Just-in-time and Adaptive Methods for Soft Sensor Design

by

Ming Ma

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

Master of Science

in

Process Control

Department of Chemical and Materials Engineering
University of Alberta

Abstract

In many industrial processes, critical variables cannot be easily measured on-line: they are either obtained from hardware analyzers which are often expensive and difficult to maintain, or carried out off-line through laboratory analysis which cannot be used in real time control. These considerations motivate the design of inferential sensors or so-called soft sensors to infer process quality variables in real time from on-line process measurements. Numerous modeling techniques have been proposed and successfully applied to soft sensors for many industrial processes. Despite the popularity of these techniques in industry, development and implementation of soft sensors are still challenging due to complexity of industrial processes. The main contribution of this thesis is the development of several soft sensing methods that can achieve and maintain satisfactory performance while handling multi-mode, nonlinear and time-varying problems.

Real time identification of local process model, also known as Just-in-time (JIT) modeling, is a special modeling technique for design of infinite-mode soft sensors. It is widely used in dealing with nonlinear and multi-mode of industrial processes. The performance of JIT model depends on parameters of the similarity function as well as the structure and parameters of the local model. A Bayesian framework is proposed to provide a systematic method for real time parameterization of the similarity function, selection of the local model structure, and estimation of the corresponding model parameters in JIT modeling methods. Another challenging issue in JIT modeling is the selection of most relevant samples from database by considering input-output information. Thus, a new input-output similarity function is defined and integrated into a Bayesian framework for JIT modeling.

To cope with time-varying behaviour of processes, on-line adaptation is usually integrated in the implementation procedure. Although there are a number of publica-

tions dealing with adaptation of soft sensors, few of them have considered the adaptation of nonlinear grey-box models which are popular in process industry. Thus, a new adaptation mechanism for nonlinear grey-box models is proposed based on recursive prediction error method (RPEM). Adaptive data preprocessing and cautious update strategy are integrated to ensure robustness and effectiveness of the adaptation.

The effectiveness and practicality of the proposed methods are verified using data from industrial processes. Some of the proposed methods have also been implemented for industrial applications.

Preface

Chapter 2 of this thesis has been published as M. Ma, S. Khatibisepehr and B. Huang, *A Bayesian Framework for Real-time Identification of Locally Weighted Partial Least Squares*, AIChE Journal (2014). I was responsible for the algorithm development, case studies and manuscript composition. Dr. Shima Khatibisepehr assisted algorithm development and contributed to manuscript composition. Dr. Biao Huang was the supervisory author and was involved with manuscript composition.

Acknowledgements

It would not have been possible for me to finish writing this thesis without help and support of the kind people around me. I owe my gratitude to all those people who have contributed to this thesis and because of whom my graduate experience has been one that I will cherish forever.

Above all, I would like to thank my supervisor Dr. Biao Huang for his constant guidance and patience that supported me at all time. His insightful comments and constructive criticisms helped me penetrate the problems and form my own ideas. His constant encouragement determined me to carry on when I encountered difficulties. His rigorous attitude towards academic work inspired me. He always spent large amounts of time meticulously reviewing and correcting my writings. I am also thankful to him for giving me opportunities to realize my ideas as well as gain valuable experiences in multiple industrial projects. My special thanks will also go to Dr. Shima Khatibisepehr who has supported me in both my research and projects. Without Shima's careful instructions and helpful advice, I could hardly complete this thesis.

I would like to thank my colleagues in Computer Process Control Group from whom I gained a lot of help: Ruben Gonzales, Swanand Khare, Kangkang Zhang, Tianbo Liu, Yaojie Lu, Hao Chen, Anahita Sadeghian, Nima Sammaknejad, Alireza Fatehi. I am also indebted to other friends and colleagues who have supported me and shared joy with me in the past two years: Shunyi Zhao, Xianqiang Yang, Liu Liu, Xiaodong Xu, Ouyang Wu, Ruomu Tan, Jing Zhang, Bing Xia, Yang Wang, Yang Zhou, Tong Chen, Yuyu Yao, Cong Jing.

I would like to acknowledge the Department of Chemical and Materials Engineering, University of Alberta, for offering me the opportunity to pursue my Master's degree. I also gratefully acknowledge the Natural Sciences and Engineering Research

Council of Canada and Alberta Innovates Technology Futures for the financial support.

Last but not least, I owe a debt of eternal gratitude to my parents for their unconditional love and support. Without their understanding and encouragement, I couldn't be where I am today.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Thesis Outline and Contributions 2

2 A Bayesian Framework for Real-time Identification of Locally Weighted Partial Least Squares **4**

2.1 Introduction 4

2.2 Problem Statement 7

2.3 Hierarchical Bayesian Optimization Framework 9

2.3.1 Inference of Model Parameters 10

2.3.2 Inference of Localization Parameter 14

2.3.3 Inference of Model Structure 15

2.4 Hierarchical Bayesian Optimization Procedure 17

2.5 Case Studies 18

2.5.1 Reid Vapor Pressure of Gasoline 18

2.5.2 Protein Content of Wheat Kernels 25

2.6 Conclusion 29

3 Bayesian Just-in-time Modeling with the Input-output Similarity Function **30**

3.1 Introduction 30

3.2 Problem Statement 32

3.3 Bayesian JIT Modeling with Input-output Similarity 34

3.3.1 Similarity Function 34

3.3.2 Bayesian JIT Modeling 35

3.4 Implementation Procedure 41
3.5 Case Study 41
3.6 Conclusion 47

4 Recursive Prediction Error Method and Its Application in Adaptive Soft Sensors Design 48

4.1 Introduction 48
4.2 Theoretical Background 51
4.2.1 Prediction Error Method 51
4.2.2 Recursive Prediction Error Method 52
4.2.3 Evaluation Criteria 52
4.3 Adaptation Mechanism 53
4.3.1 Adaptive Data Preprocessing 53
4.3.2 Reliability of Lab Measurements 55
4.3.3 Cautious Update 56
4.4 Adaptive Soft Sensor for Naphtha:Bitumen Ratio in Inclined Plate Settler 59
4.4.1 Process Description 59
4.4.2 Soft Sensor Design 61
4.4.3 Off-line Evalutaion 62
4.4.4 On-line Evaluation 65
4.5 Conclusion 71

5 Conclusions 72

5.1 Summary of Thesis 72
5.2 Future Work 74

Bibliography 76

List of Tables

2.1 Comparing prediction performance of the $1st$ layer of Bayesian LW-PLS and regular LW-PLS using data-set from reid vapor pressure of gasoline, scenario I 20

2.2 Comparing prediction performance of the $1st$ and $2nd$ layer of Bayesian LW-PLS and regular LW-PLS using data-set from reid vapor pressure of gasoline, scenario II 22

2.3 Comparing prediction performance of Bayesian LW-PLS and RMSECV-based LW-PLS using data-set from reid vapor pressure of gasoline, scenario III 25

2.4 Comparing prediction performance of the Bayesian LW-PLS and RMSECV-based LW-PLS using data-set from wheat kernels 26

2.5 Comparing prediction performance of the Bayesian LW-PLS and RMSECV-based LW-PLS using data-set from wheat kernels, extrapolation case 27

3.1 Features of different data-driven modeling methods 32

3.2 Comparing prediction performance of different methods using pharmacy tablets data-set 43

3.3 Comparing prediction performance of different methods using pharmacy tablets data-set 45

4.1 Cautious criterion 56

4.2 A summary of the influential process variables, N:B soft sensor 61

4.3 Cross-validation results of N:B soft sensors 62

4.4 On-line evaluation results for N:B soft sensors 67

List of Figures

1.1 Overview of problems and solutions 2

2.1 Reid vapor pressure of gasoline spectra data. 19
2.2 Cross-validation using data-set from reid vapor pressure of gasoline, scenario I 20
2.3 Cross-validation using data-set from reid vapor pressure of gasoline, scenario II 21
2.4 Localization parameter φ of Bayesian LW-PLS 23
2.5 Cross-validation using data-set from reid vapor pressure of gasoline, scenario III 23
2.6 Distributions of selected inputs for reid vapor pressure of gasoline example 24
2.7 Cross-validation using data-set from protein content of wheat kernels 26
2.8 Cross-validation using data-set from protein content of wheat kernels, extrapolation case 27
2.9 Distributions of selected inputs for protein content of wheat kernels example 28

3.1 Cross-validation using data-set from pharmaceutical tablets, case I . 44
3.2 Cross-validation using data-set from pharmaceutical tablets, case II . 46

4.1 Recursive adaptation flowchat 50
4.2 Reliability of lab data 56
4.3 Recursive adaptation mechanism 58
4.4 Schematic diagram of the inclined plates settler (IPS) 60
4.5 N:B ratio measured from the lab analysis and on-line analyzer 60

4.6 Auto-validation IPSA (Jan-Apr 2013) 63
4.7 Auto-validation IPSB (Jan-Apr 2013) 64
4.8 Cross-validation IPSA (Apr-Jul 2013) 65
4.9 Cross-validation IPSB (Apr-Jul 2013) 66
4.10 Data access network 68
4.11 On-line evaluation IPSA (Sep-Dec 2013) 69
4.12 On-line evaluation IPSB (Sep-Dec 2013) 70

Chapter 1

Introduction

1.1 Motivation

Industrial processes are usually equipped with a large number of sensors for process monitoring and control. Some critical process variables cannot be easily measured because of inadequacy of measurement techniques or low reliability of measuring devices. These hard-to-measure variables are usually obtained from hardware analyzers (which are often expensive and need frequent and costly maintenance) or carried out off-line by laboratory analysis (which cannot be used for real-time applications). Approximately two decades ago, work was started by taking advantage of the easy-to-measure variables to build predictive models to predict the hard-to-measure variables. This type of predictive model can be used for development of a soft sensor.

Soft sensors can fulfill a broad range of tasks. The primary and most important application of soft sensors is on-line prediction. Its task is to provide real-time estimates of quality variables on the basis of real-time process measurements. These variables are usually used as indicators for process control and process operations, thus having significant effect on the process output quality. Once soft sensors can achieve stable and satisfactory performance, they can be further used to develop advanced control strategies, such as model predictive control. The other application of soft sensors is to monitor the process state, and thus detect and diagnose process abnormalities. This is referred as fault detection and diagnosis. For more applications of soft sensors, one can refer to [1], [2], [3].

The development and implementation of soft sensors in industry is challenging due to complexity of modern processes. Although numerous modeling techniques

have been proposed for soft sensors, there are a number of issues remaining to be investigated. The main contribution of this thesis is the development of several soft sensing methods to handle multi-model, nonlinear and time-varying behaviours of processes in order to achieve and maintain satisfactory soft-sensor performance.

1.2 Thesis Outline and Contributions

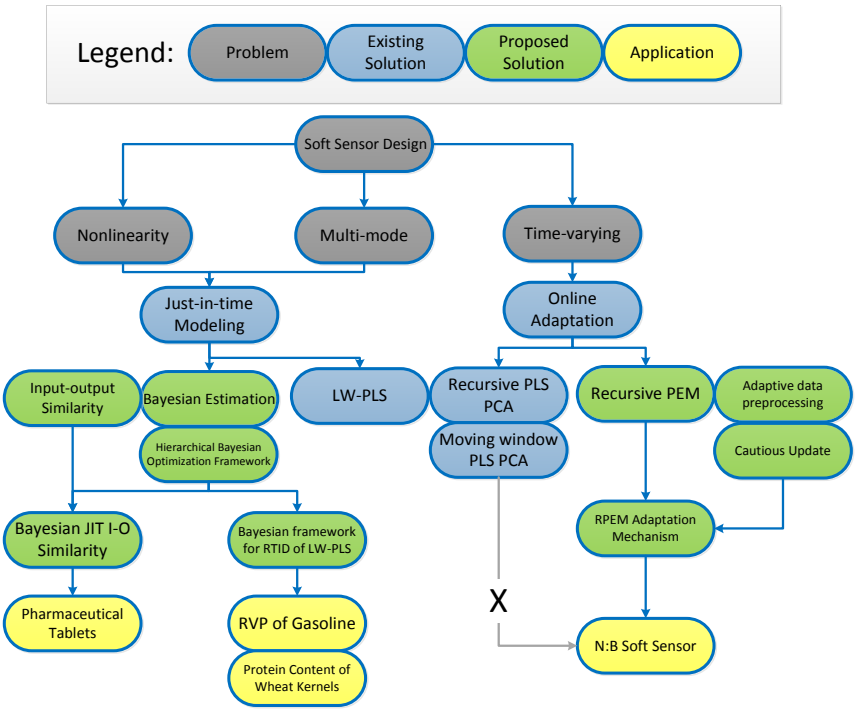


Figure 1.1: Overview of problems and solutions

The overview of problems and solutions are shown in Figure 1.1. The main contributions of this research are presented in three chapters, which are organized as follows:

In Chapter 2, a holistic Bayesian framework for the locally weighted partial least squares (LW-PLS) regression is proposed. The proposed method follows a Bayesian approach to estimate the model parameters of the LW-PLS model which makes it

have been proposed for soft sensors, there are a number of issues remaining to be investigated. The main contribution of this thesis is the development of several soft sensing methods to handle multi-model, nonlinear and time-varying behaviours of processes in order to achieve and maintain satisfactory soft-sensor performance.

1.2 Thesis Outline and Contributions

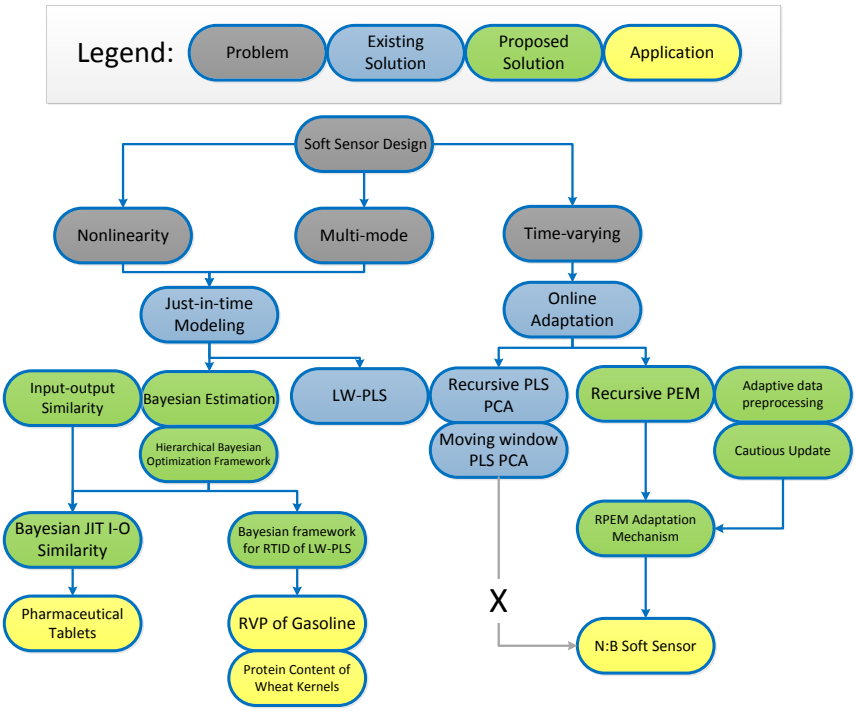


Figure 1.1: Overview of problems and solutions

The overview of problems and solutions are shown in Figure 1.1. The main contributions of this research are presented in three chapters, which are organized as follows:

In Chapter 2, a holistic Bayesian framework for the locally weighted partial least squares (LW-PLS) regression is proposed. The proposed method follows a Bayesian approach to estimate the model parameters of the LW-PLS model which makes it

possible to incorporate available prior knowledge into the identification procedure as well as take into account different magnitudes of measurement noise. Application of the hierarchical Bayesian optimization framework offers a systematic and tractable way to simultaneously obtain the optimal model structure, localization parameters and model parameters at each operating point. Moreover, the Bayesian model structure selection can automatically penalize model complexity, allowing us to avoid over-fitting. To evaluate the effectiveness of the proposed method, two industrial case studies are performed in which NIR spectra were used to provide real-time estimates of reid vapor pressure (RVP) and wheat kernels.

In Chapter 3, a novel Just-in-time modeling method is proposed based on the Bayesian framework in Chapter 2. First, a new input-output similarity function is defined to take both input and output information into account so that the noise in input data will have less negative impact and the information in output can be utilized more effectively. Furthermore, this new similarity function is integrated into a Bayesian framework which provides a systematic way to select the locally optimal model structure as well as estimate the model parameters. Bayes' theorem also makes it possible to incorporate available prior knowledge into the identification process. Various features of the proposed method are illustrated through a case study in pharmaceutical industry, where near infrared (NIR) spectra were used to provide real-time estimates of the content of active substance in tablets.

In Chapter 4, an adaptation mechanism for nonlinear grey-box model is explored based on the recursive prediction error method. Several adaptive data pre-processing methods are integrated to reduce the negative effects of noise in measurements as well as detect irregular measurements. The cautious update strategy is integrated into the adaptation mechanism to meet the need for robust adaptation, thus avoiding over-updating issues. Finally, the effectiveness of this method is demonstrated by a successful application in oil sands industry.

In Chapter 5, the main results of this thesis are summarized and opportunities for future work are discussed.

Chapter 2

A Bayesian Framework for Real-time Identification of Locally Weighted Partial Least Squares*

2.1 Introduction

Process modeling is one of the most important elements in development and implementation of advanced process monitoring and control techniques. The representativeness of process models has a significant effect on the performance of these techniques. Linear modeling techniques are commonly used to identify a model from the process variables. Ordinary least squares (OLS) regression is one of the most widely used classical modeling techniques due to its simplicity. The main assumption behind the OLS regression is that the process variables are not strongly dependent on each other. principal component regression (PCR) and partial least squares (PLS) regression have noticeable advantages over the OLS regression in dealing with the collinearity issue [4, 5, 6, 7]. The PCR first uses orthogonal transformation to convert correlated input variables into a set of uncorrelated, lower dimensional principal components. Next, the OLS is applied to reveal the parametric relationship between the principal components and the output variables. The orthogonal transformation used in the PCR only considers the relationships among input variables and fails to take into account any information about the output variables. Therefore, it may result in an ill-conditioned alignment [8]. The PLS regression overcomes this shortcoming

*This chapter is a revised version of an accepted paper "M. Ma, S. Khatibisepehr and and B. Huang, A Bayesian Framework for Real-time Identification of Locally Weighted Partial Least Squares. AIChE Journal."

by taking into account both input and output variables for finding the principal components [9]. The performance of these linear techniques will be satisfactory only if the underlying process can be assumed to be linear. To deal with processes which exhibit certain form of non-linear behaviour, several approaches have been proposed to integrate non-linear features with the linear PLS framework, thus resulting in non-linear PLS algorithms such as quadratic PLS [10], neural network PLS [11] and fuzzy PLS [12]. These approaches retain the linear latent structure of PLS model. In light of non-linear principal components, Malthouse [13] proposed a new approach named non-linear PLS (NLPLS) to extract the non-linear latent structures. However, these nonlinear PLS approaches which provide global models to describe the data from different operation modes may not achieve satisfactory performance. Considering these issues, the LW-PLS regression can be used [14]. LW-PLS combines the nature of locally weighted regression and PLS so that it can deal with the nonlinearity, multi-mode behaviour as well as the collinearity. In the locally weighted partial least squares (LW-PLS) method, local PLS models are built around each operating point through local calibration samples. In order to construct a LW-PLS model, the following aspects should be considered:

1. Selection of local calibration samples: Local calibration samples are often selected or prioritized using a certain similarity function. The similarity function takes into account the distance between a query sample and calibration ones. The similarity function is parameterized by a set of localization parameters which needs to be specified to control how steeply the similarity will decrease by increasing the distance. In this way, the localization parameters would greatly affect the selection or priorotization of local calibration samples.

2. Selection of model structure: After choosing or prioritizing proper local calibration samples, the next step is to choose a proper model structure. This could be equivalent to determining the dimensionality of the latent space that can best describe the underlying behaviour of the process.

3. Estimation of model parameters: Having selected the local calibration samples and determined the model structure, model parameters can be identified via the LW-PLS algorithm.

Therefore, the problem of identification of an LW-PLS model boils down to obtaining the optimal combination of localization parameters, model structure and model parameters. The common practice is to search for the globally optimal combination of localization parameters and model structure by minimizing the root mean square error of cross-validation (RMSECV) [15]. This approach is often computationally inefficient for on-line identification of the LW-PLS models and may also result in the over-fitting issue [16, 17]. Khatibisepehr et al. [18] have developed an off-line identification method to find locally optimal localization parameters and a model structure within a known operating space using a hierarchical Bayesian optimization framework. The idea behind this method is to first partition the operating space into a finite number of sub-spaces and then find the optimal combination of localization parameters and model structure for each sub-space. The application of Bayes' theorem makes it possible to incorporate the prior knowledge over the localization parameters and model structures. The proposed Bayesian framework can also deal with the model complexity control to avoid over-fitting. However, this method has the following shortcomings: 1. It does not utilize prior knowledge of the model parameters for modeling; 2. Like all the other existing methods, uncertainties in the parameter estimates are not taken into account in selection of the model structure and tuning of the localization parameters; 3. Due to the multi-mode behaviour of industrial processes, a finite number of sub-spaces may not cover the entire operating space especially over a long period.

Therefore, it is desired to tune the localization parameters, select the model structure, and estimate the model parameters all in a real-time phase. The main contribution of this chapter is to develop a novel integrated identification method to find locally optimal combination of model parameters, localization parameters and a model structure in a real-time manner to take full advantage of Bayesian methods. The real-time identification problem of interest is formulated under a holistic Bayesian framework consisting of consecutive levels of optimization. The resulting optimization problem is hierarchically decomposed and a layered optimization strategy is implemented. To obtain explicit solutions, an iterative hierarchical Bayesian approach is adopted to coordinate the solutions obtained in subsequent layers of optimization. The proposed method has the following advantages over the existing ones:

1. The developed hierarchical Bayesian framework offers a systematic way to select the model structure, determine the localization parameters as well as estimate the model parameters. 2. External information over the model parameters, localization parameters, and model structure can be incorporated in the identification process. 3. Sparsity and heteroscedasticity of training samples can be effectively handled. 4. Bayesian inference at a particular level takes into account the uncertainty in the estimates of the previous level. 5. Bayesian model selection can automatically penalize model complexity to avoid the over-fitting issue [19].

The remainder of this chapter is organized as follows: the following section introduces the basic formulation of the LW-PLS model and discusses the limitations of the regular LW-PLS modeling method which lead to consideration of an integrated Bayesian framework. Then, the motivation behind adopting a hierarchical approach is outlined and each level of inference is explained in detail. Next, an overall procedure to implement the proposed hierarchical framework is shown. Two industrial case studies are considered to demonstrate the effectiveness of the proposed method based on a set of real-world near infrared spectroscopy data. Finally, the paper is summarized by concluding remarks.

2.2 Problem Statement

Suppose we have a training (calibration) data-set with N samples denoted by:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_N]^T \tag{2.1}$$

$$\mathbf{y} = [y_1, y_2, \cdots y_N]^T \tag{2.2}$$

$\mathbf{X} \in \mathbb{R}^{N \times M}$ and $\mathbf{y} \in \mathbb{R}^{N \times 1}$ are the input and output matrices, respectively. The i-th sample consists of a vector of inputs, $\mathbf{x}_i = [x_{i1}, x_{i2} \cdots x_{iM}]^T$, and an output, y_i. where M is the number of input variables. The formulation of the PLS model is given by

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_X \tag{2.3}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^T + \mathbf{e}_y \tag{2.4}$$

where $\mathbf{T} \in \mathbb{R}^{N \times H}$ denotes a matrix of latent variables, $\mathbf{P} \in \mathbb{R}^{M \times H}$ is a matrix of loadings and $\mathbf{q} \in \mathbb{R}^{1 \times H}$ is a vector of regression coefficients. $\mathbf{E}_X \in \mathbb{R}^{N \times M}$ and $\mathbf{e}_y \in \mathbb{R}^{N \times 1}$ denote the matrices of input and output residuals, respectively.

