan, J. C. Oh, P. K. Varshney and M. Ali, Eds.). Vol. 6703 of *Lecture Notes in Computer Science*. pp. 115–124. Springer-Verlag. Berlin, Germany.

Kaneko, H., M. Arakawa and K. Funatsu (2010). Applicability domains and accuracy of prediction of soft sensor models. *AIChE Journal* **57**(6), 1506–1513.

Khatibisepehr, S. and B. Huang (2012). A Bayesian approach to design of adaptive multi-model inferential sensors with application in oil sand industry. *Journal of Process Control* **22**(10), 1913–1929.

Khatibisepehr, S., B. Huang, S. Khare and R. Kadali (2013*a*). Real-time performance assessment of inferential sensors. *The* 12^{th} *IFAC International Symposium on Dynamics and Control of Process Systems (DYCOPS)*.

Khatibisepehr, S., B. Huang, S. Khare, E. Domlan, F. Xu, A. Espejo and R. Kadali (2013*b*). A probabilistic framework for real-time performance assessment of inferential sensors. *Control Engineering Practice* p. in review.

McKay, M. D., J. D. Morrison and S. C. Upton (1999). Evaluating prediction uncertainty in simulation models. *Computer Physics Communications* **117**(1-2), 44–51.

Murphy, K. (2007). Conjugate Bayesian analysis of the Gaussian distribution. Technical report. University of British Columbia.

Nomikos, P. and J. F. MacGregor (1995). Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems* **30**(1), 97–108.

Pernestål, A. (2007). A Bayesian Approach To Fault Isolation With Application To Diesel Engine Diagnose. PhD thesis. KTH School of Electrical Engineering. Stockholm, Sweden.

Qi, F., B. Huang and E. C. Tamayo (2010). A Bayesian approach for control loop diagnosis with missing data. *AIChE Journal* **56**(1), 179–195.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* **34**(1), 1–47.

Sokolova, M. and G. Lapalme (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management* **45**(4), 427–437.

Soto, A. J., G. E. Vazquez, M. Strickert and I. Ponzoni (2011). Target-driven subspace mapping methods and their applicability domain estimation. *Chemometrics and Intelligent Laboratory Systems* **30**(9), 779–789.

Vries, S. De and C. J. F. Ter Braak (1995). Prediction error in partial least squares regression: A critique on the deviation used in the unscrambler. *Chemometrics and Intelligent Laboratory Systems* **30**(2), 239–245.

Yang, H. Y., S. H. Lee and M. G. Na (2009). Monitoring and uncertainty analysis of feedwater flow rate using data-based modeling methods. *IEEE Transactions on Nuclear Science* **56**(4), 2426–2433.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval* **1**(1-2), 69–90.

Chapter 8

General Discussion and Concluding Remarks

8.1 General Discussion

Several challenging issues encountered in the development and implementation of inferential sensors were addressed in this thesis. It has been shown that the problems of interest can be formulated as rigorous conditional probabilistic problems within a Bayesian framework. To summarize our discussion, the methods proposed throughout the thesis can be incorporated to lay out a novel unified Bayesian framework for the design of multi-model inferential sensors.

Given the identification data-set $\mathcal{D} = \{(s_t, \mathbf{u}_t, y_t)\}_{t=1}^{N}$, a multi-modal system can be represented as follows:

$$\begin{cases} \mathbf{u}_t = \tilde{\mathbf{u}}_t + \mathbf{e}_t \\ s_t = \tilde{s}_t + \epsilon_t \\ \tilde{y}_t^{(m)} = g^{(m)}\big(\tilde{\mathbf{u}}_t; \Theta^{(m)}\big) \quad m = 1, \cdots, M \\ y_t = \sum_{m=1}^{M} \psi_t^{(m)}(\tilde{s}_t)\tilde{y}_t^{(m)} + \varepsilon_t \end{cases} \tag{8.1}$$

Several terms in the above formulation are described below:

1. **Input variables**: $\mathbf{u}_t = \{u_{j,t}\}_{j=1}^{P} \in \mathbb{R}^P$ and $\tilde{\mathbf{u}}_t = \{\tilde{u}_{j,t}\}_{j=1}^{P} \in \mathbb{R}^P$ denote vectors of the measured (observed) and noise-free (unobserved) values of the input variables at time instant t, respectively.

2. **Output variables**: $y_t \in \mathbb{R}$ and $\tilde{y}_t \in \mathbb{R}$ denote the measured (observed) and noise-free (unobserved) values of the output variable at time instant t, respectively.

3. **Local sub-models**: Each sub-model, $m \in \{1 \cdots, M\}$, is represented by its functional form, $g^{(m)}$, and a set of corresponding parameters, $\Theta^{(m)}$.

4. **Importance weights**: A proper interpolation function is defined to assign an importance weight, $\psi^{(m)}$, to the output of each sub-model, $y^{(m)}$, in order to combine the information included in a set of local sub-models into a global predictive model. Let $\Psi = [\psi_1, \cdots, \psi_N] \in \mathbb{R}^{M \times N}$ denotes the interpolation matrix with $\psi_t = \{\psi_t^{(m)}\}_{m=1}^{M} \in \mathbb{R}^M$. A set of mode indicator variables, $\mathbf{i}_{1:N} = \{i_1, \cdots, i_N\}$, can also be introduced to denote the identity of each data pair. That is, the indicator variable represents the most probable operating mode at each time instant.

5. **Scheduling variable**: The interpolation function is parameterized by a scheduling variable that effectively determines the discrete-state dynamics at each time instant. $s_t \in \mathbb{R}$ and $\tilde{s}_t \in \mathbb{R}$ denote the measured (observed) and noise-free (unobserved) values of the scheduling variable at time instant t, respectively.

6. **Measurement noise in input variables**: The measurement noise in the input variables, $\mathbf{e}_t = \{e_{j,t}\}_{j=1}^{P} \in \mathbb{R}^P$, is assumed to be independent and identical, following a Gaussian distribution:

$$\mathbf{e}_t \sim \mathcal{N}\big(\mathbf{0}_P, \Sigma_{\mathbf{e}}\big), \quad \mathbf{0}_P = [0, \cdots, 0]^T \in \mathbb{R}^P \tag{8.2}$$

   where the covariance matrix of the input measurement noise, $\Sigma_{\mathbf{e}} \in \mathbb{R}^{P \times P}$, is diagonal. That is, $\Sigma_{\mathbf{e}} = \text{diag}(\sigma_{e_1}^2, \cdots, \sigma_{e_P}^2)$.

7. **Measurement noise in scheduling variable**: The measurement noise in the scheduling variable, $\epsilon_t \in \mathbb{R}$, is assumed to follow a Gaussian distribution:

$$\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{8.3}$$

8. **Measurement noise in output variables:** It is assumed that the measured values of the output variable are contaminated by the outliers. Thus, the measurement noise in the output variable, $\varepsilon_t \in \mathbb{R}$, is distributed as

$$\varepsilon_t \sim \delta G(\varepsilon) + (1-\delta)\mathcal{N}(0, \sigma_\varepsilon^2) \tag{8.4}$$

where δ is the unknown prior probability of appearance of an outlier and $G(\varepsilon)$ denotes the contaminating distribution. A set of quality indicator variables, $\mathbf{q}_{1:N} = \{q_1, \cdots, q_N\}$, can be introduced to denote the quality of the output measurements. That is, the indicator variable associated with each data point determines whether that observation comes from the regular or contaminating distribution.

The identification problem is to estimate the noise-free identification data, $\tilde{\mathcal{D}} = \{(\tilde{s}_t, \tilde{\mathbf{u}}_t, \tilde{y}_t)\}_{t=1}^N$, the model parameters, $\Theta = \{\Theta^{(m)}\}_{m=1}^M$, the hyperparameters, $\Phi = \{\sigma_\epsilon^{-2}, \sigma_\varepsilon^{-2}, \Sigma_{\mathbf{e}}^{-1}\}$, and the quality and mode indicator variables, $\Lambda = \{\mathbf{q}_{1:N}, \mathbf{i}_{1:N}\}$. From a Bayesian modeling point of view, the joint probability density function $p(\tilde{\mathcal{D}}, \Theta, \Phi, \Lambda|\mathcal{D})$ should be optimized. However, evaluating such posterior density functions requires a complex non-linear optimization problem to be solved. To circumvent the difficulties associated with the direct maximization of this joint probability density function, the identification problem is formulated under a layered optimization framework, as we will show in the following.

First, the chain rule of probability theory is used to factorize $p(\tilde{\mathcal{D}}, \Theta, \Phi, \Lambda|\mathcal{D})$:

$$p(\tilde{\mathcal{D}}, \Theta, \Phi, \Lambda|\mathcal{D}) = p(\tilde{\mathcal{D}}|\Theta, \Phi, \Lambda, \mathcal{D}) p(\Theta|\Phi, \Lambda, \mathcal{D}) p(\Phi|\Lambda, \mathcal{D}) p(\Lambda|\mathcal{D}) \tag{8.5}$$

Next, the layered optimization problem is formulated as follows:

1. Inference of noise-free data by maximizing the following posterior PDF

$$p(\tilde{\mathcal{D}}|\Theta, \Phi, \Lambda, \mathcal{D}) = \frac{p(\mathcal{D}|\tilde{\mathcal{D}}, \Theta, \Phi, \Lambda) p(\tilde{\mathcal{D}}|\Theta, \Phi, \Lambda)}{p(\mathcal{D}|\Theta, \Phi, \Lambda)} \tag{8.6}$$

2. Inference of model parameters by maximizing the following posterior PDF

$$p(\Theta|\Phi,\Lambda,\mathcal{D}) = \frac{p(\mathcal{D}|\Theta,\Phi,\Lambda)p(\Theta|\Phi,\Lambda)}{p(\mathcal{D}|\Phi,\Lambda)} \tag{8.7}$$

3. Inference of hyperparameters by maximizing the following posterior PDF

$$p(\Phi|\Lambda,\mathcal{D}) = \frac{p(\mathcal{D}|\Phi,\Lambda)p(\Phi|\Lambda)}{p(\mathcal{D}|\Lambda)} \tag{8.8}$$

4. Inference of indicator variables by maximizing the following posterior PDF

$$p(\Lambda|\mathcal{D}\big) = \frac{p(\mathcal{D}|\Lambda)p(\Lambda)}{p(\mathcal{D})} \tag{8.9}$$

In this Bayesian formulation, the likelihood function at a particular level corresponds to the evidence function at the previous level. Through this pattern, the optimization variables are gradually integrated out at different levels of Bayesian inference. Consequently, the optimal solutions obtained in subsequent layers of optimization are coordinated. In order to obtain a tractable explicit solution to the above layered optimization problem, a hierarchical Bayesian approach can be adopted through which the posterior PDFs are sequentially approximated in each layer and the procedure is iterated.