LW-PLS is an on-line identification method which builds a local PLS model for each query sample. Given a query sample $\mathbf{x}_q$, a similarity matrix is constructed to prioritize the calibration samples:

$$\mathbf{S}_q = diag(s_{1|q}, s_{2|q}, \cdots, s_{N|q}) \tag{2.5}$$

where $s_{i|q}$ $(i = 1, 2 \cdots N)$ is the similarity between $\mathbf{x}_q$ and $\mathbf{x}_i$.

Generally, a measurement of similarity is defined based on a notion of distance between $\mathbf{x}_q$ and $\mathbf{x}_i$. One of the widely used similarity functions is :

$$s_{i|q} = \exp\left(-\frac{d_i}{\sigma_d \lambda}\right) \tag{2.6}$$

$$d_i = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T (\mathbf{x}_i - \mathbf{x}_q)} \tag{2.7}$$

where d_i is the Euclidean distance between $\mathbf{x}_q$ and $\mathbf{x}_i$, σ_d is the standard deviation of $\mathbf{d} = \{d_1, d_2, \cdots d_N\}$ and λ is the localization parameter. Given a σ_d, the similarity decreases more steeply by increasing the distance for larger values of λ. So, λ can determine the acceptable region for selecting the local calibration samples together with σ_d.

LW-PLS models can be constructed by following *Algorithm I* in the appendix [18]. However, this regular LW-PLS algorithm implicitly assumes that the number of latent variables H, *i.e.* model structure, and localization parameter λ are given. In reality, these parameters are often unknown and have critical effects on the estimation accuracy. Even though proper combination of the model structure and localization parameter can be found in advance by using RMSECV, this method cannot maintain good estimation accuracy in a longer term. Multi-mode behaviour of processes and non-linearity of underlying mechanisms affect not only the model parameters, but also the model structure and similarity function. Furthermore, the available prior knowledge cannot be incorporated in the identification process by using the regular LW-PLS algorithm.

Considering these points, in this work, a new similarity function is defined as:

$$s_{i|q}(\varphi) = \exp\left(-d_i \varphi\right) \tag{2.8}$$

where the localization parameter is denoted by φ and treated as a hyperparameter of similarity function to be tuned for each local model. Compared with the similarity

function in the regular LW-PLS (Eqn. 6), in the new similarity function the term $\frac{1}{\sigma_d \lambda}$ has been substituted by φ. In this way, the acceptable region of calibration samples can be directly controlled by tuning φ.

The formulation of the PLS model remains the same as given by Eqns. 3 and 4. The number of retained latent variables H is treated as an unknown variable to be estimated. Therefore, the problem of identifying an LW-PLS model consists of the following steps: 1. prioritizing calibration samples, that can be equally achieved by properly tuning the localization parameter φ; 2. choosing the model structure or number of retained latent variables H; and 3. estimating the model parameters $\Theta = \{\mathbf{P}, \mathbf{T}, \mathbf{q}\}$, *i.e.* loading matrix $\mathbf{P}$, latent variable matrix $\mathbf{T}$, regression coefficient vector $\mathbf{q}$.

From a Bayesian perspective, the problem is converted to maximizing the joint posterior distribution of model parameters, localization parameter, and model structure that is defined as the conditional probability distribution of these variables given the training data-set and query sample, *i.e.* $p(\Theta, \varphi, H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q)$.

2.3 Hierarchical Bayesian Optimization Framework

A Bayesian approach to identify an LW-PLS model is to maximize the posterior probability density function of the model parameters, localization parameter and model structure, $p(\Theta, \varphi, H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q)$. Because of the difficulties associated with the direct maximization of $p(\Theta, \varphi, H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q)$, the problem of interest can be formulated and solved under an iterative hierarchical Bayesian optimization framework [20]. First, the chain rule of probability theory is used to expand the joint posterior probability distribution as:

$$p(\Theta, \varphi, H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q) = p(\Theta|\varphi, H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) p(\varphi|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) p(H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q) \tag{2.9}$$

Next, the optimization problem is decomposed hierarchically into following three layers:

$$\begin{aligned} &\max_{\Theta, \varphi, H} p(\Theta|\varphi, H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) p(\varphi|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) p(H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q) \\ &= \max_{H} \left\{ p(H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q) \left\{ \max_{\varphi} p(\varphi|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) \max_{\Theta} \left\{ p(\Theta|\varphi, H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) \right\} \right\} \right\} \end{aligned} \tag{2.10}$$

2.3.1 Inference of Model Parameters

Applying Bayes' rule, the posterior probability density function (PDF) of model parameters can be written as:

$$\begin{aligned} p(\Theta|\varphi, H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) &= \frac{p(\mathbf{X}, \mathbf{y}|\Theta, \varphi, H, \mathbf{x}_q)p(\Theta|\varphi, H, \mathbf{x}_q)}{p(\mathbf{X}, \mathbf{y}|\varphi, H, \mathbf{x}_q)} \\ &\propto p(\mathbf{X}, \mathbf{y}|\Theta, \varphi, H, \mathbf{x}_q)p(\Theta|\varphi, H, \mathbf{x}_q) \end{aligned} \tag{2.11}$$

where $p(\mathbf{X}, \mathbf{y}|\varphi, H, \mathbf{x}_q)$ is a normalizing constant.

As prior it is reasonable to assume that the model parameters are independent of the localization parameter and query sample. The prior can be explicitly expressed as the conditional joint probability of the loading matrix, regression coefficient vector, and latent variable matrix given the model structure:

$$\begin{aligned} p(\Theta|\varphi, H, \mathbf{x}_q) &= p(\Theta|H) \\ &= p(\mathbf{P}, \mathbf{T}, \mathbf{q}|H) \\ &= p(\mathbf{T}|\mathbf{P}, \mathbf{q}, H)p(\mathbf{q}|\mathbf{P}, H)p(\mathbf{P}|H) \end{aligned} \tag{2.12}$$

Given the loading matrix $\mathbf{P}$, it is reasonable to assume that $\mathbf{T}$ and $\mathbf{q}$ are independent, *i.e.* $p(\mathbf{T}|\mathbf{P}, \mathbf{q}, H) = p(\mathbf{T}|\mathbf{P}, H)$. Thus, the posterior PDF of model parameters can be explicitly written as:

$$p(\mathbf{P}, \mathbf{T}, \mathbf{q}|\varphi, H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) \propto p(\mathbf{X}, \mathbf{y}|\mathbf{P}, \mathbf{T}, \mathbf{q}, \varphi, H, \mathbf{x}_q)p(\mathbf{T}|\mathbf{P}, H)p(\mathbf{q}|\mathbf{P}, H)p(\mathbf{P}|H) \tag{2.13}$$

Following the approach of [21], a new Bayesian approach to solve the problem of LW-PLS modeling is proposed in this section.

For each calibration sample, the LW-PLS formulation is given by:

$$\mathbf{x}_i = \mathbf{P}\mathbf{t}_i + \mathbf{e}_{xi} \tag{2.14}$$

$$y_i = \mathbf{q}\mathbf{t}_i + e_{yi} \tag{2.15}$$

The noise-free inputs and output are given by:

$$\tilde{\mathbf{x}}_\mathbf{i} = \mathbf{P}\mathbf{t}_i \tag{2.16}$$

$$\tilde{y}_i = \mathbf{q}\mathbf{t}_i \tag{2.17}$$

The loading matrix $\mathbf{P}$ has the following unit orthogonal constraint:

$$\mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{2.18}$$

A vector of model parameters, $\mathbf{b} \in \mathbb{R}^{M\times 1}$, representing the relationship between the input and output variables, is defined as:

$$\mathbf{b} = \mathbf{P}\mathbf{q}^{\mathbf{T}} \tag{2.19}$$

The likelihood function relies on the nature of noise. Assume that the input and output measurements are contaminated by mutually independent Gaussian noise, $\mathbf{e}_{xi}$ and e_{yi}, with known variance $\mathbf{Q}_{e_x}$ and Q_{e_y} [21]. The estimation of these unknown variances will be discussed shortly. Given a query sample $\mathbf{x}_q$, the importance weight assigned to the i^{th} calibration sample is denoted by $s_{i|q}$. This is equivalent to saying that:

$$\mathbf{Q}_{e_{xi}} = \frac{\mathbf{Q}_{e_x}}{s_{i|q}} \tag{2.20}$$

$$Q_{e_{yi}} = \frac{Q_{e_y}}{s_{i|q}} \tag{2.21}$$

Normally, a calibration sample with large weight is strongly relevant to the local PLS model. If a calibration sample is far away from the query one, a relatively small importance weight is assigned to it in order to reduce its contribution to the local PLS model. This would be equivalent to resulting in a large noise term, meaning that this point contains more information about noise or, equivalently, less information about the model parameters. Note that if the weight is equal to zero, *i.e.* $s_{i|q} = 0$, the variance of noise will approach infinity and the corresponding point will be completely excluded in identifying the local PLS regression model. It is assumed that the measurement noises of the observations are independent. It is also assumed that the measurement noises of inputs and output are mutually independent. Thus, the likelihood can be simplified as follows:

$$p(\mathbf{X}, \mathbf{y}|\mathbf{P}, \mathbf{T}, \mathbf{q}, \varphi, H, \mathbf{x}_q) = p(\mathbf{X}|\mathbf{P}, \mathbf{T}, \varphi, H, \mathbf{x}_q)p(\mathbf{y}|\mathbf{T}, \mathbf{q}, \varphi, H, \mathbf{x}_q) \tag{2.22}$$

$$p(\mathbf{X}|\mathbf{P}, \mathbf{T}, \varphi, H, \mathbf{x}_q) = \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{P}, \mathbf{t}_i, \varphi, H, \mathbf{x}_q) \tag{2.23}$$

$$p(\mathbf{y}|\mathbf{T}, \mathbf{q}, \varphi, H, \mathbf{x}_q) = \prod_{i=1}^{N} p(y_i|\mathbf{t}_i, \mathbf{q}, \varphi, H, \mathbf{x}_q) \tag{2.24}$$

$$\mathbf{x}_i|\mathbf{P}, \mathbf{t}_i, \mathbf{q}, \varphi, H, \mathbf{x}_q \sim \mathcal{N}(\mathbf{P}\mathbf{t}_i, \frac{\mathbf{Q}_{e_x}}{s_{i|q}}) \tag{2.25}$$

$$y_i|\mathbf{P}, \mathbf{t}_i, \mathbf{q}, \varphi, H, \mathbf{x}_q \sim \mathcal{N}(\mathbf{q}\mathbf{t}_i, \frac{Q_{e_y}}{s_{i|q}}) \tag{2.26}$$

The priors over the model parameters depend on the nature of the noise-free data. The noise-free inputs are assumed to follow a multivariate Gaussian distribution, that is

$$\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mu_{\mathbf{x}}, \mathbf{Q}_{\mathbf{x}}) \tag{2.27}$$

As a result, given the loading matrix $\mathbf{P}$, the latent variable $\mathbf{t_i}$ will also follow a conditional multivariate Gaussian distribution:

$$\mathbf{t_i} = \mathbf{P}^T\tilde{\mathbf{x}}_i \tag{2.28}$$

$$\mathbf{t_i}|\mathbf{P}, H \sim \mathcal{N}(\mathbf{P}^T\mu_x, \mathbf{P}^T\mathbf{Q}_x\mathbf{P}) \tag{2.29}$$

It is also assumed that the model parameters $\mathbf{b}$ follow a multivariate Gaussian distribution:

$$\mathbf{b} \sim \mathcal{N}(\mu_b, \mathbf{Q}_b) \tag{2.30}$$

Given the loading matrix $\mathbf{P}$, and the vector of model parameters $\mathbf{b}$, the regression coefficient vector $\mathbf{q}^T$ also follow a conditional multivariate Gaussian distribution

$$\mathbf{q}^T = \mathbf{P}^T\mathbf{b} \tag{2.31}$$

$$\mathbf{q}^T|\mathbf{P}, H \sim \mathcal{N}(\mathbf{P}^T\mu_b, \mathbf{P}^T\mathbf{Q}_b\mathbf{P}) \tag{2.32}$$

In the absence of any external knowledge over the loading matrix $\mathbf{P}$, a uniform prior distribution can be specified over $\mathbf{P}$. Based on the likelihood and prior distributions, the posterior distribution can be determined as:

$$p(\mathbf{X}, \mathbf{y}|\mathbf{P}, \mathbf{T}, \mathbf{q}, \varphi, H, \mathbf{x}_q) \propto p(\mathbf{X}|\mathbf{P}, \mathbf{T}, \varphi, H, \mathbf{x}_q)p(\mathbf{y}|\mathbf{T}, \mathbf{q}, \varphi, H, \mathbf{x}_q)p(\mathbf{T}|\mathbf{P}, H)p(\mathbf{q}|\mathbf{P}, H) \tag{2.33}$$

The maximum a posteriori probability (MAP) estimates can be obtained by solving the following optimization problem:

$$\begin{aligned}\{\mathbf{P},\mathbf{T},\mathbf{q}\}_{MAP} = &\arg\max_{\mathbf{P},\mathbf{T},\mathbf{q}}\{p(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q)p(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q)\\ &\times p(\mathbf{T}|\mathbf{P},H)p(\mathbf{q}|\mathbf{P},H)\}\\ s.t.\ \ &\mathbf{P}^T\mathbf{P}=\mathbf{I}\end{aligned} \tag{2.34}$$

It is intractable to solve this optimization problem directly. The overall objective function can be decomposed into the following three simultaneous parameter-estimation and data-reconciliation optimization problem.

$$\begin{aligned}&\{\mathbf{P}\}_{MAP} = \arg\max_{\mathbf{P}} p(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q)p(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q)\\ &\{\mathbf{q}\}_{MAP} = \arg\max_{\mathbf{q}} p(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q)\ p(\mathbf{q}|\mathbf{P},H)\\ &s.t.\\ &\{\mathbf{T}\}_{MAP} = \arg\max_{\mathbf{T}} p(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q)p(\mathbf{T}|\mathbf{P},H)\\ &\mathbf{P}^T\mathbf{P}=\mathbf{I}\end{aligned} \tag{2.35}$$

Since likelihood and priors are all multivariate Gaussian, the MAP estimates can be equivalently obtained by solving the following minimization problems:

$$\begin{aligned}&\{\mathbf{P}\}_{MAP} = \arg\min_{\mathbf{P}}\{\sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)^T(\tfrac{\mathbf{Q}_{ex}}{s_{i|q}})^{-1}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)+\sum_{i=1}^{N}(y_i-\mathbf{q}\mathbf{t}_i)^T(\tfrac{\mathbf{Q}_{ey}}{s_{i|q}})^{-1}(y_i-\mathbf{q}\mathbf{t}_i)\}\\ &\{\mathbf{q}\}_{MAP} = \arg\min_{\mathbf{q}}\{\sum_{i=1}^{N}(y_i-\mathbf{q}\mathbf{t}_i)^T(\tfrac{\mathbf{Q}_{ey}}{s_{i|q}})^{-1}(y_i-\mathbf{q}\mathbf{t}_i)+(\mathbf{q}^T-\mathbf{P}^T\mu_b)^T(\mathbf{P}^T\mathbf{Q}_b\mathbf{P})^{-1}(\mathbf{q}^T-\mathbf{P}^T\mu_b)\}\\ &\{\mathbf{t}_i\}_{MAP} = \arg\min_{\mathbf{t}_i}\{(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)^T(\tfrac{\mathbf{Q}_{ex}}{s_{i|q}})^{-1}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)\ \ +(\mathbf{t}_i-\mathbf{P}^T\mu_{\mathbf{x}})^T(\mathbf{P}^T\mathbf{Q}_{\mathbf{x}}\mathbf{P})^{-1}(\mathbf{t}_i-\mathbf{P}^T\mu_{\mathbf{x}})\\ &s.t.\ \ \mathbf{P}^T\mathbf{P}=\mathbf{I}\end{aligned} \tag{2.36}$$

The first optimization function is intractable to solve because of the unit orthonormal constraint. We can first use optimization methods that have a closed form solution, to estimate $\mathbf{P}$. In this way, both of the following optimization problems can be solved analytically.

$$\{\mathbf{t}_i\}_{MAP} = [\mathbf{P}^T(\frac{\mathbf{Q}_{e_x}}{s_{i|q}})^{-1}\mathbf{P}+(\mathbf{P}^T\mathbf{Q}_{\mathbf{x}}\mathbf{P})^{-1}]^{-1}[\mathbf{P}^T(\frac{\mathbf{Q}_{e_x}}{s_{i|q}})^{-1}\mathbf{x}_i+(\mathbf{P}^T\mathbf{Q}_{\mathbf{x}}\mathbf{P})^{-1}\mathbf{P}^T\mu_{\mathbf{x}}] \tag{2.37}$$

$$\{\mathbf{q}^T\}_{MAP} = \left[\mathbf{T}^T\mathbf{S}_{\mathbf{q}}\mathbf{T}Q_{e_y}+\left(\mathbf{P}^T\mathbf{Q}_b\mathbf{P}\right)^{-1}\right]^{-1}\left[\mathbf{T}^T\mathbf{S}_{\mathbf{q}}\mathbf{Y}Q_{e_y}+\left(\mathbf{P}^T\mathbf{Q}_b\mathbf{P}\right)^{-1}\mathbf{P}^T\mu_b\right] \tag{2.38}$$

In this Bayesian modeling algorithms, $\mathbf{Q}_{e_x}, Q_{e_y}, \mu_b, \mathbf{Q}_b, \mu_x, \mathbf{Q}_x$ are assumed to be known. That means the prior density was assumed to be fully specified in advance. In the presence of limited prior knowledge of the noise variance and model parameters, a widely used alternative is the empirical Bayesian analysis which estimates the prior from the available data assuming data is representative [22]. In the empirical Bayesian analysis, there are two kinds of approaches to estimate the prior from data: parametric approach and nonparametric approach [21]. The parametric approach assuming the structures of the prior distribution are known and it only needs to estimate the hyperparameters of the prior density function . The nonparametric approach will estimate the entire prior from the data which is more complex and time-consuming. For computational convenience, the parametric approach is used to estimate the prior in light of training data using *Algorithm II* in the appendix.

2.3.2 Inference of Localization Parameter

Applying Bayes' rule, the posterior PDF of localization parameter can be expressed as:

$$p(\varphi|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) \propto p(\mathbf{X}, \mathbf{y}|\varphi, H, \mathbf{x}_q)p(\varphi|H, \mathbf{x}_q) \tag{2.39}$$

As priors, one can assume that the localization parameter φ is statistically independent of the model structure H and the query sample $\mathbf{x}_q$:

$$p(\varphi|H, \mathbf{x}_q) = p(\varphi) \tag{2.40}$$

In the absence of any external knowledge, a non-informative prior can be specified in the form of a constrained uniform distribution. To incorporate the available prior knowledge, conjugate priors are normally utilized for which the resulting posterior distribution can be conveniently evaluated. To assure generality, a Gamma prior distribution is specified over the localization parameter:

$$p(\varphi) = \frac{\varphi^{a-1}}{b^a \Gamma(a)} \exp(-\frac{\varphi}{b}) \tag{2.41}$$

where a is the shape parameter and b is the scale parameter. The likelihood in Eqn. 39, can be evaluated by integrating out the model parameters:

$$p(\mathbf{X},\mathbf{y}|\varphi,H,\mathbf{x}_q)=\int_{\Theta} p(\mathbf{X},\mathbf{y}|\Theta,\varphi,H,\mathbf{x}_q)p(\Theta|H)d\Theta \tag{2.42}$$

Since the above problem is often intractable, the integral in Eqn. 42 can be approximated by applying Laplace's method of approximation [23].

$$\int_{\Theta} p(\mathbf{X},\mathbf{y}|\Theta,\varphi,H,\mathbf{x}_q)p(\Theta|H)d\Theta \approx p(\mathbf{X},\mathbf{y}|\Theta^{MAP},\varphi,H)p(\Theta^{MAP}|H)\det(\frac{\mathbf{A}_\Theta}{2\pi})^{-\frac{1}{2}} \tag{2.43}$$

where $\mathbf{A}_\Theta=-\nabla\nabla\log p(\Theta|\varphi,H,\mathbf{X},\mathbf{y},\mathbf{x}_q)$. The inverse of Hessian matrix $\mathbf{A}_\Theta$ reflects the posterior uncertainty in Θ. Then the MAP estimate of localization parameter can be shown as:

$$\begin{aligned}\{\varphi\}_{MAP}&=\arg\max_{\varphi}\{p(\varphi|H,\mathbf{X},\mathbf{y},\mathbf{x}_q)\}\\&=\arg\max_{\varphi}\{p(\mathbf{X},\mathbf{y}|\Theta^{MAP},\varphi,H,\mathbf{x}_q)p(\Theta^{MAP}|H)\det(\tfrac{\mathbf{A}_\Theta}{2\pi})^{-\frac{1}{2}}p(\varphi)\}\end{aligned} \tag{2.44}$$

Since both the likelihood and prior probability density functions belong to the family of exponential PDFs, the MAP solution can be obtained by solving the following minimization problem:

$$\{\varphi\}_{MAP}=\arg\min_{\varphi}\left\{\begin{array}{l}\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\hat{\mathbf{x}}_i)^T(\frac{\mathbf{Q}_{e_x}}{s_{i|q}})^{-1}(\mathbf{x}_i-\hat{\mathbf{x}}_i)\\+\frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^T(\frac{Q_{e_y}}{s_{i|q}})^{-1}(y_i-\hat{y}_i)\\+(1-a)\log\varphi+\frac{1}{b}\varphi\\-\log[\det(\frac{\mathbf{A}_\Theta}{2\pi})^{-\frac{1}{2}}]-\frac{M+1}{2}\log\prod_{i=1}^{N}s_{i|q}\end{array}\right\} \tag{2.45}$$

This optimization problem can be solved by the sampling method instead of deriving a closed form solution which cannot be obtained directly. For instance, the continuous localization parameter can be discretized into a finite set of reasonable values $\{\varphi_1,\varphi_2,...\varphi_f\}$. We can next draw samples from the posterior distribution using these candidate values of the localization parameters to approximate the MAP solution.

2.3.3 Inference of Model Structure

Applying Bayes' rule, the posterior PDF of model structure can be expressed as:

$$P(H|\mathbf{X},\mathbf{y},\mathbf{x}_q)\propto p(\mathbf{X},\mathbf{y}|H,\mathbf{x}_q)P(H|\mathbf{x}_q) \tag{2.46}$$

As priors, it is reasonable to assume that the model structure is statistically independent of the query sample $\mathbf{x}_q$. Given a set of candidate model structures, *i.e* $H \in \{H_1, H_2 ... H_L\}$, the random variable H is a categorical variable and can be modelled by

$$p(H) = \prod_{l=1}^{L} p(H = H_l)^{[H=H_l]} \tag{2.47}$$

where $[H = H_l]$ equals 1 if $H = H_l$ and equals 0 otherwise. In the absence of any prior information, a uniform distribution can be used for the candidate model structures, *i.e.* $p(H = H_1) = p(H = H_2) ... = p(H = H_L)$.

The likelihood function $p(\mathbf{X}, \mathbf{y}|H, \mathbf{x}_q)$ can be obtained by integrating out the localization parameter:

$$p(\mathbf{X}, \mathbf{y}|H, \mathbf{x}_q) = \int_{\varphi} p(\mathbf{X}, \mathbf{y}|\varphi, H, \mathbf{x}_q) p(\varphi) d\varphi \tag{2.48}$$

Since it is intractable to solve the above integral directly, Laplace's method of approximation is applied again:

$$\int_{\varphi} p(\mathbf{X}, \mathbf{y}|\varphi, H, \mathbf{x}_q) p(\varphi) d\varphi \approx p(\mathbf{X}, \mathbf{y}|\varphi^{MAP}, H, \mathbf{x}_q) p(\varphi^{MAP}) \det (\frac{A_\varphi}{2\pi})^{-\frac{1}{2}} \tag{2.49}$$

where $A_\varphi = -\nabla\nabla \log p(\varphi|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q)$. The inverse of Hessian matrix A_φ reflects the posterior uncertainty in φ.