The expressions for the likelihood function and prior PDF in each level of Bayesian inference are given below.

- **Prior PDF of noise-free data**:

$$\begin{aligned} p(\tilde{\mathcal{D}}|\Theta,\Phi,\Lambda)) &= p(\tilde{\mathcal{D}} \\ &= \prod_{t=1}^{N} p(\tilde{y}_t)p(\tilde{s}_t)p(\tilde{\mathbf{u}}_t) \end{aligned} \tag{8.10}$$

The prior PDF of $\tilde{y}_t$ is expressed as

$$p(\tilde{y}_t) = \sqrt{\frac{1}{2\pi\sigma_y^2}}\exp\left(-\frac{\big(\tilde{y}_t-\mu_y\big)^2}{2\sigma_y^2}\right) \tag{8.11}$$

where μ_y and σ_y respectively denote the expected value and standard deviation of the noise-free output variable.

The prior PDF of $\tilde{s}_t$ is expressed as

$$p(\tilde{s}_t) = \sqrt{\frac{1}{2\pi\sigma_s^2}} \exp\left(- \frac{\big(\tilde{s}_t - \mu_s\big)^2}{2\sigma_s^2} \right) \tag{8.12}$$

where μ_s and σ_s denote the expected value and standard deviation of the noise-free scheduling variable, respectively.

The prior PDF of $\tilde{\mathbf{u}}_t$ is expressed as

$$p(\tilde{\mathbf{u}}_t) = \sqrt{\frac{1}{2\pi\Sigma_{\mathbf{u}}}} \exp\left(- \frac{1}{2}\Big(\tilde{\mathbf{u}}_t - \mu_{\mathbf{u}}\Big)^T \Sigma_{\mathbf{u}}^{-1} \Big(\tilde{\mathbf{u}}_t - \mu_{\mathbf{u}}\Big) \right) \tag{8.13}$$

where $\mu_{\mathbf{u}}$ and $\Sigma_{\mathbf{u}}$ respectively denote the vector of expected values and covariance matrix of the noise-free input variables.

- **Likelihood of noise-free data**:

$$\begin{aligned} p(\mathcal{D}|\tilde{\mathcal{D}}, \Theta, \Phi, \Lambda) &= \prod_{t=1}^{N} p(y_t, \mathbf{u}_t, s_t|\tilde{y}_t, \tilde{\mathbf{u}}_t, \tilde{s}_t, \Theta, \Phi, \Lambda) \\ &= \prod_{t=1}^{N} p(y_t|\tilde{\mathbf{u}}_t, \tilde{s}_t, \Theta, \sigma_\varepsilon^{-2}, q_t) p(\mathbf{u}_t|\Sigma_{\mathbf{e}}^{-1}, \tilde{\mathbf{u}}_t) p(s_t|\sigma_\epsilon^{-2}, \tilde{s}_t) \\ &= \prod_{t=1}^{N} p(\varepsilon_t|\Theta, \sigma_\varepsilon^{-2}, q_t, \tilde{s}_t) p(\mathbf{e}_t|\Sigma_{\mathbf{e}}^{-1}) p(\epsilon_t|\sigma_\epsilon^{-2}) \end{aligned} \tag{8.14}$$

The likelihood of y_t is expressed as

$$p(\varepsilon_t|\Theta, \sigma_\varepsilon^{-2}, q_t, i_t) = \sqrt{\frac{q_t}{2\pi\sigma_\varepsilon^2}} \exp\left(- \frac{q_t\big(y_t - \sum_{m=1}^{M} \psi_t^{(m)}(\tilde{s}_t)\tilde{y}_t^{(m)}\big)^2}{2\sigma_\varepsilon^2} \right) \tag{8.15}$$

if the output data is contaminated with scale outliers and

$$p(\varepsilon_t|\Theta, \sigma_\varepsilon^{-2}, q_t, i_t) = \sqrt{\frac{1}{2\pi\sigma_\varepsilon^2}} \exp\left(- \frac{\big(y_t - \sum_{m=1}^{M} \psi_t^{(m)}(\tilde{s}_t)\tilde{y}_t^{(m)} - q_t\big)^2}{2\sigma_\varepsilon^2} \right) \tag{8.16}$$

if the output data is contaminated with location outliers. See Sections 5.6.1 and 6.3 for more details.

The likelihood of $\mathbf{u}_t$ is expressed as

$$\begin{aligned} p(\mathbf{e}_t|\Sigma_{\mathbf{e}}^{-1}) &= \prod_{j=1}^{P} p\big(e_{j,t}|\sigma_{e_j}^{-2}\big) \\ &= \prod_{j=1}^{P} \sqrt{\frac{2}{2\pi\sigma_{e_j}^2}} \exp\left(- \frac{\big(u_{j,t} - \tilde{u}_{j,t}\big)^2}{2\sigma_{e_j}^2} \right) \end{aligned} \tag{8.17}$$

The likelihood of s_t is expressed as

$$p\big(\epsilon_t|\sigma_\epsilon\big) = \sqrt{\frac{2}{2\pi\sigma_\varepsilon^2}} \exp\left(- \frac{\big(s_t - \tilde{s}_t\big)^2}{2\sigma_\varepsilon^2} \right) \tag{8.18}$$

See Section 5.6.2 for more details.

- **Prior PDF of model parameters**:

$$\begin{aligned} p(\Theta|\Phi,\Lambda) &= p\big(\{\Theta^{(m)}\}_{i=1}^{m}\big) \\ &= \prod_{m=1}^{M} p\big(\Theta^{(m)}\big) \\ &\propto \prod_{m=1}^{M} \exp\left(- \frac{1}{2}\Big(\Theta^{(m)} - \Theta_0^{(m)}\Big)^T \Sigma_{\Theta_0^{(m)}}^{-1} \Big(\Theta^{(m)} - \Theta_0^{(m)}\Big)\right) \end{aligned} \tag{8.19}$$

where $\Theta_0^{(m)}$ and $\Sigma_{\Theta_0^{(m)}}$ denote the explicitly specified expected values and covariance matrix of $\Theta^{(m)}$. See Sections 5.6.1 and 6.3 for more details.

- **Likelihood of model parameters**:

$$p(\mathcal{D}|\Theta,\Phi,\Lambda) = \prod_{t=1}^{N} p(y_t, \mathbf{u}_t, s_t|\Theta,\Phi,\Lambda) \tag{8.20}$$

where

$$\begin{aligned} p(y_t, \mathbf{u}_t, s_t|\Theta,\Phi,\Lambda) = \int & p(y_t, \mathbf{u}_t, s_t|\tilde{y}_t, \tilde{\mathbf{u}}_t, \tilde{s}_t, \Theta,\Phi,\Lambda) \\ & \times p(\tilde{y}_t, \tilde{\mathbf{u}}_t, \tilde{s}_t|\Theta,\Phi,\Lambda)\, d\tilde{y}_t\, d\tilde{\mathbf{u}}_t\, d\tilde{s}_t \end{aligned} \tag{8.21}$$

The integral in Equation 8.21 often lacks a closed-form expression. Several methods are available for approximating the integrations in complex models, such

as Monte Carlo sampling-based techniques, Laplace's method, and variational Bayes approaches. The interested readers are referred to MacKay (2002) for more information.

- **Prior PDF of hyperparameters**:

$$\begin{aligned} p(\Phi|\Lambda) &= p\big(\sigma_\varepsilon^{-2}\big)p\big(\sigma_\epsilon^{-2}\big)\prod_{j=1}^{P} p\big(\sigma_{e_j}^{-2}\big) \\ &\propto \sigma_\varepsilon^{2-2k_\varepsilon}\exp\bigg(-\frac{s_\varepsilon}{\sigma_\varepsilon^2}\bigg)\sigma_\epsilon^{2-2k_\epsilon}\exp\bigg(-\frac{s_\epsilon}{\sigma_\epsilon^2}\bigg) \\ &\times \prod_{j=1}^{P}\sigma_{e_j}^{2-2k_j}\exp\bigg(-\frac{s_j}{\sigma_{e_j}^2}\bigg) \end{aligned} \tag{8.22}$$

See Section 5.6.2 for more details.

- **Likelihood of hyperparameters**:

$$\begin{aligned} p(\mathcal{D}|\Phi,\Lambda) &= \int p(\mathcal{D}|\Theta,\Phi,\Lambda)p(\Theta|\Phi,\Lambda)d\Theta \\ &\approx p(\mathcal{D}|\Theta^{\text{MP}},\Phi,\Lambda)p(\Theta^{\text{MP}}|\Phi,\Lambda)2\pi\sqrt{\det\mathbf{B}^{-1}} \end{aligned} \tag{8.23}$$

where $\mathbf{B} = -\nabla\nabla\log p(\Theta|\Phi,\Lambda,\mathcal{D})$. There are two main assumptions underlying the evidence approximation in Equation 8.23: 1. The data is not grossly at variance with the likelihood and the prior. 2. The number of well-determined parameters are large. See Section 5.6.1 for more details.

- **Prior PDF of quality and mode indicator variables**:

$$\begin{aligned} p(\Lambda) &= p(\mathbf{q}_{1:N},\mathbf{i}_{1:N}) \\ &= \prod_{t=1}^{N} p(q_t)p(i_t) \end{aligned} \tag{8.24}$$

The prior probability of q_t is expressed as

$$p(q_t) = \delta^{\Big(1-\frac{q_t-\rho}{1-q_t\rho}\Big)}(1-\delta)^{\Big(\frac{q_t-\rho}{1-q_t\rho}\Big)} \tag{8.25}$$

if the output data is contaminated with scale outliers and

$$p(q_t) = \prod_{t=1}^{N} (0.5\delta)^{\left(\frac{|q_t|}{\Delta}\right)} (1-\delta)^{\left(1-\frac{|q_t|}{\Delta}\right)} \tag{8.26}$$

if the output data is contaminated with location outliers. See Sections 5.4 and 5.6.3 for more details.

The prior probability of i_t is expressed as

$$p(i_t) = \prod_{m=1}^{M} p(i_t = m)^{[i_t=m]} \tag{8.27}$$

See Section 6.3 for more details.

- **Likelihood of quality and mode indicator variables**:

$$\begin{aligned} p(\mathcal{D}|\Lambda) &= \int p(\mathcal{D}|\Phi,\Lambda) p(\Phi|\Lambda) d\Phi \\ &\approx p(\mathcal{D}|\Phi^{\text{MP}},\Lambda) p(\Phi^{\text{MP}}|\Lambda) 2\pi \sqrt{\det \mathbf{C}^{-1}} \end{aligned} \tag{8.28}$$

where $\mathbf{C} = -\nabla\nabla \log p(\Phi|\Lambda,\mathcal{D})$. The above approximation holds good if the posterior distribution over hyperparameters is sharply peaked around Φ^{MP}. See Section 5.6.2 for more details.