Finally, the MAP estimate of the model structure can be obtained as follows:

$$\begin{aligned}
\{H\}_{MAP} &= \arg\max_H (p(H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q)) \\
&= \arg\max_H \{p(\mathbf{X}, \mathbf{y}|\varphi^{MAP}, H, \mathbf{x}_q) p(\varphi^{MAP}) \det (\tfrac{A_\varphi}{2\pi})^{-\frac{1}{2}} p(H)\} \\
&= \arg\max_H \{p(\mathbf{X}, \mathbf{y}|\Theta^{MAP}, \varphi^{MAP}, H, \mathbf{x}_q) p(\Theta^{MAP}|H) p(\varphi^{MAP}) \det (\tfrac{\mathbf{A}_\Theta}{2\pi})^{-\frac{1}{2}} \det (\tfrac{A_\varphi}{2\pi})^{-\frac{1}{2}} p(H)\}
\end{aligned} \tag{2.50}$$

Since both the likelihood and prior probability density functions belong to the family of exponential PDFs, the MAP solution can be obtained by solving the following

minimization problem:

$$\{H\}_{MAP} = \arg\min_{H} \left\{ \begin{array}{l} \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\mathbf{x}}_i)^T(\frac{\mathbf{Q}_{e_x}}{s_{i|q}})^{-1}(\mathbf{x}_i - \hat{\mathbf{x}}_i) \\ +\frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^T(\frac{Q_{e_y}}{s_{i|q}})^{-1}(y_i - \hat{y}_i) \\ +\frac{1}{2}\sum_{i=1}^{n}(\mathbf{t}_i - \mathbf{P}^T\mu_x)^T(\mathbf{P}^T\mathbf{Q}_x\mathbf{P})^{-1}(\mathbf{t}_i - \mathbf{P}^T\mu_x) \\ +\frac{1}{2}(\mathbf{q}^T - \mathbf{P}^T\mu_b)^T(\mathbf{P}^T\mathbf{Q}_b\mathbf{P})^{-1}(\mathbf{q}^T - \mathbf{P}^T\mu_b) \\ +(1-a)\log\varphi + \frac{1}{b}\varphi - \log[\det(\frac{\mathbf{A}_\Theta}{2\pi})^{-\frac{1}{2}}] - \frac{M+1}{2}\log\prod_{i=1}^{N} s_{i|q} \\ +\frac{1}{2}(1+N)H\log 2\pi - \log[\det(\frac{A_\varphi}{2\pi})^{-\frac{1}{2}}] \end{array} \right\} \tag{2.51}$$

2.4 Hierarchical Bayesian Optimization Procedure

1. Choose the similarity function given in Eqn. 2.8.

2. Select a proper set of candidate model structures $\{H_1, H_2 ... H_L\}$. If there is available prior information about the model structures, the candidates and their prior probabilities $p(H)$ can be determine based on the prior knowledge. If there is no prior information, the candidate model structures can be selected based on empirical method: select several candidate model structures around the globally optimal one obtained from off-line LOOCV, and set a uniform prior distribution over this set of candidate model structures.

3. Characterize the noise variances, $\mathbf{Q}_{e_x}$ and Q_{e_y}, and specify a prior distribution over the model parameters, $p(\Theta|H)$, using *Algorithm II*.

4. Characterize the prior distribution over localization parameter, $p(\varphi|H)$, using *Algorithm III*.

5. For $l = 1 : L$

   (1). Select H_l and choose an initial value for the localization parameter φ_l.

   (2). While $\mathbf{P}_l, \mathbf{T}_l, \mathbf{q}_l$ and φ_l converge

   (2.1). calculate the similarity matrix, $\mathbf{S}_{q_l}$, using Eqns. 2.5, 2.7, and 2.8.

   (2.2). calculate the loading matrix, $\mathbf{P}_l$, by applying the LW-PLS algorithm to $\{\mathbf{X}, \mathbf{y}, \mathbf{x}_q\}$.

(2.3). calculate the regression coefficient vector, $\mathbf{q}_l$, latent variable matrix, $\mathbf{T}_l$, using Eqns. 3.43 and 3.44;

(2.4).calculate the localization parameter φ_l using Eqn. 2.45.

(3). Calculate the posterior probability of model structure, $p(H = H_l|\mathbf{X}, \mathbf{y}, \mathbf{x}_q)$, using Eqn. 2.51.

6. Choose the model structure with the highest posterior probability as well as corresponding loading matrix, $\mathbf{P}$, and regression coefficient vector, $\mathbf{q}$.

7. Calculate output as $\hat{y} = \mathbf{x}_q\mathbf{P}\mathbf{q}^T$.

2.5 Case Studies

This section demonstrates the practical application of the Bayesian LW-PLS through case studies. To illustrate the advantages of hierarchical Bayesian optimization, two sets of near NIR data for real-time prediction of reid vapor pressure (RVP) of Gasoline and wheat kernels are used. It is noteworthy that the NIR data-sets have high dimension with strongly correlated spectra. All industrial data presented here have been normalized in order to protect proprietary information.

2.5.1 Reid Vapor Pressure of Gasoline

The objective of this study is to estimate Reid Vapor Pressure of Gasoline from NIR spectra data. The set of data is taken from [18]. The data-set consists of NIR spectra for 423 gasoline samples. The diffusion reflectance spectra of samples are measured with wavelength range of 800-1,700 nm in 1 nm intervals (Figure 2.1). The samples are divided into 296 calibration or training data-set and 127 validation or test samples. Standard ASTM testing methodologies have been used to obtain the reference measurements for RVP.

In order to show the features of the proposed method more clearly, the performance is evaluated in the following three scenarios:

Scenario I: known localization parameter and model structure, but unknown model parameters.

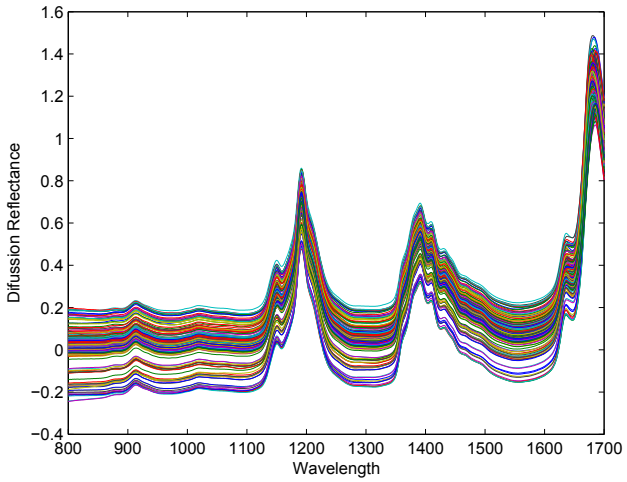


Figure 2.1: Reid vapor pressure of gasoline spectra data.

The 1*st* layer of the proposed method, inference of model parameters, is applied to develop real-time LW-PLS models for the prediction of RVP. The prediction performance of the developed models is compared with that of the models identified using regular LW-PLS regression. The similarity functions are chosen as in Eqn. 2.6 and the localization parameter λ and the number of retained latent variables H are set as 0.5 and 30 respectively and same for both methods. The prior distributions of model parameters are specified by using *Algorithm II*. The comparison results are reported in Table 2.1 and Figure 2.2. It can be observed that the Bayesian parameter estimation is more accurate than the regular LW-PLS for some of the calibration samples. A slightly higher (1%) prediction performance has been achieved by incorporating the prior knowledge and taking into account the different contributions of noise in the measurements. The challenge in using this Bayesian approach for estimation exists not only in obtaining proper prior distribution but also in specifying appropriate noise variance. Since no prior information is available, the variances of measurement noise can only be estimated from existing sources such as the calibration data. However, the main challenge in the locally weighted methods is simultaneous estimation of localization parameter, model structure and model parameters where the proposed Bayesian approach shows its great advantage, as demonstrated in the following scenarios.

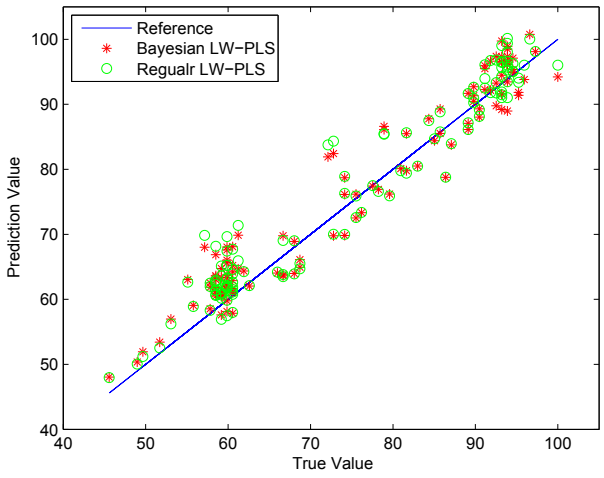


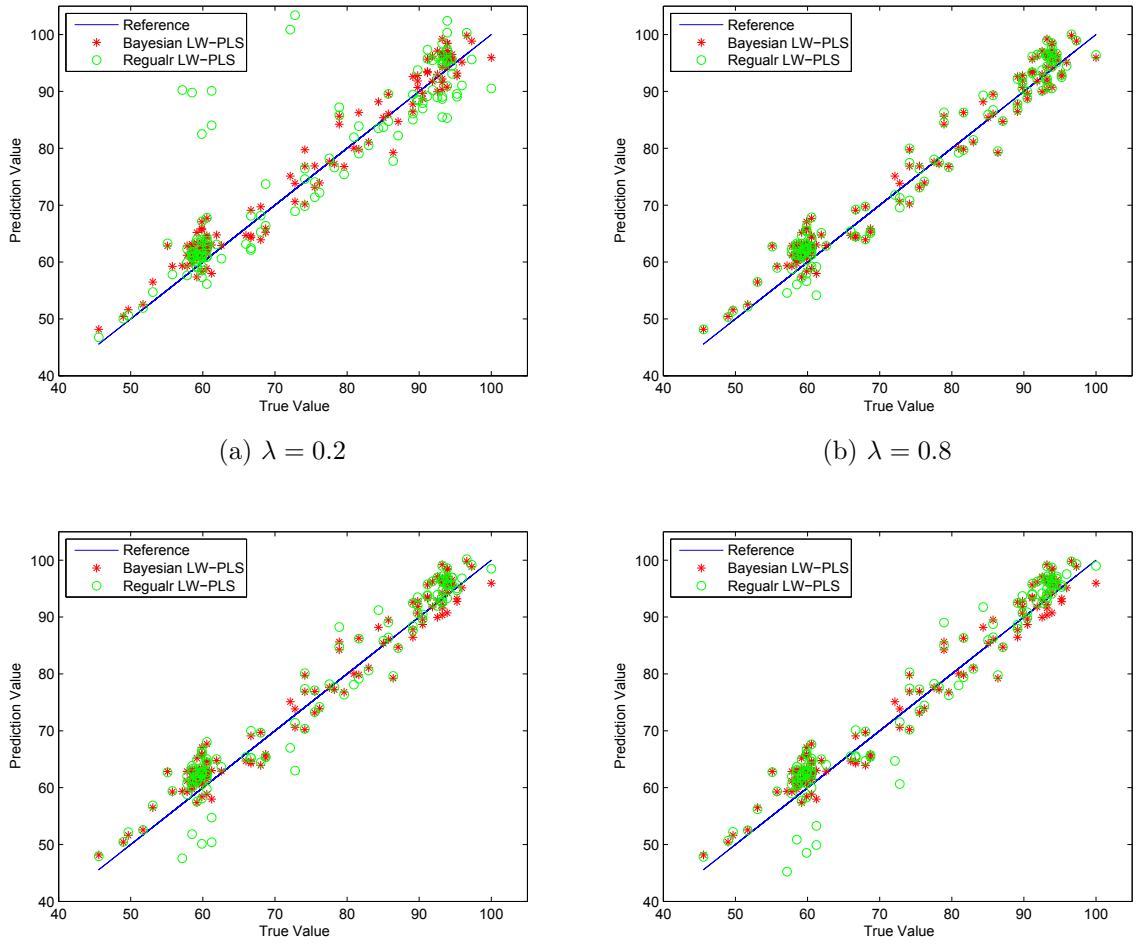
Figure 2.2: Cross-validation using data-set from reid vapor pressure of gasoline, scenario I

	Bayesian LW-PLS	Regular LW-PLS
Localization parameter $\lambda = 0.5$		
Selected Number of Retained LVs $H = 30$		
MSE of cross-validation	15.4347	15.6006
Correlation of cross-validation	0.9732	0.9731

Table 2.1: Comparing prediction performance of the $1st$ layer of Bayesian LW-PLS and regular LW-PLS using data-set from reid vapor pressure of gasoline, scenario I

Scenario II: known model structure, but unknown localization parameter and model parameters.

The $1st$ and $2nd$ layers of the proposed method, estimation of the model parameters, and selection of localization parameter are applied to identify the LW-PLS models. The number of retained latent variables is set as 30. For the regular LW-PLS, the classic similarity function (Eqn. 2.6) is used and we consider four different values for localization parameter λ: 0.2 , 0.8, 1.5 and 2. For the proposed method, the new similarity function (Eqn. 2.8) is used. The prior distribution of model parameters is specified by using *Algorithm II*. The prior distribution over the localization parameters φ is specified by using *Algorithm III* within sampling range [0.1,2]. From the



(a) $\lambda = 0.2$

(b) $\lambda = 0.8$

(c) $\lambda = 1.5$

(d) $\lambda = 2$

Figure 2.3: Cross-validation using data-set from reid vapor pressure of gasoline, scenario II

results shown in Table 2.2 and Figure 2.3, it can be observed that the performance of the regular LW-PLS method depends highly on the value of the localization parameter. Therefore, proper tuning of the localization parameters has a significant effect on the prediction performance of the LW-PLS models. Since the Bayesian LW-PLS searches for the locally optimal value of the localization parameter within the developed hierarchical optimization framework, the prediction performance of the resulting LW-PLS models is superior. Figure 2.4 shows that for different local models, different optimal localization parameters have been obtained to achieve a better performance.

	Bayesian LW-PLS	Regular LW-PLS
Selected Number of Retained LVs $H = 30$		
Localization parameter	$\varphi \in [0.1, 2]$	$\lambda = 0.2$
MSE of cross-validation	9.9807	57.5943
Correlation of cross-validation	0.9833	0.8846
Localization parameter	$\varphi \in [0.1, 2]$	$\lambda = 0.8$
MSE of cross-validation	9.9807	10.1754
Correlation of cross-validation	0.9833	0.9816
Localization parameter	$\varphi \in [0.1, 2]$	$\lambda = 1.5$
MSE of cross-validation	9.9807	14.1729
Correlation of cross-validation	0.9833	0.9743
Localization parameter	$\varphi \in [0.1, 2]$	$\lambda = 2$
MSE of cross-validation	9.9807	15.9304
Correlation of cross-validation	0.9833	0.9713

Table 2.2: Comparing prediction performance of the $1st$ and $2nd$ layer of Bayesian LW-PLS and regular LW-PLS using data-set from reid vapor pressure of gasoline, scenario II

Scenario III: unknown model structure, localization parameter and model parameters.

The proposed method, Bayesian LW-PLS and one widespread method, RMSECV-based LW-PLS are applied to develop the LW-PLS models for real-time prediction of RVP. The main idea behind RMSECV is to search for the globally optimal localization parameter and model structure by minimize the RMSE of leave-one-out cross-validation (LOOCV) in an off-line identification phase and then apply the LW-PLS to do on-line estimation of the model parameters. The candidate model structures are set as $[25, 30]$ for both methods. The result of RMSECV for optimal localization parameters and number of retained latent variables are 2 and 30 respectively. For Bayesian LW-PLS, first, the prior distributions over the model parameters are specified by using *Algorithm II*. The prior distribution over the localization parameters φ is specified by using *Algorithm III* within sampling range [0.1,2]. In the absence of the prior knowledge, a uniform distribution is used for the model structure. The comparison results are reported in Table 2.3 and illustrated in Figure 2.5. According to the results, Bayesian LW-PLS performs much better than the traditional method,

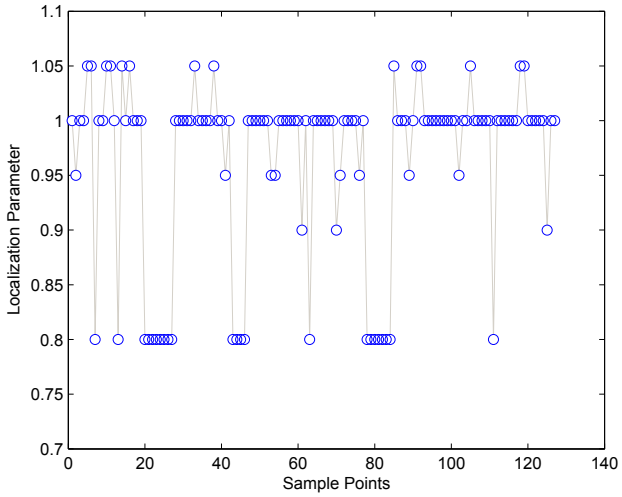


Figure 2.4: Localization parameter φ of Bayesian LW-PLS

RMSECV-based LW-PLS.

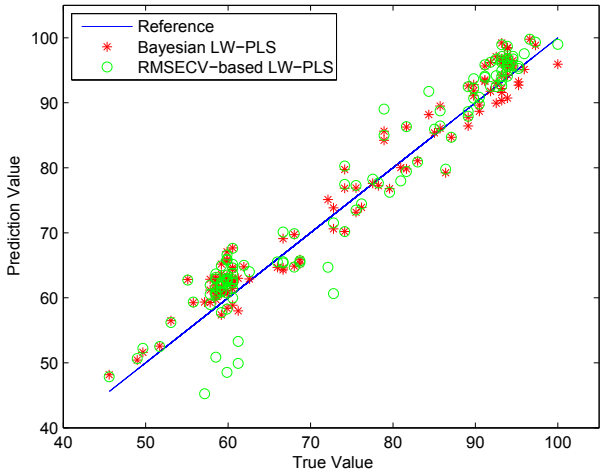
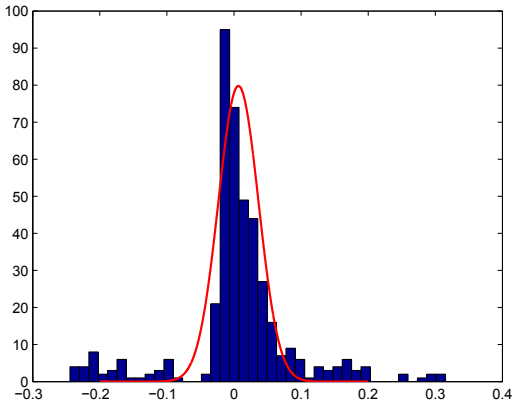
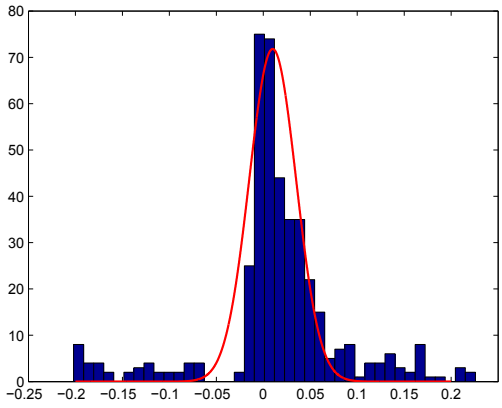


Figure 2.5: Cross-validation using data-set from reid vapor pressure of gasoline, scenario III

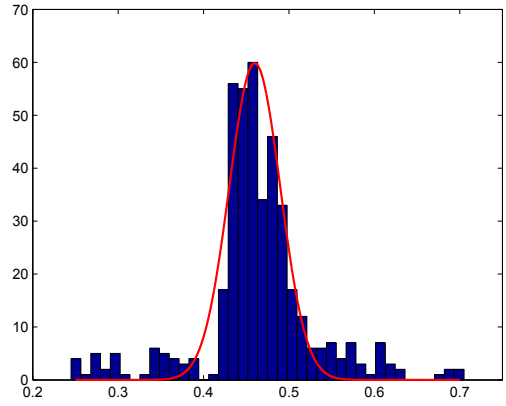
In all of these three scenarios, the priors over main parameter are obtained from estimation of empirical prior *i.e.*, *Algorithm II*. The assumption behind this approach is Gaussian distributed inputs. As shown in Figure 2.6, the distribution of the input can be well approximated by Gaussian distribution.



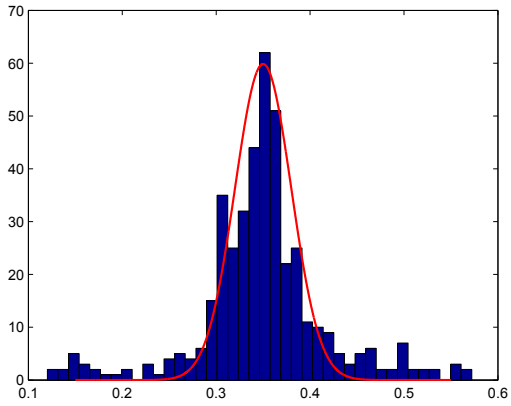
(a) NIR spectra at wavelength 800nm



(b) NIR spectra at wavelength 1100nm



(c) NIR spectra at wavelength 1400nm



(d) NIR spectra at wavelength 1650nm

Figure 2.6: Distributions of selected inputs for reid vapor pressure of gasoline example

	Bayesian LW-PLS	RMSECV
Selected Number of retained LVs	[25,30]	30
Localization parameter	$\varphi \in [0.1, 2]$	$\lambda = 2$
MSE of cross-validation	9.9499	15.9304
Correlation of cross-validation	0.9835	0.9713

Table 2.3: Comparing prediction performance of Bayesian LW-PLS and RMSECV-based LW-PLS using data-set from reid vapor pressure of gasoline, scenario III

2.5.2 Protein Content of Wheat Kernels

In this case study, the LW-PLS models are developed for on-line prediction of the protein content of wheat kernels from NIR spectra. This data set was used by [24, 25] as a standard NIR data-set. The wheat kernels were randomly chosen from bulk samples representing different varieties or various mixtures from two different locations in Denmark.

The calibration and test data-sets collected in this study consist of 100 and 105 samples with reference value ranges from 46.1 to 103.4 and 47.8 to 93.7, respectively. As stated by [24, 25], the test samples were acquired with the calibration samples, but stored for about 2 additional months before measurement in order to provide a check for temporal drift in the samples and instrumentation.

The Bayesian LW-PLS and RMSECV-based LW-PLS are applied to develop the calibration models for protein content. The candidate model structures are set as $[8, 12]$ for both methods. The optimal localization parameter, λ, and number of retained latent variables, H, obtained via RMSECV, are 0.2 and 9 respectively. For the Bayesian LW-PLS, the prior distribution of model parameters are specified by following the procedure in *Algorithm II*. A Gamma prior distribution over the localization parameter φ is extracted from calibration data by using *Algorithm III* and the corresponding sampling range is chosen as $[0.1, 2]$. From comparison results reported in Table 2.4 and illustrated in Figure 2.7, it can, again, be observed that the Bayesian LW-PLS significantly outperforms the RMSECV-based LW-PLS.

In order to further evaluate the effectiveness of the proposed method, a case of extrapolation is performed on the same NIR data-set. The calibration samples which have output value in the range of 82.3 to 103.4 are selected to form the new calibration

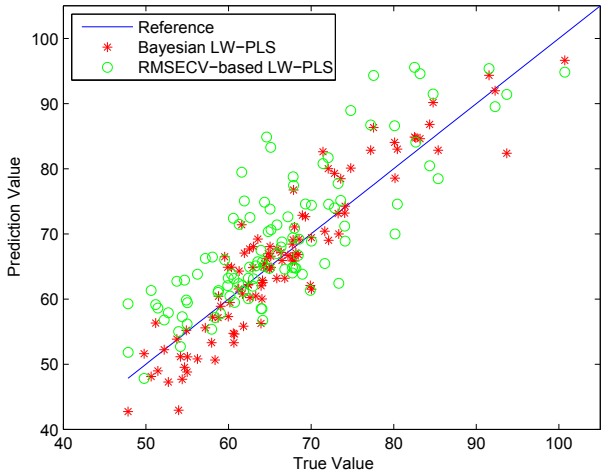


Figure 2.7: Cross-validation using data-set from protein content of wheat kernels

	Bayesian LW-PLS	RMSECV
Selected Number of retained LVs	[8,12]	9
Localization parameter	$\varphi \in [0.1, 2]$	$\lambda = 0.2$
MSE of cross-validation	20.8174	48.3352
Correlation of cross-validation	0.9365	0.8517

Table 2.4: Comparing prediction performance of the Bayesian LW-PLS and RMSECV-based LW-PLS using data-set from wheat kernels

data-set. The validation or test data sets remain unchanged which have output value between 47.8 to 93.6 so that the calibration data-set does not overlap with all the operation region of test ones. It means some of the prediction can only be carried out by extrapolation. This situation can happen in real-world application if a process is shifted to a new operation mode.

From Table 2.5 and Figure 2.8, we can see that the performances of the proposed methods are again much better than the RMSECV-based LW-PLS. Especially in the extrapolated part where outputs range from 47.8 to 82.3, the predictions of RMSECV-based LW-PLS obviously deviate from the reference value, while the predictions of Bayesian one can still follow the reference.

In this case study, the priors over model parameters are also obtained from *Algorithm II*. As shown in Figure 2.9, even though the distribution of the input does not exactly follow Gaussian distribution, the proposed method can still outperform the compared one.

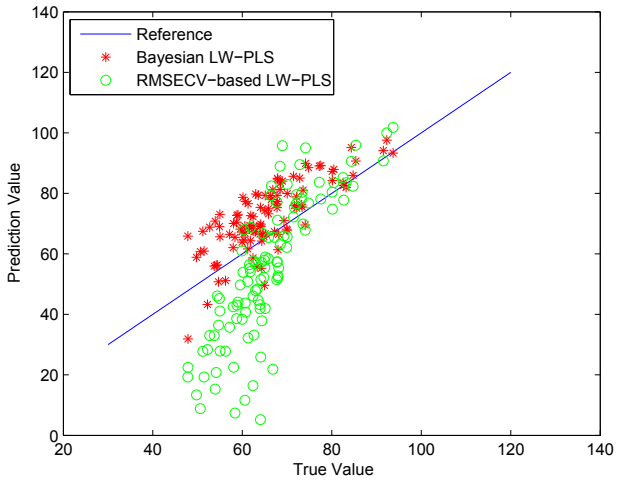


Figure 2.8: Cross-validation using data-set from protein content of wheat kernels, extrapolation case

	Bayesian LW-PLS	RMSECV
Selected Number of retained LVs	[8,12]	9
Localization parameter	$\varphi \in [0.1, 2]$	$\lambda = 0.2$
MSE of cross-validation	94.9904	380.0755
Correlation of cross-validation	0.8207	0.8163

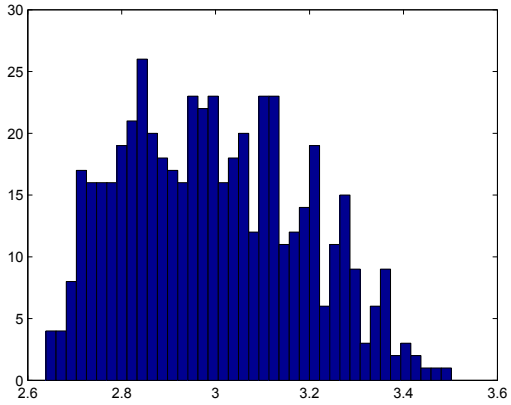
Table 2.5: Comparing prediction performance of the Bayesian LW-PLS and RMSECV-based LW-PLS using data-set from wheat kernels, extrapolation case

Revisit the optimization problem in Equation 2.34 which is equivalent to the following minimization problem:

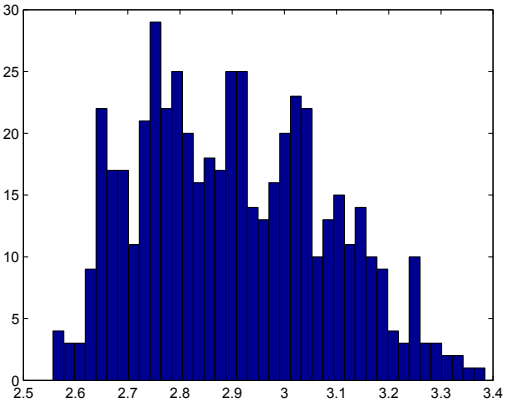
$$\{\mathbf{P},\mathbf{q},\mathbf{T}\}_{MAP} = \underset{\mathbf{P},\mathbf{q},\mathbf{T}}{\arg\min} \left\{ \begin{array}{l} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{P}\mathbf{t}_i)^T (\frac{\mathbf{Q}_{ex}}{s_{i|q}})^{-1} (\mathbf{x}_i - \mathbf{P}\mathbf{t}_i) \\ + \sum_{i=1}^{N} (y_i - \mathbf{q}\mathbf{t}_i)^T (\frac{\mathbf{Q}_{ey}}{s_{i|q}})^{-1} (y_i - \mathbf{q}\mathbf{t}_i) \\ + (\mathbf{q}^T - \mathbf{P}^T\mu_b)^T (\mathbf{P}^T\mathbf{Q}_b\mathbf{P})^{-1} (\mathbf{q}^T - \mathbf{P}^T\mu_b) \\ + \sum_{i=1}^{N} (\mathbf{t}_i - \mathbf{P}^T\mu_{\mathbf{x}})^T (\mathbf{P}^T\mathbf{Q}_{\mathbf{x}}\mathbf{P})^{-1} (\mathbf{t}_i - \mathbf{P}^T\mu_{\mathbf{x}}) \end{array} \right\} \quad (2.52)$$

$$s.t. \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}$$

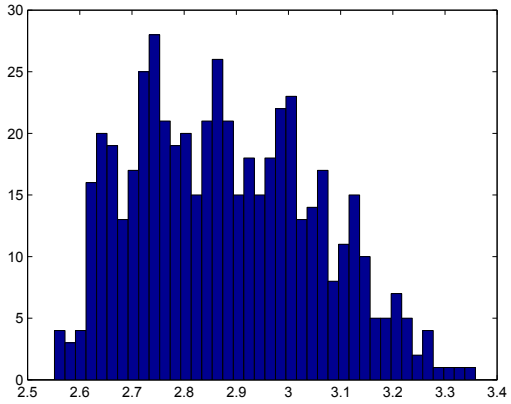
The first two quadratic terms represent the information from historical data, last two quadratic terms contain the information from available prior knowledge over model parameters. If informative priors are obtained beforehand, the last two terms will make a contribution to a more accurate estimation. If the prior contains little helpful information (situation in this case), a fairly good estimation of model parameters can



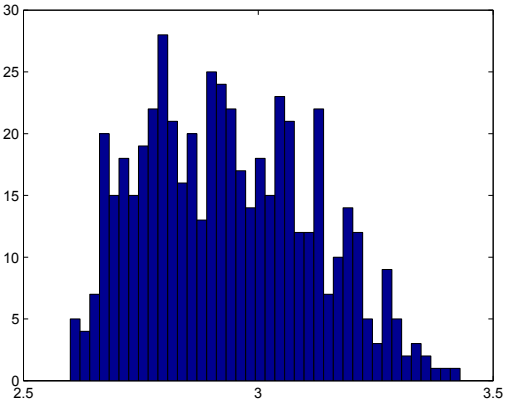
(a) NIR spectra at wavelength 900nm



(b) NIR spectra at wavelength 925nm



(c) NIR spectra at wavelength 950nm



(d) NIR spectra at wavelength 975nm

Figure 2.9: Distributions of selected inputs for protein content of wheat kernels example

be achieved by taking advantage of information contained in data. Moreover, the next two layers of Bayesian framework, *i.e.* inference of localization parameter and model structure can further improve the performance.

2.6 Conclusion

This chapter proposed a holistic Bayesian framework for the LW-PLS regression. The proposed method has the following advantages over the regular LW-PLS regression: 1. By following a Bayesian approach to estimate the model parameters of the LW-PLS model, available prior knowledge can be incorporated into the identification process. 2. Different contributions of measurement noise can be taken into account. 3. Application of the hierarchical Bayesian optimization framework offers a systematic and tractable way to get the optimal combination of the model structure, localization parameters as well as model parameters for each operating point. 4. Bayesian model structure selection can automatically deals with the model complexity problem to avoid the over-fitting issue. The attractive features of the proposed framework were illustrated through two industrial case studies in which NIR spectra were used to provide real-time estimates of RVP and wheat kernels using the LW-PLS models. In the first case study, different scenarios were investigated not only to illustrate the advantages of each layer of the proposed Bayesian formulation of the LW-PLS regression problem, but also to clearly demonstrate the integration mechanism adopted in the developed hierarchial Bayesian optimization framework.

Chapter 3

Bayesian Just-in-time Modeling with the Input-output Similarity Function

3.1 Introduction

Soft sensors have proved to be useful for the task of on-line prediction, process monitoring and fault detection. In many industrial processes, critical variables cannot be easily measured on-line [1, 26, 27]. They have to be obtained from hardware analyzers which are often expensive and difficult to maintain, or through off-line laboratory analysis which cannot be used in real time control. Soft sensors are the key technology for estimating these hard-to-measure process and quality variables in real time. There are two types of modeling methods for soft sensor design, namely, model-driven and data-driven methods. The model-driven method takes advantage of mechanism of underlying processes which are usually difficult to acquire and/or intractable for modeling. In the absence of process knowledge, data-driven may be applied to developing the model. These models are trained on collected data by means of statistic modeling techniques. The most popular data-driven techniques are principal component regression (PCR) and partial least squares (PLS). They have been successfully applied to the development of soft sensors for industrial processes[28, 29, 30, 31].

Although these techniques have gained popularity in industry, development and maintenance of soft sensors are laborious[32]. Most data-driven techniques build only a single global model based on historical data to deal with process in different operating conditions. Even if a good soft sensor is obtained at the beginning, the

performance will deteriorate due to the multi-mode and time-varying behaviours. In order to maintain the performance, soft sensors need to be regularly updated off-line which is a time-consuming effort. To deal with the problem, recursive versions of these methods are developed to update the soft sensor on-line and automatically, such as recursive PCR, recursive PLS and recursive prediction error method (PEM)[33, 34, 35, 36]. However, these recursive methods are meant to deal with slow drifts in parameters but not abrupt changes (which often accompany operating condition changes in non-linear systems). Due to these considerations, Just-in-time (JIT) modeling was proposed to deal with both multi-mode and nonlinear behaviour of the process[37, 38, 39]. Instead of building a global model, the JIT modeling method builds a localized model (using the most relevant data) whenever new query samples become available. In this way, the JIT local models create a piece-wise approximation of the non-linear model which can also deal with multi-mode process behaviour

However, the performance of traditional JIT models is not always satisfactory. It is determined by following aspects: local calibration sample, model structure and model parameters. In traditional JIT modeling framework, such as locally weighted PLS (LW-PLS), the local calibration samples are selected only based on the distance in input space. Even though the local calibration samples are close to the query sample with respect to the input space, they may not be close to the query with respect to the output space. The local calibration samples should be determined by taking account the information in both input and output spaces. Second, the model structure of each local model is assumed to be known beforehand and is kept fixed. Third, traditional approaches, such as OLS, PCR, PLS to estimate local model parameters, do not fully utilize available information. For such applications, information within the historical data and prior process knowledge cannot be fully incorporated into the parameter estimation. Chen [40] proposed an Orthogonal Signal Correction (OSC) based LW-PLS method . The main idea of OSC-LW-PLS is to filter input space by removing the information uncorrelated to the output space (using OSC); after that, the similarity is calculated based on filtered data. However, this method leaves the second and third problems unaddressed. Several optimization methods are applied to obtain a globally optimal model structure resulting in the lowest root mean square error of cross validation (RMSECV)[15]. But, it is possible that the optimal model structures

Modeling technique	Handle Collinearity	Handle non-linearity and time-varying behaviour	Accounts for input-output in Similarity	Automatically select model structure for each local model	Incoporate prior knowledge over model structure and parameters
OLS	No	No	N/A	N/A	No
PCR	Yes	No	N/A	No	No
PLS	Yes	No	N/A	No	No
LW-PLS	Yes	Yes	No	No	No
OSC-LW-PLS	Yes	Yes	Yes	No	No
Bayesian JIT	Yes	Yes	Yes	Yes	Yes

Table 3.1: Features of different data-driven modeling methods

of local models are different from each other and from the global one. It is desired to find the optimal model structure for each local model in a real time identification phase.

The features of these modeling techniques are summarized in Table 3.1. None of the techniques mentioned above can address all the problems simultaneously. Especially, none of the techniques can provide a way to search for a proper model structure and incorporate prior knowledge into estimation. To cope with these challenges, a new JIT modeling method is proposed in this work. In the proposed method, local calibration samples are specified by a new similarity function which can extract hidden information in the inputs and outputs. As a result, noise in input will have less negative impact and information in output can be utilized in identification more effectively. Next, the problem of searching for the optimal model structure for each local model and estimating the corresponding model parameters is formulated under an iterative hierarchical Bayesian optimization framework. This Bayesian framework offers a systematic way to search for the optimal model structure as well as estimate the model parameters. Bayes' theorem also makes it possible to incorporate statistical information from historical data and prior process knowledge which can further enhance the accuracy of prediction.

3.2 Problem Statement

Suppose we have a training (calibration) data-set with N samples denoted by:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_N]^T \tag{3.1}$$

$$\mathbf{y} = [y_1, y_2, \cdots y_N]^T \tag{3.2}$$

$\mathbf{X} \in \mathbb{R}^{N \times M}$ and $\mathbf{y} \in \mathbb{R}^{N \times 1}$ are the input and output matrices, respectively. The i-th sample consists of inputs, $\mathbf{x}_i = [x_{i1}, x_{i2} \cdots x_{iM}]^T$, and an output, y_i. where M is the number of input variables. The model takes the form of the PLS model structure, given by:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_X \tag{3.3}$$

$$\mathbf{y} = \mathbf{T}\mathbf{q}^T + \mathbf{e}_y \tag{3.4}$$

where $\mathbf{T} \in \mathbb{R}^{N \times H}$ denotes a matrix of latent variables, and H denotes the number of latent variables. $\mathbf{P} \in \mathbb{R}^{M \times H}$ is a matrix of loadings and $\mathbf{q} \in \mathbb{R}^{1 \times H}$ is a vector of regression coefficients. $\mathbf{E}_X \in \mathbb{R}^{N \times M}$ and $\mathbf{e}_y \in \mathbb{R}^{N \times 1}$ denote the matrices of input and output residuals, respectively.

JIT modeling is an on-line identification method under which data-driven modeling techniques can be applied in a local modeling perspective. When a query sample $\mathbf{x}_q$ becomes available, the solution of on-line identification of a local model consists of the following main steps:

1. Selection of local calibration samples: Search for the most relevant samples in historical data-set using a pre-defined similarity function. The similarity function should account for the information in both input and output spaces.

2. Selection of model structure: after specifying the local calibration samples, the next key step is to determine a proper model structure to capture the underlying behaviour of the process. In order to select a proper model structure for PLS model, the number of latent variables H needs to be selected. The number of latent variables selected affects the model complexity.

3. Estimation of model parameters: Once the model structure is determined, data-driven modeling techniques are applied to estimate the model parameter, $\Theta = \{\mathbf{P}, \mathbf{T}, \mathbf{q}\}$, *i.e.* loading matrix $\mathbf{P}$, latent variable matrix $\mathbf{T}$ and regression coefficient vector $\mathbf{q}$.

3.3 Bayesian JIT Modeling with Input-output Similarity

3.3.1 Similarity Function

The similarity function plays a key role in JIT modeling. It determines the method to select local calibration samples. A proper similarity function can choose representative calibration samples for modeling, thus improving the accuracy of prediction.

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

$$\mathbf{S}_q = diag(s_{1|q}, s_{2|q}, \cdots, s_{N|q}) \tag{3.5}$$

where $s_{i|q}$ $(i = 1, 2 \cdots N)$ is the similarity between $\mathbf{x}_q$ and $\mathbf{x}_i$.

Generally, a measurement of similarity is defined based on a notion of distance between $\mathbf{x}_q$ and $\mathbf{x}_i$. One of the widely used similarity functions is :

$$s_{i|q} = \exp\left(-\frac{d_i}{\sigma_d \lambda}\right) \tag{3.6}$$

$$d_i = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T (\mathbf{x}_i - \mathbf{x}_q)} \tag{3.7}$$

where d_i is the Euclidean distance between $\mathbf{x}_q$ and $\mathbf{x}_i$, σ_d is the standard deviation of $\mathbf{d} = \{d_1, d_2, \cdots d_N\}$ and λ is the localization parameter. By tuning the localization parameter, we can control how quick the similarity will decrease with increasing distance, which controls the degree at which local samples are prioritized. Furthermore, in order to balance the weight of each input, Mahalanobis distance is applied. The Mahalanobis distance between $\mathbf{x}_q$ and $\mathbf{x}_i$ is calculated using the following equation:

$$d_i = \sqrt{(\mathbf{x}_i - \mathbf{x}_q)^T \mathbf{Q}^{-1} (\mathbf{x}_i - \mathbf{x}_q)} \tag{3.8}$$

where $\mathbf{Q}$ is the covariance matrix of input $\mathbf{X}$. The drawback of this method is that it only applies information from the input space to measure similarity. If the information in output space can be incorporated into similarity measurement, the representativeness of local calibration sample and accuracy of prediction may be improved. Wang [41] proposed a new similarity measurement which takes both input and output information into account. The similarity is calculated as:

$$s_{i|q} = \omega d_{i,x} + (1 - \omega) d_{i,y} \tag{3.9}$$

$$d_{i,y} = \frac{|y_i - \hat{y}_q|}{\sum_{i=1}^{N} |y_i - \hat{y}_q|} \tag{3.10}$$

where $d_{i,x}$ is the Euclidean distance in the input space and $\hat{y}_q$ is the estimated output of the query sample obtained by an initial global PLS model. ω is a hyperparameter to tune the weight of input and output distance. In this way, the information of output can be incorporated into similarity calculation. However, the existing similarity cannot control the degree at which local samples are prioritized. Considering these points, in this work, a new similarity function is developed by synthesizing these two similarity functions. It is defined as:

$$s_{i|q} = \exp\left(-\frac{d_i}{\sigma_d \lambda}\right) \tag{3.11}$$

$$d_i = \omega d_{i,x} + (1-\omega) d_{i,y} \tag{3.12}$$

$$d_{i,x} = \frac{\tilde{d}_{i,x}}{\sum_{i=1}^{N} \tilde{d}_{i,x}} \tag{3.13}$$

$$d_{i,y} = \frac{|y_i - \hat{y}_q|}{\sum_{i=1}^{N} |y_i - \hat{y}_q|} \tag{3.14}$$

where $\tilde{d}_{i,x}$ is the Euclidean distance in the input space and $\hat{y}_q$ is the estimated output of the query sample obtained by an initial global PLS model. In this similarity function, there are two hyperparameters ω, λ which make the similarity more flexible to specific cases. The parameter ω can be tuned to balance the information in input and output space. The localization parameter λ can control the prioritization of calibration samples.

3.3.2 Bayesian JIT Modeling

Hierarchical Bayesian Optimization Framework

Under the Bayesian framework, the problem is converted to maximizing the joint posterior probability function of model parameters and model structure, *i.e.* $p(\Theta, H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q)$. Because it is intractable to maximize $p(\Theta, H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q)$ directly. The problem can be formulated and solved under an iterative Hierarchical Bayesian Optimization framework. Based on the chain rule of probability theory, the joint posterior probability

function is expanded as:

$$p(\Theta, H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q) = p(\Theta|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q)p(H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q) \tag{3.15}$$

Now, the maximization of the posterior probability function can be transformed as a hierarchical optimization problem:

$$\begin{aligned} &\max_{\Theta,H} [p(\Theta|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q)p(H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q)] \\ &= \max_{H} \left\{ p(H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q) \left[\max_{\Theta} p(\Theta|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) \right] \right\} \end{aligned} \tag{3.16}$$

Inference of Model Parameters

Applying Bayes' rule, the posterior PDF of model parameters can be written as:

$$p(\Theta|H, \mathbf{X}, \mathbf{y}, \mathbf{x}_q) \propto p(\mathbf{X}, \mathbf{y}|H, \Theta, \mathbf{x}_q)p(\Theta|H, \mathbf{x}_q) \tag{3.17}$$

By expanding Θ, we have

$$p(\mathbf{P}, \mathbf{T}, \mathbf{q}|\mathbf{X}, \mathbf{y}, \mathbf{x}_q, H) \propto p(\mathbf{X}, \mathbf{y}|\mathbf{P}, \mathbf{T}, \mathbf{q}, \mathbf{x}_q, H)p(\mathbf{T}|\mathbf{P}, H)p(\mathbf{q}|\mathbf{P}, H)p(\mathbf{P}|H) \tag{3.18}$$

Following the approach of Bayesian Latent Variable regression [21], a new Bayeisan approach to solve the problem of JIT modeling is derived next.

For each calibration sample, the LW-PLS formulation is given by:

$$\mathbf{x}_i = \mathbf{P}\mathbf{t}_i + \mathbf{e}_{xi} \tag{3.19}$$

$$y_i = \mathbf{q}\mathbf{t}_i + e_{yi} \tag{3.20}$$

The noise-free inputs and output are given by:

$$\tilde{\mathbf{x}}_\mathbf{i} = \mathbf{P}\mathbf{t}_i \tag{3.21}$$

$$\tilde{y}_i = \mathbf{q}\mathbf{t}_i \tag{3.22}$$

The loading matrix $\mathbf{P}$ has the following constraint:

$$\mathbf{P}^T\mathbf{P} = \mathbf{I} \tag{3.23}$$

A vector of model parameters, $\mathbf{b} \in \mathbb{R}^{M\times 1}$, representing the relationship between the input and output variables, is defined as:

$$\mathbf{b} = \mathbf{P}\mathbf{q}^\mathbf{T} \tag{3.24}$$

The likelihood function relies on the nature of noise. Assume that the input and output measurements are contaminated by mutually independent Gaussian noise, $\mathbf{e}_{xi}$ and e_{yi}, with known variance $\mathbf{Q}_{e_x}$ and Q_{e_y}. The estimation of these variances will be discussed shortly. Given a query sample $\mathbf{x}_q$, the importance weight assigned to the i^{th} calibration sample is denoted by $s_{i|q}$. This is equivalent to saying that:

$$\mathbf{Q}_{e_{xi}} = \frac{\mathbf{Q}_{e_x}}{s_{i|q}} \tag{3.25}$$

$$Q_{e_{yi}} = \frac{Q_{e_y}}{s_{i|q}} \tag{3.26}$$

where $s_{i|q}$ can be calculated using the new similarity function (Eqns. 3.11, 3.12, 3.13 and 3.14). Since both the output and input information is integrated in Eqn.3.12, the input and output will both participate to determine the similarity. A calibration sample having small distances in both input and output dimensions will yield large similarity. This will result in a small noise variance, meaning that the point contains less information for noise, equivalently more information for identification. Note that if a sample point has a small distance from query sample in the input dimension, but a large distance in the output dimension, this will result in large noise variance under this formulation of the similarity function so that this data point will be discounted in the subsequent identification process.