8.2 Concluding Remarks

In many industrial applications, real-time analysis of key performance indicators constitutes an essential prerequisite for advanced monitoring and control of industrial processes. However, on-line measurement of process quality variables is often restricted by inadequacy of measurement techniques, low reliability of measuring devices, and significant time-delays associated with laboratory sample analysis. In industrial processing plants, such limitations can have a severe influence on the quality of products, production of waste, and safety of operations. These concerns motivate the development of theoretical and practical methods of **inferential sensing**, also called **soft sensing**, to provide frequent

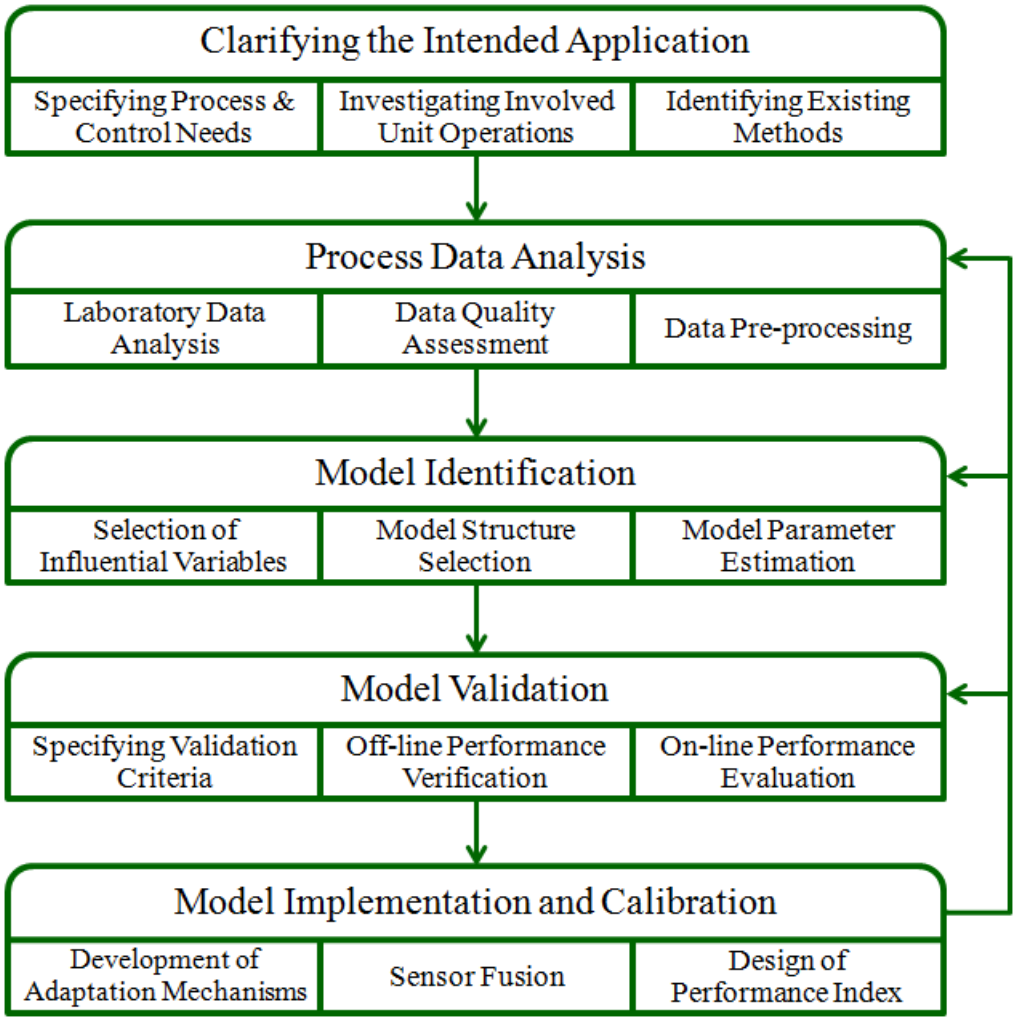


Figure 8.1: Flowchart of the inferential sensor design procedure

on-line estimates of quality variables on the basis of their correlation with real-time process measurements (Fortuna *et al.*, 2007; Kadlec *et al.*, 2009). Figure 8.1 presents a flowchart of the inferential sensor design procedure (Khatibisepehr *et al.*, 2013).

Development and implementation of inferential sensors entail many challenges that are often addressed in a rather *ad hoc* manner (Hangos and Cameron, 2001; Paoletti *et al.*, 2007; Kadlec *et al.*, 2009; Pani and Mohanta, 2011; Kano and Fujiwara, 2013). Despite the increasing number of publications dealing with industrial applications, several issues require further investigation. The main objective of this research was to develop advanced inferencing paradigms to provide rigorous and general solutions to certain outstanding

inferential sensing problems. The main contributions of this thesis, as explained in each of the earlier chapters, are summarized below.