It is assumed that the measurement noises in the observations are independent in the time sequence and the measurement noises in inputs and output are mutually independent. Thus, the likelihood can be simplified as follows:

$$p(\mathbf{X}, \mathbf{y}|\mathbf{P}, \mathbf{T}, \mathbf{q}, \mathbf{x}_q, H) = p(\mathbf{X}|\mathbf{P}, \mathbf{T}, \mathbf{x}_q, H)p(\mathbf{y}|\mathbf{T}, \mathbf{q}, \mathbf{x}_q, H) \tag{3.27}$$

$$p(\mathbf{X}|\mathbf{P}, \mathbf{T}, \mathbf{x}_q, H) = \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{P}, \mathbf{t}_i, \mathbf{x}_q, H) \tag{3.28}$$

$$p(\mathbf{y}|\mathbf{T}, \mathbf{q}, \mathbf{x}_q, H) = \prod_{i=1}^{N} p(y_i|\mathbf{t}_i, \mathbf{q}, \mathbf{x}_q, H) \tag{3.29}$$

$$\mathbf{x}_i|\mathbf{P}, \mathbf{t}_i, \mathbf{q}, \mathbf{x}_q, H \sim \mathcal{N}(\mathbf{P}\mathbf{t}_i, \frac{\mathbf{Q}_{e_x}}{s_{i|q}}) \tag{3.30}$$

$$y_i|\mathbf{P}, \mathbf{t}_i, \mathbf{q}, \mathbf{x}_q, H \sim \mathcal{N}(\mathbf{q}\mathbf{t}_i, \frac{Q_{e_y}}{s_{i|q}}) \tag{3.31}$$

The priors over the model parameters depend on the distribution of noise-free data. The noise-free inputs are assumed to follow a multivariate Gaussian distribution, that

is

$$\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mu_{\mathbf{x}}, \mathbf{Q}_{\mathbf{x}}) \tag{3.32}$$

As a result, given the loading matrix $\mathbf{P}$, the latent variable $\mathbf{t_i}$ will also follow a multivariate Gaussian distribution:

$$\mathbf{t_i} = \mathbf{P}^T \tilde{\mathbf{x}}_i \tag{3.33}$$

$$\mathbf{t_i}|\mathbf{P}, H \sim \mathcal{N}(\mathbf{P}^T \mu_x, \mathbf{P}^T \mathbf{Q}_x \mathbf{P}) \tag{3.34}$$

Let the regression parameters $\mathbf{b}$ follow a multivariate Gaussian distribution:

$$\mathbf{b} \sim \mathcal{N}(\mu_b, \mathbf{Q}_b) \tag{3.35}$$

Given the loading matrix $\mathbf{P}$, and the vector of model parameters $\mathbf{b}$, the regression coefficient vector $\mathbf{q}^T$ will also follow a multivariate Gaussian distribution

$$\mathbf{q}^T = \mathbf{P}^T \mathbf{b} \tag{3.36}$$

$$\mathbf{q}^T|\mathbf{P}, H \sim \mathcal{N}(\mathbf{P}^T \mu_b, \mathbf{P}^T \mathbf{Q}_b \mathbf{P}) \tag{3.37}$$

In the absence of any priori knowledge over the loading matrix $\mathbf{P}$, a uniform prior distribution can be specified.

The maximum a posteriori probability (MAP) estimates can then be obtained by solving the following optimization problem:

$$\begin{aligned} &\{\mathbf{P}, \mathbf{T}, \mathbf{q}\}_{MAP} = \arg\max_{\mathbf{P},\mathbf{T},\mathbf{q}} \{p(\mathbf{X}|\mathbf{P}, \mathbf{T}, \mathbf{x}_q, H) p(\mathbf{y}|\mathbf{T}, \mathbf{q}, \mathbf{x}_q, H) p(\mathbf{T}|\mathbf{P}, H) p(\mathbf{q}|\mathbf{P}, H)\} \\ &s.t.\ \ \mathbf{P}^T\mathbf{P} = \mathbf{I} \end{aligned} \tag{3.38}$$

The solution can be obtained by solving the following three simultaneous parameter-estimation and data-reconciliation optimization problems, where outer optimization accounts for estimation of the model parameters and inner optimization accounts for estimation of the latent variables.

$$\begin{aligned} &\{\mathbf{P}\}_{MAP} = \arg\max_{\mathbf{P}} p(\mathbf{X}|\mathbf{P}, \mathbf{T}, H, \mathbf{x}_q) p(\mathbf{y}|\mathbf{T}, \mathbf{q}, H, \mathbf{x}_q) \\ &\{\mathbf{q}\}_{MAP} = \arg\max_{\mathbf{q}} p(\mathbf{y}|\mathbf{T}, \mathbf{q}, H, \mathbf{x}_q)\ p(\mathbf{q}|\mathbf{P}, H) \\ &s.t. \\ &\{\mathbf{T}\}_{MAP} = \arg\max_{\mathbf{T}} p(\mathbf{X}|\mathbf{P}, \mathbf{T}, H, \mathbf{x}_q) p(\mathbf{T}|\mathbf{P}, H) \\ &\mathbf{P}^T\mathbf{P} = \mathbf{I} \end{aligned} \tag{3.39}$$

All posteriors follow multivariate Gaussian distribution and thus, the MAP estimates can be equivalently obtained by solving the following minimization problems:

$$\begin{aligned}
\{\mathbf{P}\}_{MAP} &= \arg\min_{\mathbf{P}}\{\sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)^T(\tfrac{\mathbf{Q}_{ex}}{s_{i|q}})^{-1}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)+\sum_{i=1}^{N}(y_i-\mathbf{q}\mathbf{t}_i)^T(\tfrac{\mathbf{Q}_{ey}}{s_{i|q}})^{-1}(y_i-\mathbf{q}\mathbf{t}_i)\}\\
\{\mathbf{q}\}_{MAP} &= \arg\min_{\mathbf{P}}\{\sum_{i=1}^{N}(y_i-\mathbf{q}\mathbf{t}_i)^T(\tfrac{\mathbf{Q}_{ey}}{s_{i|q}})^{-1}(y_i-\mathbf{q}\mathbf{t}_i)+(\mathbf{q}^T-\mathbf{P}^T\mu_b)^T(\mathbf{P}^T\mathbf{Q}_b\mathbf{P})^{-1}(\mathbf{q}^T-\mathbf{P}^T\mu_b)\}\\
\{\mathbf{t}_i\}_{MAP} &= \arg\min_{\mathbf{t}_i}\{(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)^T(\tfrac{\mathbf{Q}_{ex}}{s_{i|q}})^{-1}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i) \quad +(\mathbf{t}_i-\mathbf{P}^T\mu_{\mathbf{x}})^T(\mathbf{P}^T\mathbf{Q}_{\mathbf{x}}\mathbf{P})^{-1}(\mathbf{t}_i-\mathbf{P}^T\mu_{\mathbf{x}})\\
s.t. \quad & \mathbf{P}^T\mathbf{P}=\mathbf{I}
\end{aligned} \tag{3.40}$$

The first optimization problem is intractable because of the unit orthonormal constraint. We can first use a closed form optimization solution, to estimate $\mathbf{P}$. Both of the following optimization problems can be solved analytically.

$$\{\mathbf{t}_i\}_{MAP} = [\mathbf{P}^T(\frac{\mathbf{Q}_{e_x}}{s_{i|q}})^{-1}\mathbf{P}+(\mathbf{P}^T\mathbf{Q}_{\mathbf{x}}\mathbf{P})^{-1}]^{-1}[\mathbf{P}^T(\frac{\mathbf{Q}_{e_x}}{s_{i|q}})^{-1}\mathbf{x}_i+(\mathbf{P}^T\mathbf{Q}_{\mathbf{x}}\mathbf{P})^{-1}\mathbf{P}^T\mu_{\mathbf{x}}] \tag{3.41}$$

$$\{\mathbf{q}\}_{MAP} = (\mathbf{T}^T\mathbf{T}+\mathbf{P}^T\mathbf{Q}_b\mathbf{P}S_q{Q_{e_y}}^{-1})(\mathbf{T}^TY+\mathbf{P}^T\mathbf{Q}_b\mathbf{P}S_q{Q_{e_y}}^{-1}\mathbf{P}^T\mu_b) \tag{3.42}$$

In the above derivations, $\mathbf{Q}_{e_x}, Q_{e_y}, \mu_b, \mathbf{Q}_b, \mu_x, \mathbf{Q}_x$ are assumed to be known. This requirement means that the prior density must be fully specified in advance. In the presence of limited prior knowledge over the noise variance and model parameters, a commonly used alternative is the empirical Bayesian analysis which estimates the prior from the available data. Again, the parametric approach is used to estimate the prior. The approach is presented as *Algorithm II* in the appendix.

Inference of Model Structure

Applying Bayes rule, the posterior PDF of hyperparameter can be expressed as:

$$p(H|\mathbf{X},\mathbf{y},\mathbf{x}_q) \propto p(\mathbf{X},\mathbf{y}|H,\mathbf{x}_q)p(H|\mathbf{x}_q) \tag{3.43}$$

When considering the prior of the model structure, it is reasonable to assume that it is statistically independent of the query sample $\mathbf{x}_q$. If a set of candidate model structures are given, i.e $H \in \{H_1, H_2...H_L\}$, then the random variable H is a

categorical variable and can be modelled by

$$p(H) = \prod_{l=1}^{L} p(H = H_l)^{[H=H_l]} \tag{3.44}$$

where $[H = H_l]$ equals 1 if $H = H_l$ and equals 0 otherwise. Without any prior information, a uniform distribution, i.e. $p(H = H_l) = ... = p(H = H_L)$ is adopted.

The likelihood function $p(\mathbf{X}, \mathbf{y}|H, \mathbf{x}_q)$ can be obtained by integrating over the model parameter:

$$p(\mathbf{X}, \mathbf{y}|H, \mathbf{x}_q) = \int_{\Theta} p(\mathbf{X}, \mathbf{y}|\Theta, H, \mathbf{x}_q) p(\Theta) d\Theta \tag{3.45}$$

Since it is intractable to solve this equation directly, Laplace's method of approximation is applied here:

$$\int_{\Theta} p(\mathbf{X}, \mathbf{y}|\Theta, H, \mathbf{x}_q) p(\Theta) d\Theta \approx p(\mathbf{X}, \mathbf{y}|\Theta^{MAP}, H, \mathbf{x}_q) p(\Theta^{MAP}) \det (\frac{A_\Theta}{2\pi})^{-\frac{1}{2}} \tag{3.46}$$

where $A_\Theta = -\frac{\partial^2}{\partial \Theta^2} \log p(\Theta|\mathbf{X}, \mathbf{y}, \mathbf{x}_q, H)$. Then the MAP estimate of model structure can be derived as:

$$\begin{aligned} \{H\}_{MAP} &= \arg\max_H p(H|\mathbf{X}, \mathbf{y}, \mathbf{x}_q) \\ &= \arg\max_H \left\{ p(\mathbf{X}, \mathbf{y}|\Theta^{MAP}, H, \mathbf{x}_q) p(\Theta^{MAP}|H) \det (\frac{A_\Theta}{2\pi})^{-\frac{1}{2}} p(H) \right\} \end{aligned} \tag{3.47}$$

$$\{H\}_{MAP} = \arg\min_H \left\{ \begin{array}{l} \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{P}\mathbf{t}_i)^T (\frac{\mathbf{Q}_{e_x}}{s_{i|q}})^{-1} (\mathbf{x}_i - \mathbf{P}\mathbf{t}_i) \\ + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{q}\mathbf{t}_i)^T (\frac{Q_{e_y}}{s_{i|q}})^{-1} (y_i - \mathbf{q}\mathbf{t}_i) \\ + \frac{1}{2} \sum_{i=1}^{n} (\mathbf{t}_i - \mathbf{P}^T \mu_x)^T (\mathbf{P}^T \mathbf{Q}_x \mathbf{P})^{-1} (\mathbf{t}_i - \mathbf{P}^T \mu_x) \\ + \frac{1}{2} (\mathbf{q}^T - \mathbf{P}^T \mu_b)^T (\mathbf{P}^T \mathbf{Q}_b \mathbf{P})^{-1} (\mathbf{q}^T - \mathbf{P}^T \mu_b) \\ + \log[\det (\frac{\mathbf{A}_\Theta}{2\pi})^{-\frac{1}{2}}] + \log \prod_{i=1}^{N} s_{i|q} \\ + \frac{1}{2}(1 + N) H \log 2\pi \end{array} \right\} \tag{3.48}$$

3.4 Implementation Procedure

1. Choose Eqn 3.11-3.14 as similarity function.

2. Select a proper set of candidate model structures $\{H_1, H_2...H_L\}$.

3. Characterize the localization parameter λ and balance parameter ω by Leave One Out Cross Validation (LOOCV).

4. Characterize the noise variances $\mathbf{Q}_{e_x}$, Q_{e_y} and prior distribution of model parameters, $p(\Theta|H)$ by using *Algorithm II*.

5. Characterize the prior distribution of model structure, $p(H)$, based on prior knowledge. If there is no prior information over model structure, a uniform distribution can be used to describe the prior.

6. For $l = 0 : L$

   (1). Set $H = H_l$.

   (2). While $\mathbf{P}_l, \mathbf{T}_l, \mathbf{q}_l$ converge

   (2.2). Calculate loading matrix $\mathbf{P}_l$ by applying LW-PLS algorithm to $(\mathbf{X}, \mathbf{y}, \mathbf{x}_q)$.

   (2.3). Calculate regression coefficient vector $\mathbf{q}_l$ and latent variable matrix $\mathbf{T}_l$ by using Eqn 3.41 and Eqn 3.42;

   (3). Calculate objective function of model structure H_l by using Eqn 48.

7. Choose the model structure H with the lowest value of objective function as well as corresponding loading matrix $\mathbf{P}$, regression coefficient vector $\mathbf{q}$.

8. Calculate model output as $\hat{y} = \mathbf{x}_q\mathbf{P}\mathbf{q}^T$.

3.5 Case Study

To illustrate the advantages of Bayesian JIT modeling, a set of Pharmacy NIR data of Escitalopram tablets is used. The objective is to develop a reliable model for predicting the active substance content in Pharmaceutical tablets from the NIR spectra

of samples. This NIR data set is a public benchmark for multivariate data analysis which has high dimension with strongly correlated spectra. The data has been used by [40, 42]. This data-set consists of NIR spectra for 310 tablet samples. As stated by [42], the tablet samples have four different dosage values (5, 10, 15 and 20 mg tablets) and samples in each type come from three different scales of batch processes (full scale, pilot scale and laboratory scale). The variety of samples will result in nonlinear behaviour of the data set. The JIT method is considered to provide an effective solution to this case.

The spectra of 310 tablet samples are divided into 124 calibration samples and 184 validation or test samples. The calibration and test data set consists of samples with reference value ranging from 4.84% to 9.79% and 4.61% to 9.38%, respectively.

The Bayesian JIT model is applied to develop the calibration models for active substance. The advantages of the proposed method is demonstrated by comparing results with other popular methods, *i.e.*,PLS, LW-PLS, OSC-LW-PLS. Furthermore, to illustrate the features of Bayesian JIT model, each layer of Bayesian JIT is applied individually to show its necessity and advantage:

The following methods are applied and compared to build NIR spectroscopic model.

1. PLS: Global PLS approach.

2. LW-PLS: locally weighted PLS; the similarity is calculated based on distance in input space only.

3. OSC-LW-PLS: OSC based locally weighted PLS proposed by [40];

4. JIT-I: JIT approach using the new similarity function (Eqn. 3.11 3.12 3.13 and 3.14) and then using LW-PLS to calculate the model parameters Θ with fixed model structure H;

5. JIT-II: JIT approach using the new similarity function (Eqn. 3.11 3.12 3.13 and 3.14) to calculate similarity and then using Bayesian approach to calculate model parameters with fixed model structure H.

6. Bayesian-JIT: Bayesian Just-in-time approach proposed in this work; the similarity is calculated by a new function and integrated in a hierarchical Bayesian Optimization framework which can automatically search for the optimal model structure H for each local model and estimate the model parameters Θ.

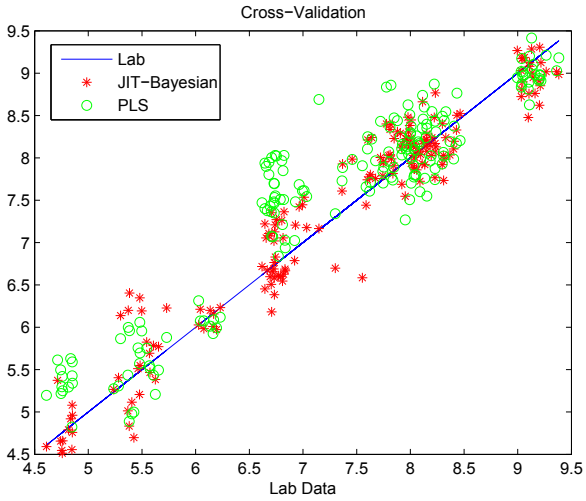
Before applying these methods, the number of LVs and some hyper parameters need to be specified. In this study, the optimal hyper parameters for the last five methods and the number of latent variables for the first five methods shown in the table are determined by Leave One Out Cross-Validation (LOOCV).

Methods	H	λ	ω	OSC factors	RMSEP	R
PLS	7	-	-	-	0.5188	0.9334
LW-PLS	5	0.85	-	-	0.4095	0.9537
OSC-LW-PLS	4	1.1	-	3	0.3499	0.9635
JIT-1	4	0.1	0.6	-	0.3591	0.9619
JIT-2	4	0.18	0.55	-	0.3300	0.9668
Bayesian JIT	2-5	0.18	0.55	-	0.3199	0.9708

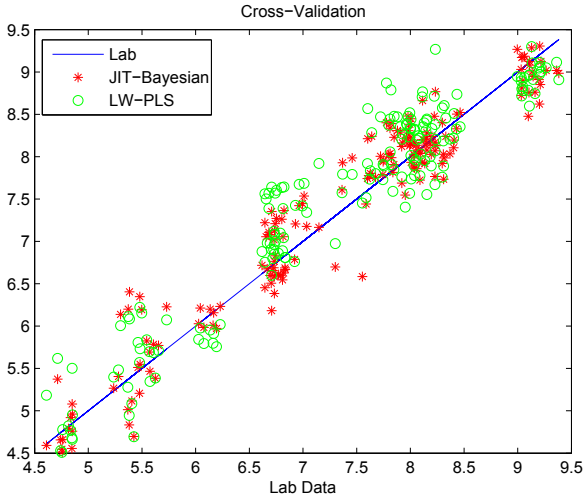
Table 3.2: Comparing prediction performance of different methods using pharmacy tablets data-set

According to Table 3.2 and Figure 3.1, all the JIT methods outperform the global PLS, meaning that JIT methods are more effective when dealing with non-linearity in data. JIT-OSC, JIT-1, JIT-2 and Bayesian JIT, which take advantage of both input and output information when calculating the similarity, have superior performance to the traditional LW-PLS. Among all the tested methods, Bayesian JIT has the best performance with lowest RMSE and highest R. It is interesting to see that JIT-1, JIT-2 and Bayesian JIT perform better consistently which shows that the improvement of each step in JIT modeling, *i.e.*, selection of local calibration samples, selection of model structures and estimation of model parameters, results in significant improvement in predictive ability of the model.

In order to further investigate the generalization of this method, a different case study was performed on the same 310 tablet samples. In this case study, the 124 samples(which are set as calibration samples) are transferred to the set of test samples. The remaining 184 samples are used for calibration. Bayesian JIT and other popular data-driven methods are applied to develop calibration models for active substances



(a) PLS vs Bayesian JIT



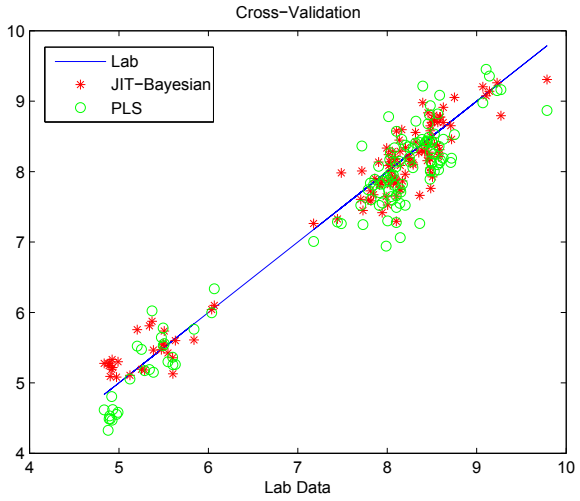
(b) LW-PLS vs Bayesian JIT

Figure 3.1: Cross-validation using data-set from pharmaceutical tablets, case I

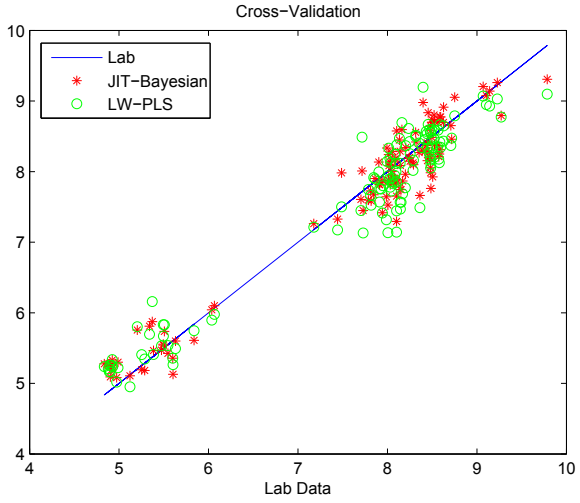
in pharmaceutical tablets. Again, the hyper parameters for the last five methods and the number of LVs for the first five methods are specified by LOOCV.

Methods	H	λ	ω	OSC factors	RMSEP	R
PLS	9	-	-	-	0.3817	0.9652
LW-PLS	7	0.4	-	-	0.3313	0.9693
OSC-LW-PLS	4	0.2	-	3	0.3332	0.9679
JIT-1	5	0.2	0.5	-	0.2945	0.9749
JIT-2	5	0.2	0.5	-	0.2841	0.9762
Bayesian JIT	2-5	0.2	0.4	-	0.2825	0.9762

Table 3.3: Comparing prediction performance of different methods using pharmacy tablets data-set



(a) PLS vs JIT-Bayesian



(b) LW-PLS vs JIT-Bayesian

Figure 3.2: Cross-validation using data-set from pharmaceutical tablets, case II

From Table 3.3 and Figure 3.2 we can see that the performance of JIT methods are again much better than the global PLS method. This proves the effectiveness of JIT method in this case. Moreover, the JIT methods which used both input and output information in similarity calculations outperform others. Among them, Bayesian JIT achieves the best performance with lowest RMSEP and highest R. The result also shows that JIT-1, JIT-2 and Bayesian JIT perform consistently better which again proves that improvement of each step of JIT modeling can lead to improvement in the predictive ability of the model.

3.6 Conclusion

In this work, a new Just-in-time (JIT) modeling approach is proposed to achieve higher accuracy of prediction. The proposed method has the following advantages over regular JIT method: 1. It takes both input and output information into account when calculating similarities. 2. A Bayesian optimization framework is proposed for real-time selection of local model structure and estimation of corresponding model parameters. This optimization framework offers a systematic way to search for the optimal model structure for each local model. 3. Using the Bayesian method also makes it possible to incorporate the statistical information in historical data and prior process knowledge into the estimation procedure, which further enhances the accuracy of prediction. The advantages of the proposed approach were illustrated through a case study based on real-world NIR data from pharmaceutical industry. Multiple modeling approaches, *i.e*, PLS, LW-PLS, JIT-OSC along with the proposed Bayesian JIT method were applied to estimate the content of active substance in tablet from spectra data. Compared with traditional PLS, regular JIT (LW-PLS) methods and the improved JIT method (OSC-LW-PLS), the proposed approach achieved the best performance in both case studies.

Chapter 4

Recursive Prediction Error Method and Its Application in Adaptive Soft Sensors Design

4.1 Introduction

Soft sensors have proved to be an effective alternative to traditional approach for the acquisition of critical process variables[43, 44, 45, 46, 47]. There are generally two kinds of soft sensors, namely model-driven and data-driven soft sensor [48]. The model-driven soft sensors take advantage of first principles which describe the physical and chemical phenomena of the process, such as mass balance, thermal balance and reaction kinetics. One drawback of the first-principles technique is that this method requires a lot of expert knowledge about the process, and is often intractable. Thus, as an effective alternative, data-driven soft sensors have gained increasing popularity in industry [1]. Since it is based on the historical data from operating plants, a data-driven model can be developed more quickly. However, data-driven soft sensors are less reliable, because the data themselves cannot fully explain the underlying mechanism of the process. Grey-box soft sensors make use of first-principles knowledge, and black-box techniques to fill the knowledge gap [49]. It is developed based on the knowledge describing the chemical and physical principles underlying the process as well as the statistical information from the data. A typical example using this modeling technique is a model-driven soft sensor making use of the data-driven approach to identify the unknown portions which cannot be modelled easily in terms of available process knowledge. These unknown portions are usually treated as the

black-box model which will be identified by using historical data. Since the complex underlying mechanism of process usually results in a nonlinear model structure, some data-driven modeling techniques such as ordinary least squares (OLS), principle component regression (PCR), and partial least squares (PLS) cannot be directly applied for identification. A widely used approach for grey-box modeling is prediction error method (PEM). The main idea underlying the PEM is to minimize the prediction errors so that it can be applied to complex model parameterizations [50].