Chapter 2 provided a general introduction to the main steps involved in development and implementation of industrial inferential sensors, presented an overview of the relevant Bayesian literature as well as a review of the industrial applications of Bayesian inferential sensors. Since the use of Bayesian techniques in industrial applications, in particular in design of inferential sensors for process industries, is relatively new, the potential Bayesian solutions to some of the main issues associated with inferential sensor design were also discussed. This chapter was not intended to provide a comprehensive review of the great variety of methods used in the design of inferential sensors, but was rather focused on the techniques that have their origin in Bayesian Statistics. Therefore, the main contribution of this chapter is complementing the existing reviews in the field.

Chapter 3 provided a classical non-Bayesian framework for real-time inferential modeling of complex processes. Given a query point, a search algorithm was applied to select spatial and temporal nearest neighbors within the identification data-set. The selected sub-set of identification data was then used to identify a local ν-SVR model, which could effectively handle small identification data-sets. The proposed just-in-time/space modeling techniques can cope with variations in process characteristics and handle non-linearity of underlying mechanisms. The method was implemented to facilitate real-time modeling and prediction of cytotoxicity effects on living cells induced by certain water contaminants. The developed framework enabled us to analyze intrinsic cell behavior and predict the trajectory of its progress (growth or death) over a considerable time horizon.

Chapter 4 presented a novel Bayesian framework for real-time model structure selection and similarity function parameterization in just-in-time/space modeling methods. The locally weighted partial least squares (LW-PLS) algorithm was adopted as the main modeling technique. A Bayesian procedure was developed to partition the operating space

into a finite number of sub-spaces and characterize them during the off-line identification phase. The problem of finding the locally optimal LW-PLS model structure and similarity function hyperparameters was formulated and solved under an iterative hierarchical Bayesian optimization framework for each sub-space. Two industrial case-studies were considered to demonstrate the effectiveness of the proposed method: 1. real-time prediction of Reid vapor pressure of gasoline in a petrochemical refinery, and 2. real-time prediction of the active substance content of a pharmaceutical tablet. The method was successfully applied to identify inferential LW-PLS models for real-time prediction of these quality variables using near-infrared (NIR) transmittance spectra.

Chapter 5 proposed a novel unified Bayesian framework for robust identification of inferential models in the presence of outliers. First, the most common contaminating distributions and outlier models were introduced. Next, a unified objective function was proposed and a layered optimization strategy was implemented. The solutions obtained in subsequent layers of optimization were coordinated within an iterative hierarchical Bayesian framework. The proposed optimization strategy not only yields maximum *a posteriori* estimates of model parameters, but also provides an automated mechanism for determining the hyper-parameters and investigating the quality of each observation. Using a simulated continuous fermentation reactor, it was shown that the proposed robust Bayesian framework outperforms the traditional robust regression techniques in terms of the accuracy of model parameters and noise variance estimates. The robustness of the method was further demonstrated using a pilot-scale continuous stirred tank heater.

In Chapter 6, the problem of identification of multi-modal systems switching among non-linear continuous-state dynamics was investigated. A novel Bayesian procedure for the development of multi-model inferential sensors was developed to meet the specific requirements of the process industries. The importance of adaptation mechanisms for maintaining the on-line performance of inferential sensors was discussed. A Bayesian

decision-support scheme for real-time implementation of the multi-model inferential sensors was presented. The developed scheme includes a global adaptation mechanism, within the envelope of previously identified operating conditions. The implementation of the proposed procedures was demonstrated through a simulation case study. An adaptive multi-model inferential sensor was developed to predict the product concentration of a simulated continuous stirred tank reactor. The effectiveness of the method was further highlighted through a successful industrial application of an adaptive multi-model inferential sensor designed for real-time monitoring of a key quality variable in an oil sands processing unit.

Chapter 7 presented a novel data-driven Bayesian approach for real-time performance assessment of inferential sensors. The main contribution of this chapter is to rigorously formulate the problem of reliability analysis of real-time predictions under a Bayesian framework utilizing both discrete and continuous operating statuses. The main focus was to characterize the effect of the operating space on the prediction accuracy in the absence of target measurements. The real-time performance assessment of multi-model inferential sensors was also discussed. A simulated continuous fermentation reactor was used as a working example to outline the ideas throughout this chapter. The proposed Bayesian approach was successfully applied for performance assessment of two industrial inferential sensors. These inferential sensors had been designed to provide real-times predictions of the two quality variables of an oil sands processing unit.

The main problems involved in inferential sensing were outlined and addressed through a unified Bayesian framework in this thesis. The thesis can be used as a guide to Bayesian inferential sensing practice in process industries.

8.3 Future Research

Future research in this area can be taken into multiple different directions. Some of the challenging issues that foreshadow interesting topics for future research are summarized below.

1. Although the problems of process data analysis and model identification are interconnected, most of the existing solutions are disconnected. It is desired to seek for a unified framework that simultaneously considers different aspects of data analysis and inferential modeling. As shown in Chapter 5, there is a potential in formulating the problems of interest as rigorous conditional probabilistic problems within a Bayesian framework. To derive analytical expressions for all levels of inference, the popular autoregressive with exogenous input (ARX) model was used to illustrate the design of a robust unified Bayesian framework. However, the application of the ideas presented in Chapter 5 is not limited to ARX models. The derivations can be directly extended to other classes of dynamic models, though numerical optimization may be required.