To develop grey-box soft sensors, the collected on-line measurements, i.e. historical data, can be exploited by PEM for off-line model identification. However, even if a good model is identified initially, the accuracy of a soft sensor is guaranteed for only a specific operating region in which the model has been identified and so the performance of the model will deteriorate over time. This is because, in most cases, historical data cannot contain all the possible future conditions of the process. Furthermore, the process may exhibit a certain form of time-variant behaviour due to fouling and/or abrasion in the process equipment, variation in catalyst activity, changes in weather and so on, which are difficult to take into account during the modeling phase [51].

To deal with these issues, an on-line adaptation is often integrated in the implementation procedure. The procedure for adaptive soft sensor is shown in Figure 1. At first, off-line modeling techniques are used to build initial models based on the historical data. Expert knowledge about the process helps to search for a proper model structure and identify the influential variables as well as time delays. During the operational phase, the real-time predictions of the target properties are generated by the soft sensor based on fast-rate on-line inputs. The reference for target properties is obtained by lab analysis. Together with the corresponding on-line inputs and predictions, they provide sources for on-line adaptation. The effectiveness of on-line adaptation is significantly affected by the following aspects:

1. Adaptive data pre-processing: In on-line application, in order to achieve more accurate estimation of quality variables and effective model updates, both on-line input and lab data need to be preprocessed in an on-line manner. This consists of several operations, such as abnormal point detection, data de-noising

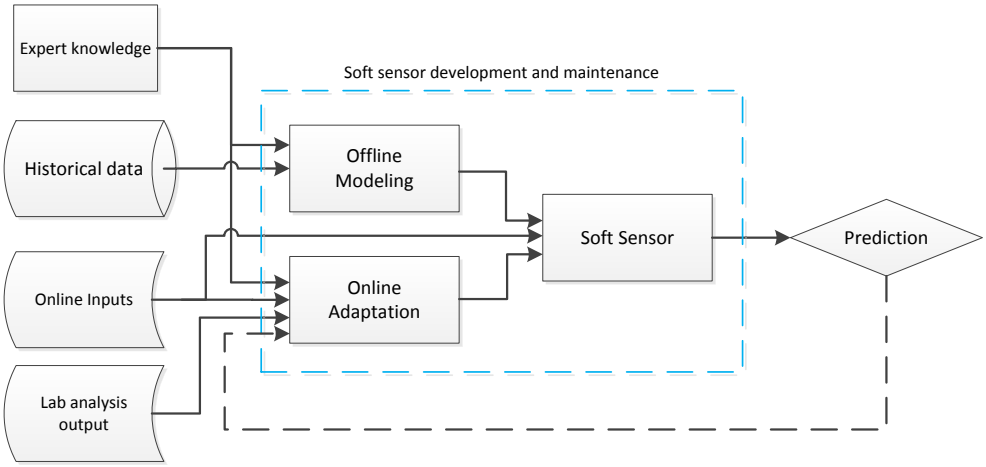


Figure 4.1: Recursive adaptation flowchat

and data scaling.

2. Performance feedback: The need for adaptation should be triggered by monitoring the performance of the soft sensor. There are different criteria to evaluate the performance and determine when to apply adaptation to the soft sensor.

3. Adaptation algorithm: Once the data are available, and there is a need for adaptation, the adaptation algorithm can be applied to adjust the soft sensor.

Despite the increasing number of publications dealing with adaptation of soft sensors, several issues remain open. Most of the existing methods for adaptation are moving window methods or methods involving recursive updating of OLS, PCA and PLS which are designed for adaptation of linear models[52, 35, 53]. They cannot be applied directly to nonlinear models. Moreover, in practice, the second aspect, *i.e.*, performance feedback is often ignored. Instead, the adaptation is performed at a certain frequency or once new lab analysis output is available. This may result in over-updating issues.

Considering these challenges, the recursive prediction error method (RPEM) is adopted in this chapter as an adaptation algorithm for grey-box models which can

calibrate all the model parameters on-line. Based on RPEM, an integrated framework of on-line adaptation for grey-box model is proposed. Several adaptive data pre-processing methods are applied to reduce the negative effects of measurement noise, as well as detect irregular measurements. The cautious update strategy is integrated into the adaptation mechanism to trigger the need for adaptation, thus avoiding over-updating issues. Then, this adaptive framework is applied for adaptive soft sensors designs in oil sands process. The developed soft sensors can cope with the time-varying behaviour of the process as well as process nonlinearity, thereby reducing the burden of model maintenance.

4.2 Theoretical Background

4.2.1 Prediction Error Method

Prediction Error Methods are a broad family of parameter estimation methods. The main idea underlying the PEM is to minimize the one-step-ahead prediction errors:

$$\begin{aligned} \min_{\Theta} \quad J_N(\Theta) &= \frac{1}{N}\sum_{t=1}^{N}\left(y_t^{Lab} - \hat{y}_t\right)^2 \\ &= \frac{1}{N}\sum_{t=1}^{N}\varepsilon(t,\Theta)^2 \end{aligned} \tag{4.1}$$

where $J_N(\Theta)$ is the objective function to be minimized. y_t^{Lab} is the reference value conducted by lab analysis at time "t". $\hat{y}_t$ is the value of predicted variable and N is the number of calibration samples. The scheme of prediction error method is basically a recursive algorithm with a gradient-type iteration:

$$\Theta^{(i+1)} = \Theta^{(i)} - \mu_i\left(R_N\left(\Theta^{(i)}\right)\right)^{-1} J'_N\left(\Theta^{(i)}\right) \tag{4.2}$$

$$\phi(t,\Theta) = \left[\begin{array}{cccc} \frac{\partial\hat{y}_t}{\partial\theta_1} & \frac{\partial\hat{y}_t}{\partial\theta_2} & \cdots & \frac{\partial\hat{y}_t}{\partial\theta_n} \end{array}\right]^T \tag{4.3}$$

$$R_N(\Theta) := \frac{1}{N}\sum_{t=1}^{N}\phi(t,\Theta)\,\phi^T(t,\Theta) + \lambda I \tag{4.4}$$

$$J'_N(\Theta) = -\frac{1}{N}\sum_{t=1}^{N}\phi(t,\Theta)\,\varepsilon(t,\Theta) \tag{4.5}$$

where $\Theta^{(i)}$ is the parameter estimate at the ith iteration, μ_i is the step size parameter to accelerate the algorithm and $J'_N\left(\Theta^{(i)}\right)$ is the gradient of the objective function.

4.2.2 Recursive Prediction Error Method

The recursive prediction error method was proposed by [36] based on the off-line prediction error method. The algorithm is introduced as follows.

First the following term is obtained to represent the initial state.

$$\mathcal{P}(0) = (\frac{1}{N}\sum_{k=1}^{N}\psi(k,\Theta^{[0]})\psi^T(k,\Theta^{[0]}) + \lambda I)^{-1} \tag{4.6}$$

where,

$$\psi(t,\Theta) = \frac{\partial \hat{y}_k(\Theta)}{\partial \Theta} \tag{4.7}$$

N is the number of identification data points, λ is a tuning parameter adjusted to prevent numerical issues , and $\Theta^{[0]}$ is the initial parameter. Let us also introduce the following expressions:

$$\mathcal{S}(t) = 1 + \psi(t,\Theta^{[t-1]})\mathcal{P}(t-1)\psi^T(t,\Theta^{[t-1]}) \tag{4.8}$$

$$\mathcal{P}(t) = \mathcal{P}(t-1) - \mathcal{P}(t-1)\psi^T(t,\Theta^{[t-1]})\mathcal{S}(t)^{-1}\psi(t,\Theta^{[t-1]})\mathcal{P}(t-1) \tag{4.9}$$

The RPEM algorithm starts with the initial value of parameters $\Theta^{[0]}$ and iterates through the following three steps when new lab data is available:

1. When new lab data y_t is available, given the existing estimate of parameters $\Theta^{[t-1]}$, calculate $\mathcal{S}(t)$ from Equation 4.8, and the prediction error $\varepsilon(k,\Theta)$.

2. Update parameters $\Theta^{[t]}$ as follows:

$$\Theta^{[t]} = \Theta^{[t-1]} - \beta\mathcal{P}(t-1)\psi^T(t,\Theta^{[t-1]})\mathcal{S}(t)^{-1}\varepsilon(t,\Theta^{[t-1]}) \tag{4.10}$$

3. Calculate $\mathcal{P}(t)$ from Equation 4.9 for the next iteration.

4.2.3 Evaluation Criteria

RMSE

Root mean squares error of prediction (RMSEP) is a measure of the extent of agreement between a predicted variable and the reference value. Smaller RMSE values imply higher accuracy. They are a consequence of high precision and low bias.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t^{Lab} - \hat{y}_t\right)^2} \tag{4.11}$$

where n is the number of validation samples. Ideally, the root mean of prediction error distributions is equal to zero.

Correlation coefficient

Correlation coefficient is used to describe how correlated the predictions are with reference value (lab data). It can be calculated as:

$$R = \frac{\text{cov}(Y^{Lab}, \hat{Y})}{\sqrt{\text{var}(Y^{Lab})\text{var}(\hat{Y})}} \tag{4.12}$$

where cov (.) and var(.) indicate the covariance and variance, respectively. R can vary between -1 and 1. Values close to 1 or -1 show strong correlation between the two variables. In fact, good predictions should have an R close to 1.

Graphical Techniques

The graphical techniques used in analysis of residuals are listed below:

1. Scatter plot of predicted values versus target values: The ideal case would be for all the data points to lie on the 45 degree line, indicating perfect agreement between the measured/predicted values and target values.

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

4.3 Adaptation Mechanism

This section first outlines the methods which are used to design the scheme of adaptation. Next, a overall framework for on-line adaptation is proposed based on these methods.

4.3.1 Adaptive Data Preprocessing

Real-world data which is collected from process operation is inevitably corrupted by different disturbances in process, malfunctions and errors in sensors and data transmissions. Data corruption will introduce undesired changes to the original data, and thus have a negative effect on modeling and prediction. In off-line model

identification, data preprocessing is introduced to clean the raw data. For on-line implementation, to achieve good estimation of the quality variable and effective model adaptation, it is desired to preprocess the data in an on-line manner. This consists of abnormal point detection and data de-noising.

On-line Abnormal Point Detection

Abnormal points are generated from the process when it behaves in some new operating modes those are not recorded in history. The grey-box models which are identified based on historical data may not provide accurate estimations in those abnormal situations. Instead of providing unreliable estimations, the grey-box models will give "bad-value" alarms in those situations. A widely used method for off-line abnormal points detection is the 3-σ rule which assumes that the process variable follows a Gaussian distribution with mean value μ and standard deviation σ. A data point x_i is labeled as abnormal point if $|x_i - \mu| > 3\sigma$. During on-line implementation, the abnormal point detection and replacement can be implemented in a moving window. The criteria to label an abnormal point are:

$$\begin{aligned} &|x_i - \mu| > 3\delta \\ &\mu = \tfrac{1}{N_w} \sum_{t=i-N_w}^{i-1} x_t \\ &\delta = \sqrt{\tfrac{1}{N_w} \sum_{t=i-N_w}^{i-1} (x_t - \mu)} \end{aligned} \tag{4.13}$$

where N_w is the window size. Once an abnormal point is detected, it will be replaced by some appropriate value. In a continuous process, the abnormal point is replaced by the previous data point, i.e. $x_i = x_{i-1}$.

On-line data de-noising

Measurement noise causes errors in model estimation and therefore has to be dealt with by increasing the Signal to Noise Ratio (SNR) of the data. The easiest way to achieve this goal is to smooth the data using a linear filter. A process variable is smoothed by using a weighted sum of previous measurements.

$$x_i^d = \alpha x_i + (\alpha - 1) x_{i-1} \tag{4.14}$$

where α is an adjustable smoothing parameter with values between 0 and 1, and x_i^d is the de-noised sample. This is called Exponentially Weighted Moving Average (EWMA) filter.

4.3.2 Reliability of Lab Measurements

Although lab measurements are considered to be target values, their reliability and accuracy could be affected by several factors:

1. Sampling Error: Lab samples are manually taken from the process. The exact sampling time may not be recorded.

2. Human Error: Inaccurate or even wrong lab analysis may be carried out due to human error. This possible factor can have significant effect on the quality of lab measurements.

3. Device inaccuracy: Lab devices need to be calibrated regularly. Even when proper calibrations are applied to lab devices, there may still exist certain inaccuracies.

So it is desired to evaluate the reliability of lab data before using them for adaptation. As shown in Figure 4.2, the lab data are fitted by Gaussian distribution. The reliability of lab data can be evaluated based on 3-σ rule:

$$\mu_Y = \frac{1}{N}\sum_{t=1}^{N} y_t \tag{4.15}$$

$$\delta_Y = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \mu_Y)} \tag{4.16}$$

$$d = \left| y^{Lab} - \mu_Y \right| \tag{4.17}$$

$$\mathrm{d} < 3\sigma \rightarrow Reliable \tag{4.18}$$

$$\mathrm{d} > 3\sigma \rightarrow Unreliable \tag{4.19}$$

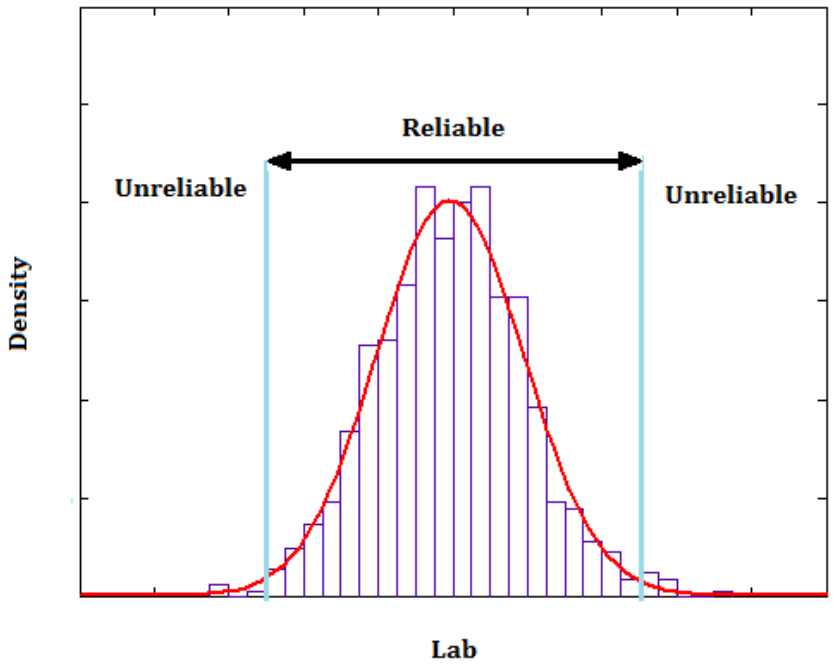


Figure 4.2: Reliability of lab data

4.3.3 Cautious Update

Having preprocessed the on-line measurements and ensured the reliability of lab measurements, recursive prediction error method is applied to implement adaptation. But the adaptation should not be immediately performed, once new lab data is available. As stated in introduction, the need for adaptation should be triggered by monitoring the performance of the soft sensor. So the following criterion is proposed to evaluate the need for adaptation.

Judgement	Execute
$\lvert\Theta^{[t]'} - \Theta^{[t-1]}\rvert > \delta_a$	$T = T + 1, \Theta^{[t]} = \Theta^{[t-1]}$
$\delta_b < \lvert\Theta^{[t]'} - \Theta^{[t-1]}\rvert < \delta_a$ and $0 \leq T \leq 3$	$T = T + 1, \Theta^{[t]} = \Theta^{[t-1]}$
$\delta_b < \lvert\Theta^{[t]'} - \Theta^{[t-1]}\rvert < \delta_a$ and $T > 3$	$T = T + 1, \Theta^{[t]} = \Theta^{[t]'}$
$\lvert\Theta^{[t]'} - \Theta^{[t-1]}\rvert \leq \delta_b$	$T = 0, \Theta^{[t]} = \Theta^{[t]'}$

Table 4.1: Cautious criterion

Each time we get updated $\Theta^{[t]'}$ from previous $\Theta^{[t-1]}$ by recursive PEM, calculate the difference between them, which is called update step size, and compare it with a

certain threshold δ_a, δ_b. T is the frequency count of large errors; a large count would indicate that a large parameter change is needed. $\Theta^{[t]}$ is the new parameter. The underlying rule behind this criterion is that a small update step size, which means a little variation in process, is tolerable. This indicates that no adaptation is needed. A single large update step size is treated as the result of abnormal measurements which should be ignored. Several persistent large update sizes indicate there has been an abrupt change in process behaviour and proper adaptation needs to be applied. As a result, an overall framework for on-line adaptation is summarized in Figure 4.3.

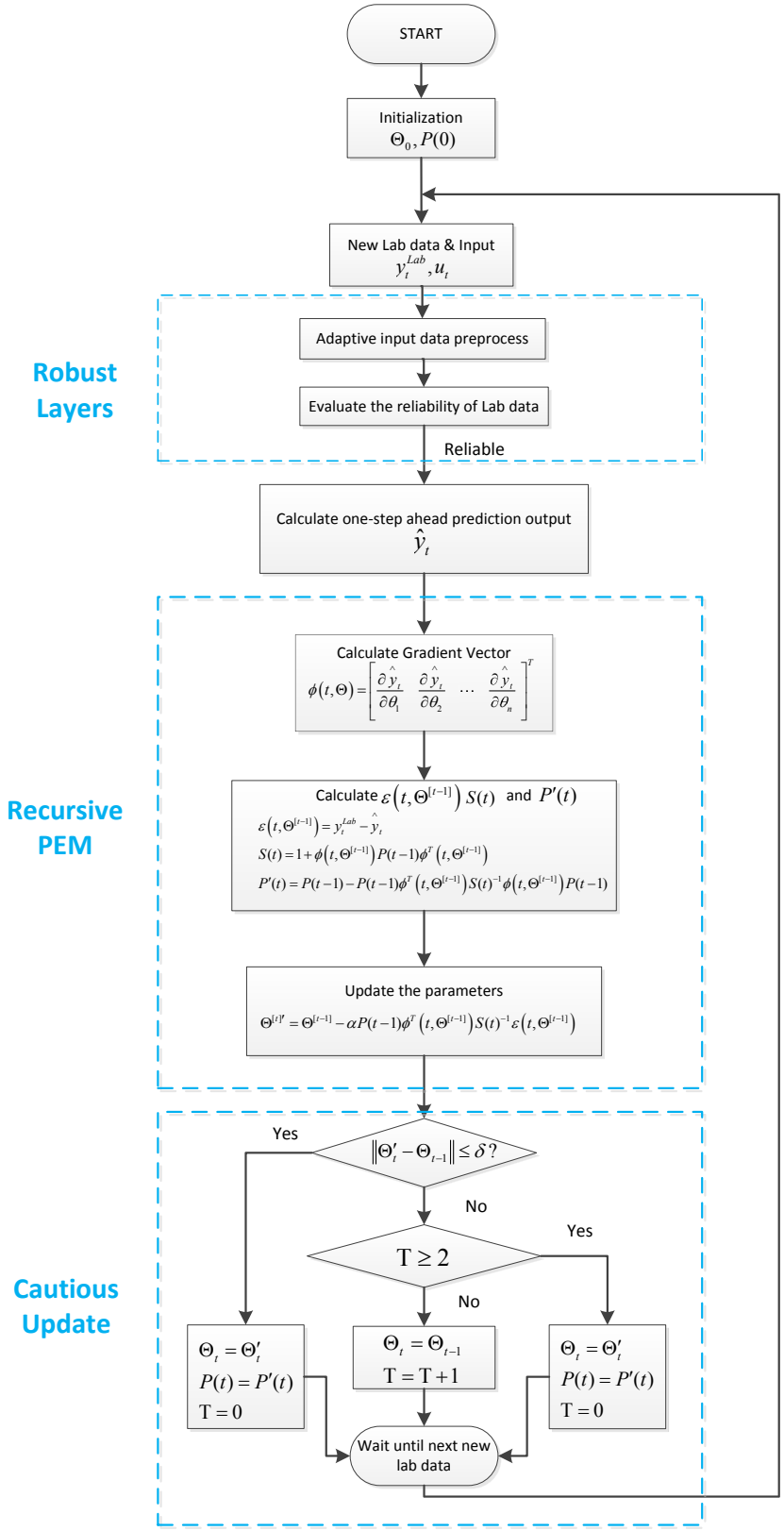


Figure 4.3: Recursive adaptation mechanism

4.4 Adaptive Soft Sensor for Naphtha:Bitumen Ratio in Inclined Plate Settler

4.4.1 Process Description

The main objective for the oil sands industry is the separation of bitumen from other components in oil sands, such as water and minerals. First, the oil sand is mixed with hot water and fed into the Primary Separation Vessel (PSV). Under the force of gravity, the resulting mixture is separated into three layers in the PSV, namely bitumen froth, middling and sand. The lightest layer, *i.e.* bitumen froth, floats to the top of PSV and then is sent to the froth treatment plant to remove residual water and fine solids. Bitumen froth is first diluted by mixing with process aids, *i.e.*, naphtha so that the density differences between bitumen, water and solids are increased. Then, the diluted froth is fed into the Inclined Plate Settler (IPS) which is used to separate mixture of naphtha, bitumen and water from minerals.

The Naphtha to Bitumen(N:B) ratio in the product stream of IPS indicates the quality of bitumen froth, thus serving as an important quality variable. In order to achieve an effective separation at affordable cost, the N:B ratio needs to be maintained at a certain level. However, the N:B measurements are not available on demand. As shown in Figure 4.5, the on-line hardware analyzers cannot provide accurate and reliable measurement of N:B. They are also expensive and difficult to maintain. Accurate N:B measurement available through off-line laboratory analysis cannot be used in real-time monitoring or control. Therefore, there is a need to develop a reliable soft sensor to provide more accurate estimate of N:B in product streams.

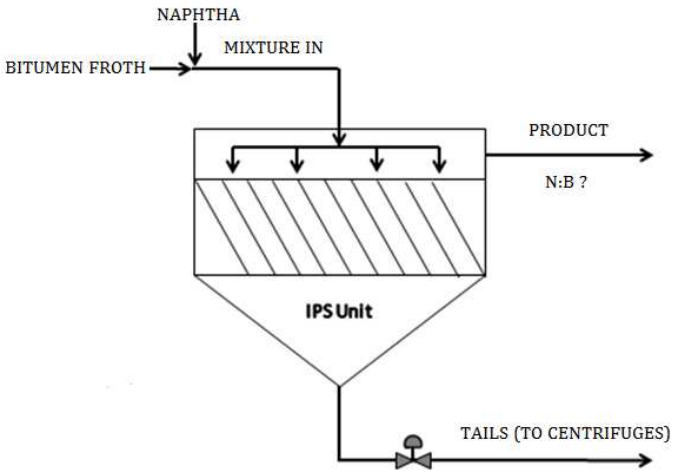


Figure 4.4: Schematic diagram of the inclined plates settler (IPS)

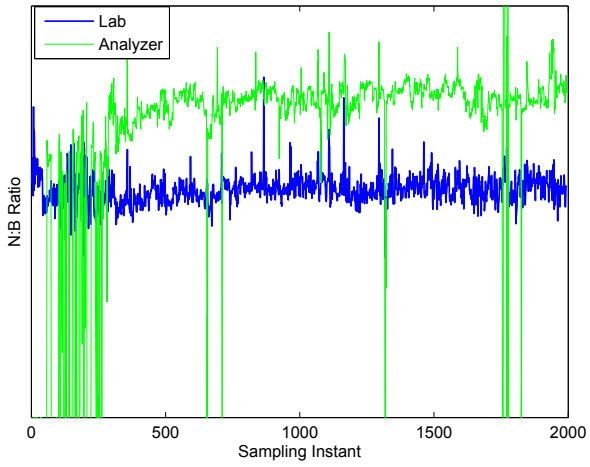


Figure 4.5: N:B ratio measured from the lab analysis and on-line analyzer

Process Variable	Symbol
F_n	Naphtha flow-rate
F_f	IPS feed flow-rate
ρ_f	IPS feed density
F_u	IPS underflow flow-rate
V	Accumulated volume of the mixture
ρ_v	Density of the accumulated volume

Table 4.2: A summary of the influential process variables, N:B soft sensor

4.4.2 Soft Sensor Design

Based on process knowledge, process variables identified to be influential in soft sensor design, are listed in Table 4.2. Sampling interval for the on-line measurements is 1 min, while the laboratory analysis of N:B is recorded every 2 hours. The identification dataset consists of the records of on-line measurements collected from May 1st 2012-Jan 1st 2013.