2. In order to maintain the reliability of an inferential sensor, it is required to track its on-line performance. However, the main body of research in this area has been focused on exploiting advanced strategies for development of inferential sensors. Hence, it is of paramount importance to search for general criteria and techniques for on-line performance assessment of inferential models. Model uncertainty (plausible alternative model structures/parameters) and input uncertainty (plausible alternative input values) are the major sources of prediction uncertainty. In Chapter 7, the effect of model uncertainty on the prediction performance of the inferential sensors was explored. Further investigation is required to capture the conditional dependence of the reliability of inferential sensor predictions on the uncertainty of

input measurements.

3. Maintenance of inferential sensors is another important topic for future research. There have been several efforts to develop real-time and recursive identification methods as well as local adaptation mechanisms. Yet, proper maintenance of the identification data-set remains a challenging task. Theoretical and practical developments are required to effectively assess the reliability of operational and laboratory measurements in real-time.

4. There is a growing realization that off-line operation assistance tools can play a significant role in improving plant-wide operations. The main research challenge is to develop information synthesis schemes that can coordinate processing of diverse forms of knowledge. Further research is imperative to effectively synthesize qualitative and quantitative information provided by operations personnel, inferential and physical sensors, laboratory analysis, and many other sources.

5. Long and uncertain time-delays in reference data (*e.g.* lab data) constitute one of the main practical problems in inferential sensor development. Samples are frequently collected from the operational field and the recorded sampling time can deviate significantly from the actual time. Laboratory analysis of each sample can be time consuming, thereby introducing a significant time-delay. Therefore, modeling, filtering, and information synthesis in the presence of long and uncertain time-delays are of great research interest in inferential sensor development.

6. Bias update has been common practice in inferential sensor applications. Advanced updating strategies, as reviewed in Chapter 2, include the multi-rate information fusion method and the filtering method. However, due to the slow rate of sampling of laboratory data, these updates are often associated with an abrupt change of the prediction, introducing undesired bumps to the inferential sensor predictions.

Optimal synthesis of multi-rate data is another topic of interest.

Bibliography

Fortuna, L., S. Graziani, A. Rizzo and M. G. Xibilia (2007). *Soft Sensors for Monitoring and Control of Industrial Processes*. first ed.. Springer-Verlag. London, UK.

Hangos, K. M. and I.T. Cameron (2001). *Process Modelling and Model Analysis*. first ed.. Academic Press. San Diego, USA.

Kadlec, P., B. Gabrys and S. Strandt (2009). Data-driven soft sensors in the process industry. *Computers and Chemical Engineering* **33**(4), 795–814.

Kano, M. and K. Fujiwara (2013). Virtual sensing technology in process industries: Trends and challenges revealed by recent industrial applications. *Journal of Chemical Engineering of Japan* **46**(1), 1–17.

Khatibisepehr, S., B. Huang and S. Khare (2013). Design of inferential sensors in the process industry: A review of bayesian methods. *Journal of Process Control* p. in press.

MacKay, D. J. C. (2002). *Information Theory, Inference, and Learning Algorithm*. first ed.. Cambridge University Press. New York, USA.

Pani, A. K. and H. K. Mohanta (2011). A survey of data treatment techniques for soft sensor design. *Chemical Product and Process Modeling*.

Paoletti, S., A. L. Juloski, G. Ferrari-Trecate and R. Vidal (2007). Identification of hybrid systems: A tutorial. *European Journal of Control* **13**(2-3), 242–260.

Appendix A

Guide to Soft Sensor Design Procedure

Project:

Key Investigator:

Industrial Contacts:

Problem Statement

1. Define the problem to be solved in this project.
2. Clarify the intended application of the inferential sensor to be designed.
3. Identify the existing approaches and pinpoint their shortcomings.
4. Specify the control and process needs.
5. Determine the potential benefits of successfully accomplishing the stated objectives. Calculate the dollar value of the benefits if possible.

Process Description

1. Prepare a schematic diagram of the process under investigation.
2. Investigate the involved unit operations.

3. Analyse the underlying chemical and physical phenomena.

4. Choose the influential process variables based on first principles analysis.

Process Data Analysis

1. Evaluate adequacy of the available historical data for accurate identification and reliable validation of the query inferential sensor. Divide the collected data into different subsets for identification, tuning, and validation purposes.

2. Assess the laboratory data in order to ensure the adequacy of quality and variability.

3. Find out the sampling frequency as well as the method (*e.i.* snapshot or composite) used for collecting laboratory samples.

4. Collect information about the followed procedures, measuring devices used, and time required to conduct laboratory analysis. This information would help to evaluate the extent of reliability of laboratory measurements.

5. Evaluate the performance of the field instruments and measuring devices.

6. Assess the accuracy, reliability, completeness, and representativeness of the collected operational and laboratory data. Identify the required data pre-processing techniques.

7. Perform data pre-processing in order to improve the quality of the collected historical data.

8. Choose the influential process variables based on statistical data analysis. Check consistency with the prior process knowledge.

9. Identify the possible frequent and infrequent operating conditions.

10. Prepare a table summarizing the description, operating range, sampling frequency, and measurement unit of each of the selected influential variables.

Remark. *It is critical to conduct interviews with the plant experts and operators in order to fully exploit the wealth of historical data and completely understand the system under investigation.*

Inferential Model Identification

1. Determine the key considerations in designing the query inferential sensor.

2. Choose appropriate system boundaries for derivation of mass, momentum, and energy balance equations.

3. Perform quantitative and/or qualitative first principles analysis.

4. State and justify the simplifying assumptions made for first principles analysis.

5. Select a model structure that would sufficiently approximate the behavior of the system under investigation. One may need to decide on the following features: 1. Linear or non-linear structure, 2. Static or dynamic relations, 3. Single or multiple models, 4. State-space or input-output representation.

6. Identify some of the suitable model identification approaches or algorithms that could be used. State the main strengths, required assumptions, and major limitations of each approach.

7. Select the identification approach that best meets the operational requirements while handling the modeling challenges. In the absence of any prior process knowledge, the main criteria to be considered in model selection are simplicity, generality, and flexibility.

Model Validation

1. Evaluate the size of the validation data-set and choose a reliable validation procedure accordingly.
2. Specify the main criteria to be considered for inferential model validation.
3. Define a set of descriptive performance measures to be evaluated.
4. If the off-line performance validation is satisfactory, develop an implementation framework for on-line performance verification.
5. Design an user-friendly interface (*e.g.* excel spreadsheet) to facilitate performance monitoring while the inferential sensor is tested on-line.

Model Implementation and Calibration

1. Develop proper procedures to evaluate and refine the quality of input measurements during the real-time implementation.
2. Develop a procedure for real-time performance assessment of inferential sensors.
3. Design an appropriate performance index indicating the reliability of predictions in real-time.
4. Develop a procedure for evaluating the reliability of laboratory measurements in real-time.
5. Develop a framework to synthesize the multiple sources of information and redundant measurements.
6. Develop on-line calibration procedures to detect and handle potential unknown drifts of the process operating conditions.

Appendix B

Comments on Bayesian Software Packages

This section comments on particular features of some of the major Bayesian software packages with which we have had experience through our research and application works. These include Netica, Bayes Net Toolbox (BNT), and WinBUGS. Although these packages have a long list of features, we endeavour to point out particular features that relate to the design of inferential sensors.

B.1 Netica

Netica (available at `http://www.norsys.com/`) is a user-friendly Bayesian tool that has been commercially available since 1995. It can be used to build and learn Bayesian models, as well as perform different types of inference tasks. It is also capable of representing Dynamic Bayesian models. Netica can learn probabilistic relations from data through the application of Spiegelhalter & Lauritzen parametrization, EM, or gradient descent algorithms; missing values are automatically handled. The relationships between variables may be entered as individual probabilities or in the form of equations. Both parameter and structure learning are supported by this package. Netica discretizes the continuous variables by partitioning their domain into some finite number of sub-sets. Since representation of the resulting discretized model is exponential in the number of

variables, this approach becomes problematic in large or complex models. For example, the number of parameters representing a Gaussian distribution over N variables is $O(N^2)$. If these variables are discretized into m ranges, then $O(m^N)$ elements are required to be learnt and stored. Exact general probabilistic inference performed by Netica is based on message passing in a junction tree of cliques, which is the fastest available algorithm. Once an inferential model is identified, we can answer queries or find optimal decisions. Given a case of new observations, both posterior probability of queries and most probable explanation (MPE) can be found. Netica allows the user to enter and update only individual cases; it does not handle sets of cases.

In summary, Netica is suitable for application in the following areas: diagnosis, prediction, decision analysis, sensor fusion, expert system building, probabilistic modeling, and certain kinds of statistical analysis.

B.2 Bayes Net Toolbox

Bayes Net Toolbox (BNT) (available at `http://bnt.googlecode.com/`) is another Bayesian modeling and inference package. Taking advantages of MATLAB features, BNT has become a widely used and powerful Bayesian software since 2002. BNT suffers from the lack of GUI, which is currently made up by MATLAB visualization tools; a preliminary attempt to make a GUI has been done by Murphy. The software can build and learn static and dynamic Bayesian networks, answer queries, or find optimal solutions using its powerful inference engine. BNT does not allow the entry of probabilistic relations by equation. BNT supports both parameter learning and structure learning by several learning algorithms such as EM and MCMC algorithms. BNT deals with continuous variables directly without attempting to discretize them. It allows only linear relations between the continuous variables and does not allow discrete nodes to have continuous parents. In addition, non-Gaussian probability distributions of continuous variables are not supported.

Inference tasks in static and dynamic Bayesian networks are performed by various exact and approximate inference algorithms.

Finally, BNT is applicable for implementation of the following probabilistic models: linear regression, logistic regression, mixtures of Gaussian distributions, DBNs (such as hidden Markov models (HMMs), Kalman filters, switching Kalman filters, and ARMAX models), factor analysis, probabilistic PCA, and many others.

B.3 WinBUGS

WinBUGS (available at `http://www.mrc-bsu.cam.ac.uk/bugs/`) is the most advanced version of BUGS (Bayesian Inference Using Gibbs Sampling) that provides Bayesian analysis of statistical models using Markov Chain Monte Carlo (MCMC) methods. Since MCMC is inherently less robust to the prior information than analytic statistical methods, prior knowledge plays an important role in the accuracy of a Bayesian model identified by WinBUGS. A wide range of non-Gaussian probability distributions for discrete and continuous variables are provided. WinBUGS allows the entry of probabilistic relations by equation, and supports non-linear relations between the continuous variables. However, high correlation among parameters may lead to slow convergence. Therefore, this program is inefficient for time series structures such as hidden Markov model (HMM). A new prediction can be obtained by specifying the query variable as missing in the data-set and assigning it a uniform prior.

WinBUGS is suitable for identification and implementation of generalized linear mixed models, latent variable models, and measurement error models.