In order to estimate N:B ratio in the product stream, the mass balance of the process was analysed. Obviously, applying a first principle model requires the density and composition measurements. In the absence of these measurements, a grey-box model of the process was considered where available knowledge of the process was applied to determine an appropriate model structure. After the structure was defined, historical data collected from the operating process was used to reveal the parametric relationship between N:B ratio and other on-line measurements. Based on first-principle and data analysis, a grey-box model for N:B prediction was constructed.

$$NB^{product} = \frac{\frac{F_n}{\theta_1 F_f - \theta_5 F_n}(1 + \frac{d(V\rho_v)/dt}{F_f \rho_f}) + \theta_2 + \frac{\theta_3 F_u}{F_f}}{1 + \frac{d(V\rho_v)/dt}{F_f \rho_f} + \theta_4 \frac{F_u}{F_f}} + \theta_6 \tag{4.20}$$

4.4.3 Off-line Evalutaion

Auto-validation

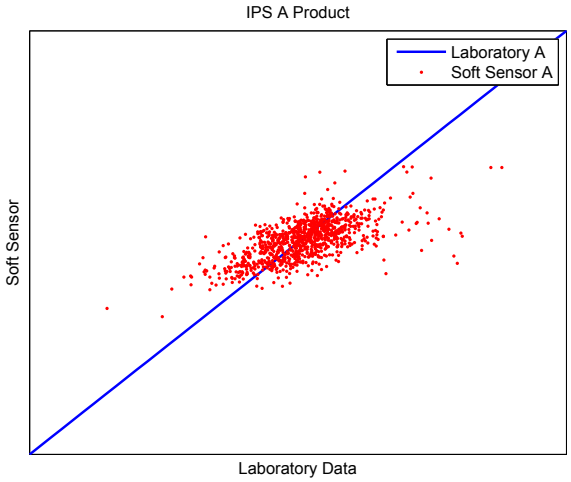
The performance of developed soft sensor was first verified on the identification dataset collected from January 1, 2013 to June 1, 2013. The scatter plots for product N:B predictions vs. lab measurements are shown in Figure 4.6.a and Figure 4.7.a. For proprietary reasons, all units and magnitudes of the plots are removed. The ideal case would be for all the data points to lie exactly along the diagonal line, indicating that the real-time predictions and the lab data are exactly the same. It can be observed that N:B predictions from the soft sensor without adaptation can fit the lab data well in the identification data-set. Also, Figure 4.6.b and Figure 4.7.b display the run-sequence plots of the predicted and target values.

Cross-validation

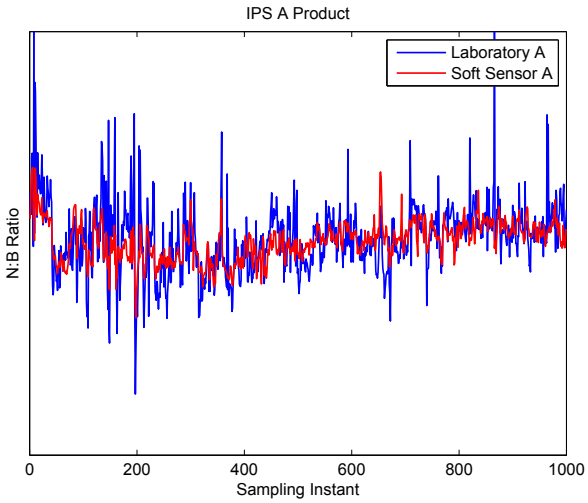
Next, the prediction performances of the designed soft sensors are evaluated on the validation data collected from April 01, 2013 to July 1, 2012. The scatter plots for N:B predictions obtained from the developed soft sensors vs. lab measurements are shown in Figure 4.8.a and 4.9.a. Moreover, the time trends of different N:B predictions are presented in Figure 4.8.b and 4.9.b. It is observed that the scatter plot of the adaptive soft sensor is closer to the lab data line when compared with the soft sensor without adaptation. In terms of time trends of predictions, it is observed that even when a good model is identified (shown in Figure 4.6 and 4.7), without adaptation, predictions of soft sensors can match the lab data well at the beginning, (first 200 points) but afterwards (200-1000 points) the performance gradually deteriorates. By contrast, the adaptive soft sensor can better match the trend of lab data and has smaller bias for a longer time period, which illustrates that on-line adaptation can significantly enhance the performance of the soft sensor

Methods	RMSEP	R
Soft Sensor A	0.0403	0.2462
Soft Sensor A with RPEM update	0.0221	0.5753
Soft Sensor B	0.0410	0.2456
Soft Sensor B with RPEM update	0.0215	0.5894

Table 4.3: Cross-validation results of N:B soft sensors

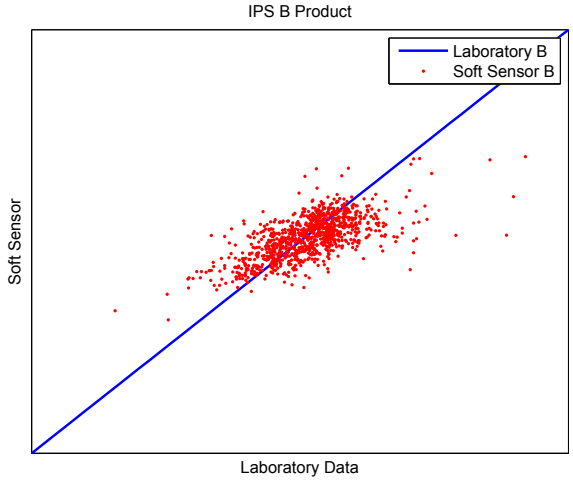


(a) Scatter plot

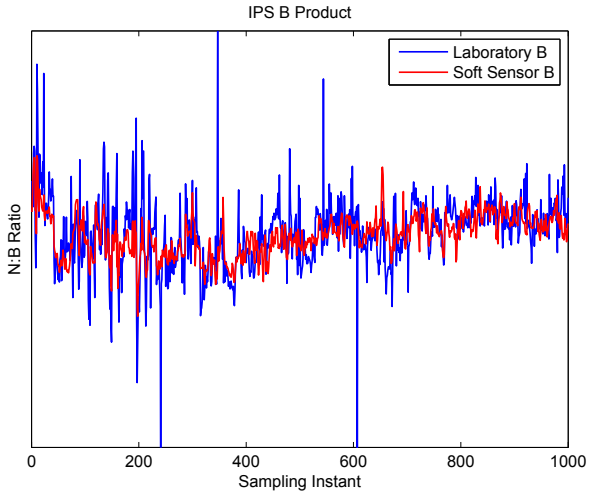


(b) Time trend

Figure 4.6: Auto-validation IPSA (Jan-Apr 2013)

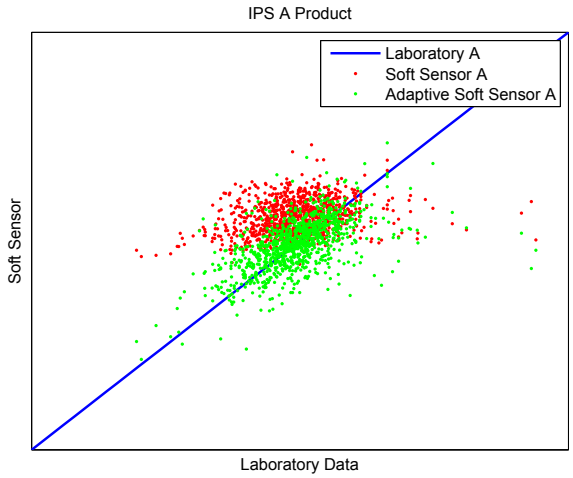


(a) Scatter plot

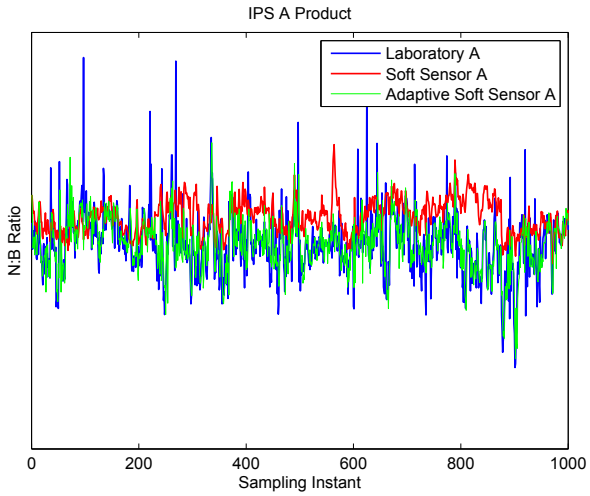


(b) Time trend

Figure 4.7: Auto-validation IPSB (Jan-Apr 2013)



(a) Scatter plot

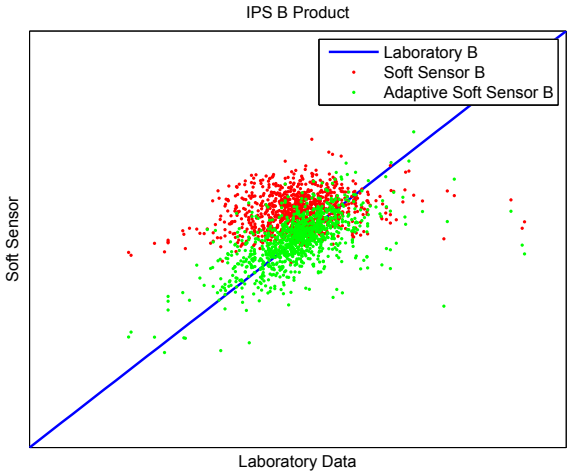


(b) Time trend

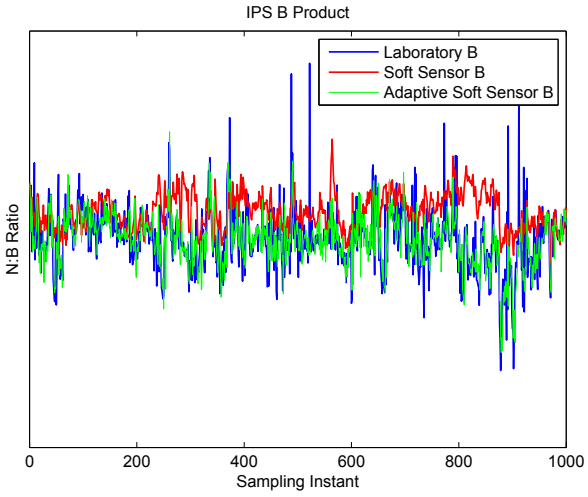
Figure 4.8: Cross-validation IPSA (Apr-Jul 2013)

4.4.4 On-line Evaluation

Off-line evaluation showed that the performance of the adaptive soft sensors was satisfactory. Thus, there were further tests performed in IPS units of extraction process. To implement the soft sensor, an Object linking and embedding for Process Control (OPC) platform was built in MATLAB to communicate with various devices in control systems. As shown in Figure 4.10, in Distributed Control System (DCS) different tags were created for the on-line measurements, such as valves, transmitters, analyzers and so on. Through Local Control Network (LCN), OPC servers had access to communicate with DCS to get on-line measurements. As a result, the OPC object



(a) Scatter plot



(b) Time trend

Figure 4.9: Cross-validation IPSB (Apr-Jul 2013)

in MATLAB could retrieve data in OPC servers through Process Control Network (PCN). In this way, all the necessary process variables for computation of predictions were available and the soft sensor predictions were sent back to DCS for control applications. To implement on-line adaptation, lab data was collected from Process Information (PI) Datalink System through a Local Area Network (LAN) and PCN, then adaptation was performed in MATLAB.

Two developed soft sensors have been tested on-line since September 10, 2013. Figure 4.11 and 4.12 show snapshots of the scatter plots and run sequence plots for N:B measurements versus laboratory measurements for IPSA and IPSB, respectively. It can be observed again that the adaptive soft sensor can provide accurate N:B predictions and outperform the soft sensor without adaptation. According to Table 4.4, the adaptive soft sensor can achieve better performance with smaller RMSE and higher correlation coefficient R.

Methods	RMSEP	R
Soft Sensor A	0.0456	0.3205
Soft Sensor A with RPEM update	0.0260	0.5746
Soft Sensor B	0.0490	0.2997
Soft Sensor B with RPEM update	0.0259	0.5985

Table 4.4: On-line evaluation results for N:B soft sensors

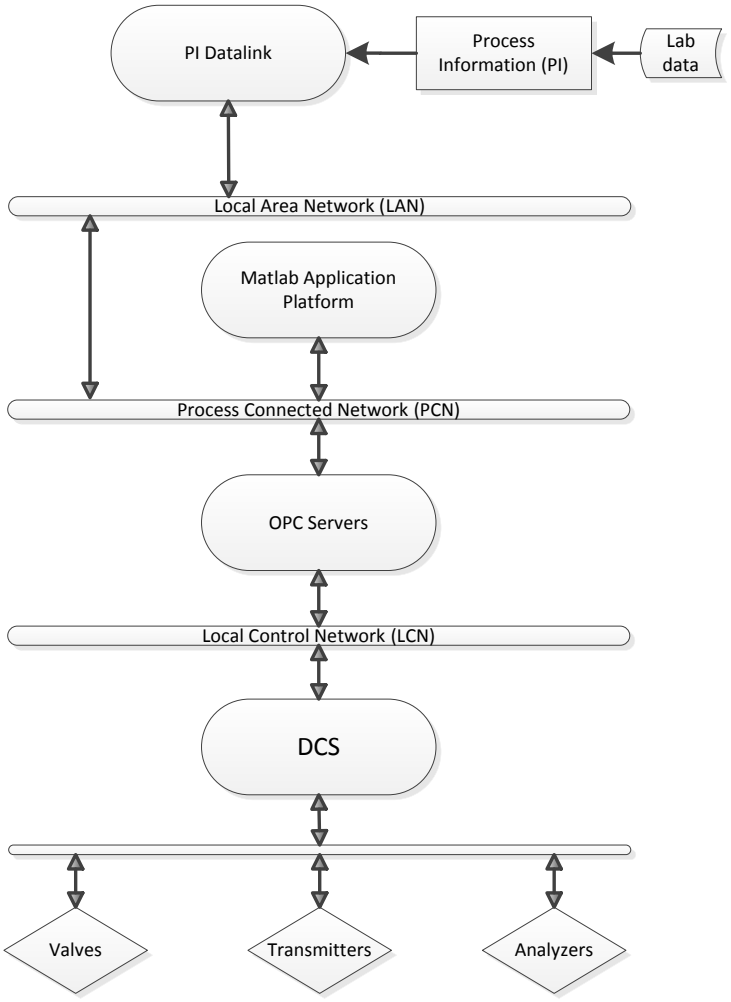
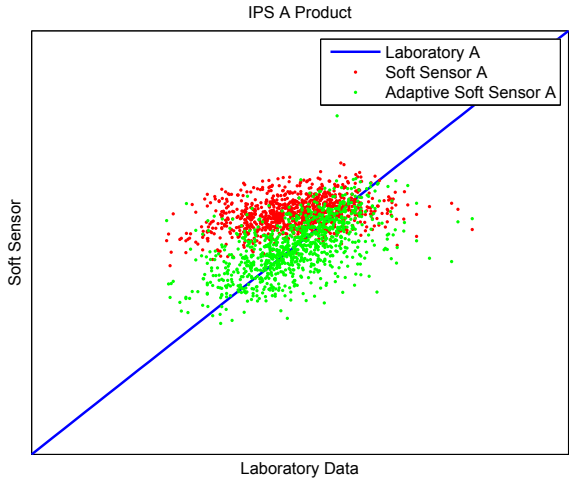
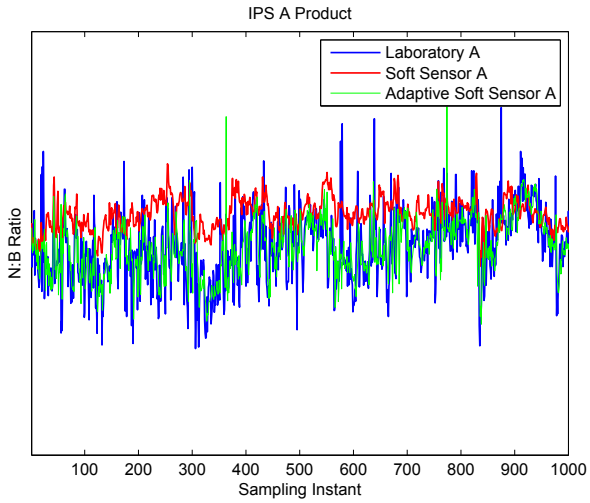


Figure 4.10: Data access network

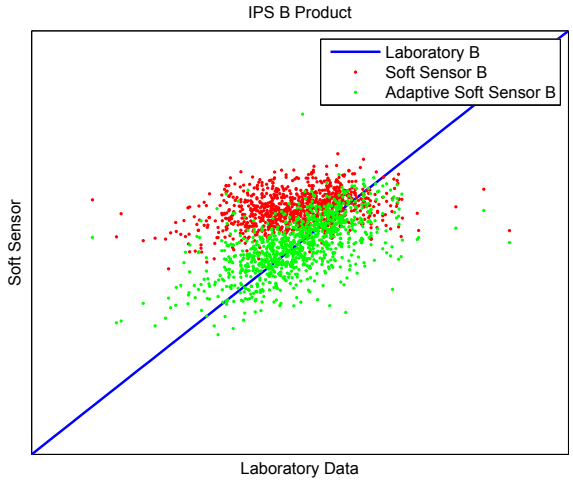


(a) Scatter plot

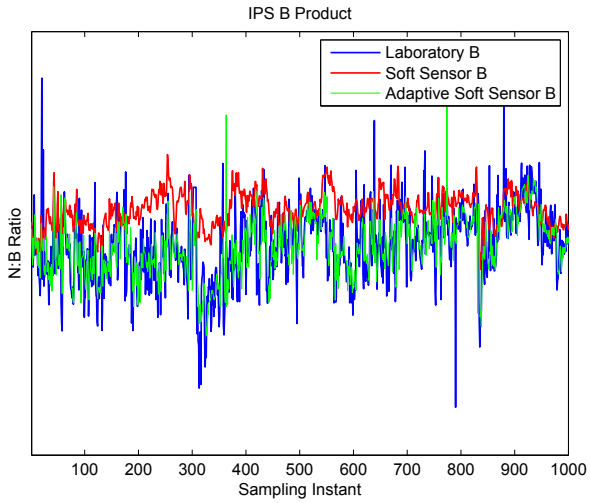


(b) Time trend

Figure 4.11: On-line evaluation IPSA (Sep-Dec 2013)



(a) Scatter plot



(b) Time trend

Figure 4.12: On-line evaluation IPSB (Sep-Dec 2013)

4.5 Conclusion

Grey-box soft sensors which are developed based on first principles of processes are widely used in industry due to their higher reliability and practicability. They can provide timely and crucial information that can help real-time process control. Despite the usefulness of the fundamental statistical modeling method - PEM, the on-line implementation in industry still remains a challenge because of the time-varying behaviour of processes. Industrial processes often vary due to changes in the environment and materials, variation in catalyst activity, fouling and/or abrasion in the process equipment. Due to these time-varying factors, it is difficult for data-driven methods to sustain long-term performance. Thus, it is desirable to integrate on-line adaptation in the implementation procedure. In this chapter, a recursive prediction error method (RPEM) based adaptation mechanism is adopted. Adaptive data preprocessing and cautious update strategy are integrated to ensure the robustness and effectiveness of the adaptation. This adaptation mechanism was applied to the design of adaptive soft sensors for oil sands industry. Based on off-line evaluations, these adaptive soft sensors can perform more accurate prediction of target variables, IPS product N:B ratio. Moreover, the adaptive soft sensors have already been implemented on-line and demonstrated superior prediction performance.

Chapter 5

Conclusions

5.1 Summary of Thesis

This thesis focuses on some key challenges to the development and implementation of soft sensors. Despite increasing number of publications concerning these fields, development and maintenance of successful soft sensors applications is still laborious.

In the development phase of a soft sensor, traditional methods, such as data-driven methods (OLS, PCR, PLS) as well as some model-driven methods, may not achieve satisfactory performance. Thus, the locally weighted modeling method (also called Just-in-time (JIT) modeling) was proposed to deal with this issue. JIT methods can deal with multiple operating modes issues as well as nonlinearity. The performance of JIT method is affected by following three aspects: selection of local calibration samples, selection of model structure and estimation of model parameters. Considering these points, a Bayesian framework for just-in-time modeling was proposed to offer a systematic way to search for the optimal combination of model structures, local calibration samples as well as model parameters for each operating point. Bayesian model structure selection can automatically penalize model complexity, thus avoiding over-fitting. The Bayesian approach also makes it possible to incorporate prior knowledge into estimation of the main parameter which can enhance the accuracy of estimation. To further improve the Bayesian JIT method, a new similarity function was proposed and integrated into the developed Bayesian framework. This new input-output similarity function takes both input and output information into account in order to ensure the representativeness of local calibration samples.

Even if a good soft sensor is developed initially, after implementing for a certain

period of time, the performance can deteriorate due to time-varying behaviours of processes. To deal with this issue, during the implementation phase of a soft sensor project, on-line adaptation is often integrated. In this thesis, an adaptation mechanism for nonlinear grey-box models was proposed based on recursive prediction error method. Adaptive data preprocessing was applied to deal with the negative effect of noise and abnormal values in the measurement. A cautious update strategy was integrated to ensure the reliability of lab data and recognize the need for adaptation in order to guarantee the robustness and effectiveness of the adaptation. This proposed framework was successfully applied to adaptive soft sensors designs within the oil sands industry: naphtha : bitumen ratio soft sensors. To develop the Naphtha to Bitumen soft sensor, mass balance equations were used to determine a proper model structure. In the absence of density and composition measurements, again, a nonlinear grey-box model was built and prediction error method was used to identify the unknown parameters. Due to abrasion in the process equipment and changes in operating region, the process exhibited a form of time-varying behaviour. Thus, to deal with this issue, on-line adaptation was applied to develop adaptive soft sensors. Based on these results, these adaptive soft sensors can produce accurate predictions for target variables for a long period of time.

5.2 Future Work

In this thesis, the proposed Bayesian framework was developed to solve Just-in-time modeling problems. To further improve the performance and efficiency, one can consider following aspects in future work:

1. In Chapter 2, the proposed Bayesian framework is developed under assumption that the noise-free input follows a Gaussian distribution. Such assumptions are commonly made in Bayesian methods, such as Kalman filtering. Even when this assumption is not satisfied the results can still be better than those obtained by other methods. However, in order to improve the accuracy of assumption and to further improve the performance, one can assume Student's T distribution which will make this method more robust to outliers.

2. In Chapter 2, Bayesian framework offers a way to take into account the different contribution of noise in measurements, based on variance. Although there is an empirical estimation method for these variances, a more systematic way to estimate the variances can be considered.

3. In Chapter 3, a new input-output similarity function was proposed to specify the local calibration samples. In order to calculate the distance in output space, query outputs were estimated by the global PLS model using corresponding query inputs. This method is effective when behaviour of the process tends to be linear, or it cannot provide fairly accurate estimates of query outputs. However, the JIT methods are mainly used to deal with nonlinearity. The situation may contradict with the assumption behind the alternative which is used to get estimated query outputs. Instead of using the global PLS model, one may consider other nonlinear modeling methods, such as Kernel PCA, PLS, Supported Vector Machine (SVM) and *etc.* to build the model for estimation of query outputs.

4. In Chapter 3, the balance parameter ω and localization parameter λ were determined by LOOCV and kept fixed for each different operating points. However, it is possible that at each operating point, the optimal balance parameter and

localization parameter may be different, so one can consider it as a hyperparameter in hierarchical Bayesian framework which will be specified for each operating point during real-time operation.

5. Although the proposed Bayesian framework can be used for real-time identification, the large computational burden is still a problem.

6. The abnormal point detection criteria in adaptation mechanism assumes that process variables follows Gaussian distributions. Although this assumption is usually true for some cases, in order to strength generalizability of the method, one may consider other different distributions of process variables, thus to develop a robust recursive PEM which can deal with the abnormal points automatically and more systematically .

Bibliography

[1] Petr Kadlec, Bogdan Gabrys, and Sibylle Strandt. Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4):795–814, 2009.

[2] Shima Khatibisepehr, Biao Huang, and Swanand Khare. Design of inferential sensors in the process industry: A review of bayesian methods. *Journal of Process Control*, 23(10):1575–1596, 2013.

[3] Luigi Fortuna, Salvatore Graziani, Alessandro Rizzo, and Maria Gabriella Xibilia. *Soft sensors for monitoring and control of industrial processes*. Springer, 2007.

[4] Hervé Abdi. Partial least squares regression and projection on latent structure regression (pls regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):97–106, 2010.

[5] Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

[6] Bao Lin and Sten Bay Jørgensen. Soft sensor design by multivariate fusion of image features and process measurements. *Journal of Process Control*, 21(4):547–553, 2011.

[7] Xinguang Shao, Fangwei Xu, Biao Huang, and Aris Espejo. 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, 2012.

[8] LLdiko E Frank and Jerome H Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[9] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185(0):1 – 17, 1986.

[10] Svante Wold, Nouna Kettaneh-Wold, and Bert Skagerberg. Nonlinear pls modeling. *Chemometrics and Intelligent Laboratory Systems*, 7(1):53 – 65, 1989.

[11] S.J. Qin and T.J. McAvoy. Nonlinear pls modeling using neural networks. *Computers & Chemical Engineering*, 16(4):379 – 391, 1992.

[12] Yoon Ho Bang, Chang Kyoo Yoo, and In-Beum Lee. Nonlinear pls modeling with fuzzy inference system. *Chemometrics and intelligent laboratory systems*, 64(2):137–155, 2002.

[13] E.C. Malthouse, A.C. Tamhane, and R.S.H. Mah. Nonlinear partial least squares. *Computers & Chemical Engineering*, 21(8):875 – 890, 1997.

[14] Sanghong Kim, Ryota Okajima, Manabu Kano, and Shinji Hasebe. Development of soft-sensor using locally weighted pls with adaptive similarity measure. *Chemometrics and Intelligent Laboratory Systems*, 124(0):43 – 49, 2013.

[15] David Perez-Guaita, Julia Kuligowski, Guillermo Quints, Salvador Garrigues, and Miguel de la Guardia. Modified locally weightedpartial least squares regression improving clinical predictions from infrared spectra of human serum samples. *Talanta*, 107(0):368 – 375, 2013.

[16] Riccardo Leardi. *Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks: genetic algorithms and artificial neural networks*, volume 23. Elsevier, 2003.

[17] Sanghong Kim, Manabu Kano, Shinji Hasebe, Akitoshi Takinami, and Takeshi Seki. Long-term industrial applications of inferential control based on just-in-time soft-sensors: Economical impact and challenges. *Industrial & Engineering Chemistry Research*, 52(35):12346–12356, 2013.

[18] Shima Khatibisepehr, Sanghong Kim, Mulang Chen, Biao Huang, and Manabu Kano. A probabilistic framework for model structure selection and hyperparameters tuning in locally weighted partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 2013, in review.

[19] Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley. Model selection: Beyond the bayesian/frequentist divide. *The Journal of Machine Learning Research*, 11:61–87, 2010.

[20] David JC MacKay. Comparison of approximate methods for handling hyperparameters. *Neural computation*, 11(5):1035–1068, 1999.

[21] Mohamed N Nounou, Bakshi, and Bhavik R. Process modeling by bayesian latent variable regression. *AIChE Journal*, 48(8):1775–1793, 2002.

[22] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2003.

[23] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[24] Dorthe Kjær Pedersen and Martens. Near-infrared absorption and scattering separated by extended inverted signal correction (eisc): Analysis of near-infrared transmittance spectra of single wheat seeds. *Applied spectroscopy*, 56(9):1206–1214, 2002.

[25] Jesper Pram Nielsen, Dorthe Kjær Pedersen, and Lars Munck. Development of nondestructive screening methods for single kernel characterization of wheat. *Cereal chemistry*, 80(3):274–280, 2003.

[26] Manabu Kano and Yoshiaki Nakagawa. Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry. *Computers & Chemical Engineering*, 32(1?2):12 – 24, 2008.

[27] Shima Khatibisepehr and Biao Huang. Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Industrial & Engineering Chemistry Research*, 47(22):8713–8723, 2008.

[28] JV Kresta, TE Marlin, and JF MacGregor. Development of inferential process models using pls. *Computers & Chemical Engineering*, 18(7):597–611, 1994.

[29] Manabu Kano, Koichi Miyazaki, Shinji Hasebe, and Iori Hashimoto. Inferential control system of distillation compositions using dynamic partial least squares regression. *Journal of Process Control*, 10(2):157–166, 2000.

[30] Cheng-Wen Chang, David A Laird, Maurice J Mausbach, and Charles R Hurburgh. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. *Soil Science Society of America Journal*, 65(2):480–490, 2001.

[31] Shima Khatibisepehr, Biao Huang, Elom Domlan, Elham Naghoosi, Yu Zhao, Yu Miao, Xinguang Shao, Swanand Khare, Marziyeh Keshavarz, Enbo Feng, Fangwei Xu, Aris Espejo, and Ramesh Kadali. Soft sensor solutions for control of oil sands processes. *The Canadian Journal of Chemical Engineering*, 91(8):1416–1426, 2013.

[32] Lei Xie, Jiusun Zeng, and Chuanhou Gao. Novel just-in-time learning-based soft sensor utilizing non-gaussian information. *Control Systems Technology, IEEE Transactions on*, 22(1):360–368, 2014.

[33] S Joe Qin. Recursive pls algorithms for adaptive data modeling. *Computers & Chemical Engineering*, 22(4):503–514, 1998.

[34] Stefan Rännar, John F MacGregor, and Svante Wold. Adaptive batch monitoring using hierarchical pca. *Chemometrics and intelligent laboratory systems*, 41(1):73–81, 1998.

[35] Weihua Li, H Henry Yue, Sergio Valle-Cervantes, and S Joe Qin. Recursive pca for adaptive process monitoring. *Journal of process control*, 10(5):471–486, 2000.

[36] John B Moore and Haim Weiss. Recursive prediction error methods for adaptive estimation. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(4):197–205, 1979.

[37] Christopher G Atkeson, Andrew W Moore, and Stefan Schaal. Locally weighted learning for control. In *Lazy learning*, pages 75–113. Springer, 1997.

[38] Gianluca Bontempi, Mauro Birattari, and Hugues Bersini. Lazy learning for local modelling and control design. *International Journal of Control*, 72(7-8):643–658, 1999.

[39] Koichi Fujiwara, Manabu Kano, Shinji Hasebe, and Akitoshi Takinami. Soft-sensor development using correlation-based just-in-time modeling. *AIChE Journal*, 55(7):1754–1765, 2009.

[40] Mulang Chen, Swanand Khare, and Biao Huang. Orthogonal signal correction based input-output similarity for just-in-time modeling. 2014, under review.

[41] Ziyi Wang, Tomas Isaksson, and Bruce R Kowalski. New approach for distance measurement in locally weighted regression. *Analytical Chemistry*, 66(2):249–260, 1994.

[42] M Dyrby, SB Engelsen, L Nørgaard, M Bruhn, and L Lundsberg-Nielsen. Chemometric quantitation of the active substance (containing c≡ n) in a pharmaceutical tablet using near-infrared (nir) transmittance and nir ft-raman spectra. *Applied Spectroscopy*, 56(5):579–585, 2002.

[43] Shima Khatibisepehr and Biao Huang. A bayesian approach to robust process identification with arx models. *AIChE Journal*, 59(3):845–859, 2013.

[44] Daniel Sbarbaro, Pedro Ascencio, Pablo Espinoza, Felipe Mujica, and Guillermo Cortes. Adaptive soft-sensors for on-line particle size estimation in wet grinding circuits. *Control Engineering Practice*, 16(2):171–178, 2008.

[45] Tirtha Chatterjee and Deoki N Saraf. On-line estimation of product properties for crude distillation units. *Journal of Process Control*, 14(1):61–77, 2004.

[46] Theodora Kourti. Process analysis and abnormal situation detection: from theory to practice. *Control Systems, IEEE*, 22(5):10–25, 2002.

[47] SH Yang, XZ Wang, C McGreavy, and QH Chen. Soft sensor based predictive control of industrial fluid catalytic cracking processes. *Chemical Engineering Research and Design*, 76(4):499–508, 1998.

[48] Jialin Liu, Ding-Sou Chen, and Jui-Fu Shen. Development of self-validating soft sensors using fast moving window partial least squares. *Industrial & Engineering Chemistry Research*, 49(22):11530–11546, 2010.

[49] Herbert JAF Tulleken. Grey-box modelling and identification using physical knowledge and bayesian techniques. *Automatica*, 29(2):285–308, 1993.

[50] Lennart Ljung. Prediction error estimation methods. *Circuits, Systems and Signal Processing*, 21(1):11–21, 2002.

[51] Petr Kadlec, Ratko Grbić, and Bogdan Gabrys. Review of adaptation mechanisms for data-driven soft sensors. *Computers & chemical engineering*, 35(1):1–24, 2011.

[52] Bhupinder S Dayal and John F MacGregor. Recursive exponentially weighted pls and its applications to adaptive control and prediction. *Journal of Process Control*, 7(3):169–179, 1997.

[53] Pierantonio Facco, Franco Doplicher, Fabrizio Bezzo, and Massimiliano Barolo. Moving average pls soft sensor for online product quality estimation in an industrial batch polymerization process. *Journal of Process Control*, 19(3):520–529, 2009.

[54] Rashad Javadli and Arno de Klerk. Desulfurization of heavy oil. *Applied Petrochemical Research*, 1(1-4):3–19, 2012.

[55] Shima Khatibisepehr. Bayesian solutions to multi-model inferential sensing problems. *Ph.D. Thesis, University of Alberta*, 2013.

Algorithm I:Regular Locally Weighted Partial Least Square

1. Determine the number of latent variables H, localization parameters λ

2. When query sample $\mathbf{x}_q$ arrives, calculate the similarity matrix $\mathbf{S}_q$ using Eqns. 2.5, 2.6 and 2.7 .

3. Calculate the weight matrix, loading matrix and regression coefficient vector by:

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2 \cdots \mathbf{w}_H] \tag{1}$$

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2 \cdots \mathbf{p}_H] \tag{2}$$

$$q = [q_1, q_2 \cdots q_H] \tag{3}$$

$$\mathbf{w}_h = \frac{(\mathbf{X} - \sum_{j=1}^{l-1} \mathbf{t}_j {\mathbf{p}_j}^T)^T \mathbf{S}_q (\mathbf{X} - \sum_{j=1}^{h-1} \mathbf{t}_j {\mathbf{p}_j}^T)}{\left\| (\mathbf{X} - \sum_{j=1}^{h-1} \mathbf{t}_j {\mathbf{p}_j}^T)^T \mathbf{S}_q (\mathbf{X} - \sum_{j=1}^{h-1} \mathbf{t}_j {\mathbf{p}_j}^T) \right\|} \tag{4}$$

$$\mathbf{p}_h = \frac{(\mathbf{X} - \sum_{j=1}^{h-1} \mathbf{t}_j {\mathbf{p}_j}^T)^T \mathbf{S}_q \mathbf{t}_h}{{\mathbf{t}_h}^T \mathbf{S}_q \mathbf{t}_h} \tag{5}$$

$$q_h = \frac{(y - \sum_{j=1}^{h-1} \mathbf{t}_j {q_j}^T)^T \mathbf{S}_q \mathbf{t}_h}{{\mathbf{t}_h}^T \mathbf{S}_q \mathbf{t}_h} \tag{6}$$

where the columns of $\mathbf{W} \in \mathbb{R}^{M \times H}$ are orthonormal weight vectors and $\mathbf{t}_h$ denotes the h^{th} column of $\mathbf{T}$ which is calculated by:

$$\mathbf{t}_h = (\mathbf{X} - \sum_{j=1}^{h-1} \mathbf{t}_j {\mathbf{p}_j}^T)\mathbf{w}_h \tag{7}$$

4. Calculate output of the local PLS model by:

$$\hat{\mathbf{y}} = \mathbf{x}_q \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q}^T \tag{8}$$

Algorithm II: Estimation of Empirical Prior Over Main Parameters

1. For industry data, it is rational to assume in a short period of time (*i.e.* one sampling interval), the input and output are kept constant. The incremental input output measurements are resulted from the measurement noise. So the noise variance $\mathbf{Q}_{e_x}$, Q_{e_y} is calculated by the variances of the distribution of incremental input and output measurements:

$$\mathbf{J}_x = [\mathbf{x}_1 - \mathbf{x}_2, \mathbf{x}_2 - \mathbf{x}_3, \cdots \mathbf{x}_{N-1} - \mathbf{x}_N]^T \tag{9}$$

$$\mathbf{j}_y = [y_1 - y_2, y_2 - y_3, \cdots y_{N-1} - y_N]^T \tag{10}$$

$$\mathbf{Q}_{e_x} = \frac{1}{2}\text{var}(\mathbf{J}_x) \tag{11}$$

$$Q_{e_y} = \frac{1}{2}\text{var}(\mathbf{j}_y) \tag{12}$$

2. Solve the Bayesian LW-PLS modeling problem with a uniform priori for all the main parameters.

3. Estimate the set of hyperparameters $\mu_b, \mathbf{Q}_b, \mu_x, \mathbf{Q}_x$ as follows:

$$\mu_b = E[\mathbf{P}\mathbf{q}^T] \tag{13}$$

$$\mathbf{Q}_b = c(\hat{\mathbf{X}}^T \mathbf{S}_q \hat{\mathbf{X}})^{-1} \tag{14}$$

$$\mu_x = E[\hat{\mathbf{X}}] \tag{15}$$

$$\mathbf{Q}_x = Cov[\hat{\mathbf{X}}] \tag{16}$$

4. Solve the Bayesian LW-PLS modeling problem using the empirically estimate priori.

Algorithm III: Estimation of Empirical Prior Over Localization Parameter

1. Determine proper model structure H.

2. Choose the similarity function given in Eqn. 6 and determine a proper set of localization parameters $[\lambda_1, \lambda_2, \cdots \lambda_f]$.

3. For $f = 1 : F$

   (1). Choose λ_f as localization parameter.

   (2). For $n = 1 : N$

   Let $\{\mathbf{X}_{-n}, \mathbf{y}_{-n}\}$ denote calibration samples except $\{\mathbf{x}_n, y_n\}$. Choose $\{\mathbf{X}_{-n}, \mathbf{y}_{-n}\}$ as calibaration samples and $\mathbf{x}_n$ as query sample, and apply the regular LW-PLS algorithm (*Algorithm I*) to $\{\mathbf{X}_{-n}, \mathbf{y}_{-n}, \mathbf{x}_n\}$ get the output prediction $\hat{y}_n$.

   (3). Calculate the prediction error

$$E_f = \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2 \tag{17}$$

4. Choose the localization parameter that results in the lowest prediction error, denote it as λ_k and record the value for each point in similarity function as

$$\varphi_i = \frac{1}{\lambda_k \sigma_{d_i}} \ (i = 1, 2 \cdots N) \tag{18}$$

5. Determine a Gamma prior distribution over φ based on $\{\varphi_1, \varphi_2 \cdots \varphi_N\}$.

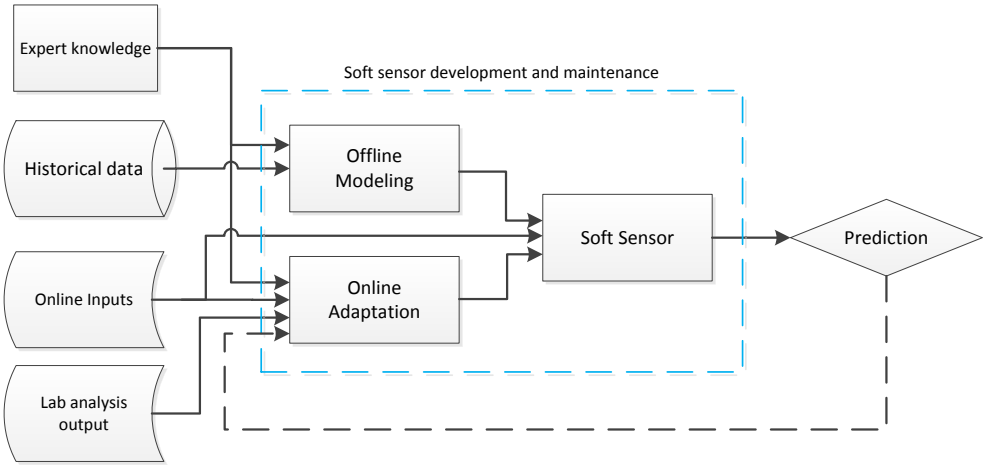


Figure 4.1: Recursive adaptation flowchat

and data scaling.

2. Performance feedback: The need for adaptation should be triggered by monitoring the performance of the soft sensor. There are different criteria to evaluate the performance and determine when to apply adaptation to the soft sensor.

3. Adaptation algorithm: Once the data are available, and there is a need for adaptation, the adaptation algorithm can be applied to adjust the soft sensor.

Despite the increasing number of publications dealing with adaptation of soft sensors, several issues remain open. Most of the existing methods for adaptation are moving window methods or methods involving recursive updating of OLS, PCA and PLS which are designed for adaptation of linear models[52, 35, 53]. They cannot be applied directly to nonlinear models. Moreover, in practice, the second aspect, *i.e.*, performance feedback is often ignored. Instead, the adaptation is performed at a certain frequency or once new lab analysis output is available. This may result in over-updating issues.

Considering these challenges, the recursive prediction error method (RPEM) is adopted in this chapter as an adaptation algorithm for grey-box models which can

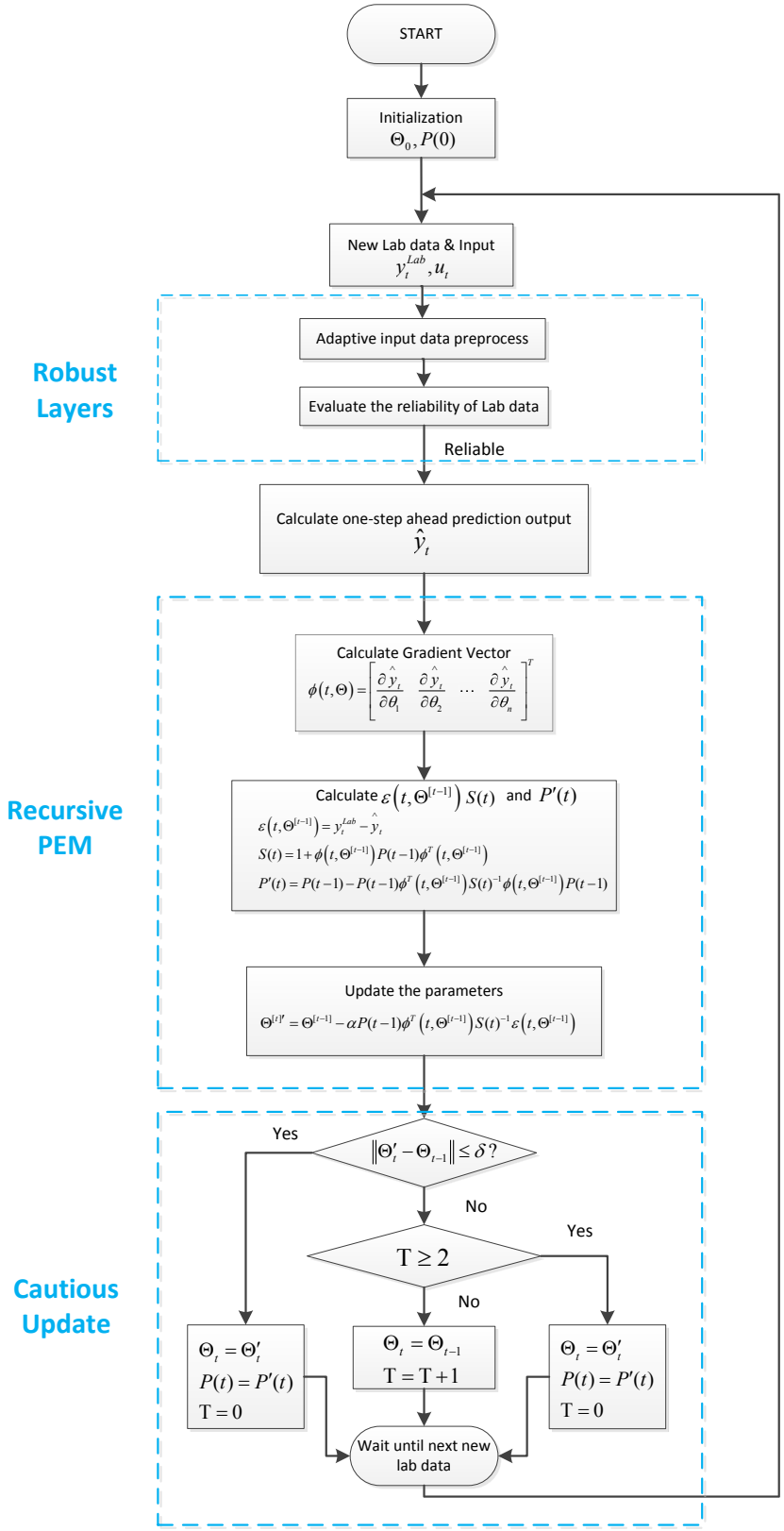


Figure 4.3: Recursive adaptation mechanism

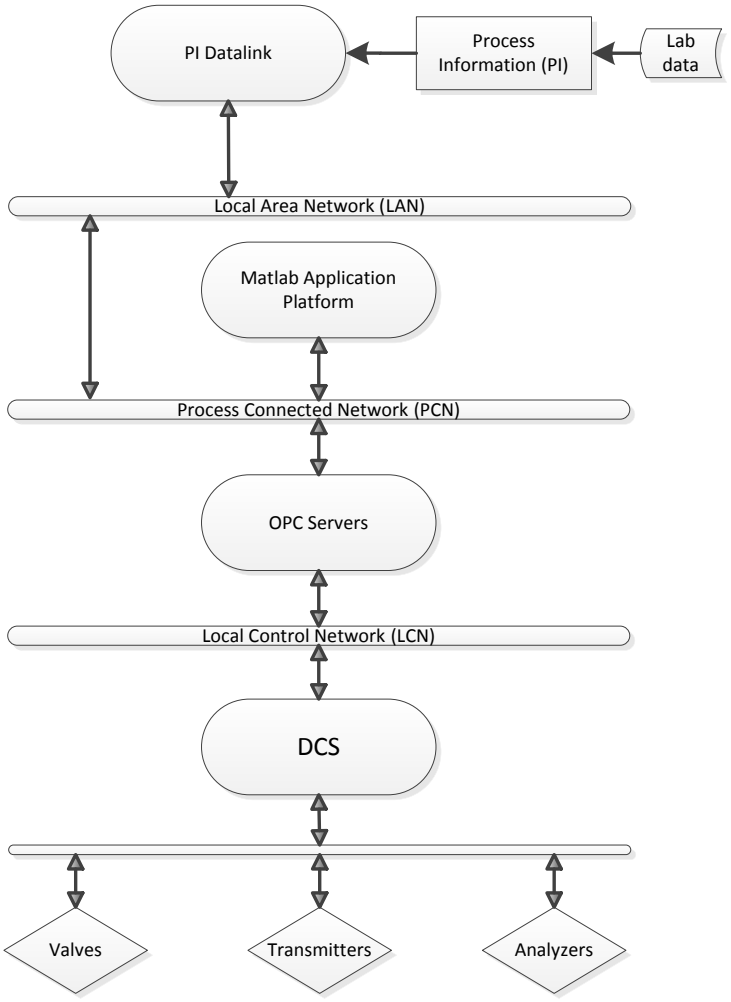


Figure 4.10: Data access network