2000

Probabilistic process data rectification

Morad, Kamalaldin

http://hdl.handle.net/1880/42649
doctoral thesis

University of Calgary graduate students retain copyright ownership and moral rights for their thesis. You may use this material in any way that is permitted by the Copyright Act or through licensing that has been assigned to the document. For uses that are not allowable under copyright legislation or licensing, you are required to seek permission.

Downloaded from PRISM: https://prism.ucalgary.ca
THE UNIVERSITY OF CALGARY

Probabilistic Process Data Rectification

by

Kamalaldin Morad

A DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMICAL AND PETROLEUM ENGINEERING

CALGARY, ALBERTA

MARCH, 2000

© Kamalaldin Morad 2000
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-49520-5
Abstract

Process measurements provide a window of information through which the state of a chemical plant can be derived and the plant’s operation described. When these measurements are flawed by noise or outliers, the conclusions drawn from these measurements can be unreliable and can lead to poor process control and plant operation. Data rectification is the process of removing errors from the measured process data and estimating the true state of the plant. In this dissertation, data rectification is posed in a probabilistic framework and historical plant data is utilized to learn the parameters of the plant model. This approach finds the most likely estimates of the true process states by maximizing the probability of the process states given the measurements. Using Bayes’ theorem, this maximization is redefined as the product of the prior probability density function (pdf) of process states and the probability distribution of measurements given the true process states. The technique exploits the existing trade off between these two terms to find the most likely values for the measured process variables.

The estimation of the covariance matrix is an important issue in data rectification. The research underlying this dissertation suggests the probabilistic approach toward data rectification and the use of principal component analysis for size reduction, emphasizes the need for a reliable estimate of the covariance matrix. A robust direct approach for covariance matrix estimation is proposed and its ability to reject outliers and provide reliable estimates of the covariance matrix is demonstrated.

The measured process data used in setting up the pdf of the process states are often imperfect, missing some components. An unsupervised method of
learning the parameters of a probability density mixture in the presence of missing data using the Expectation-Maximization (EM) algorithm is developed. The reliability of the estimates is examined as the amount of missing values in the process measured data increases.

The method of adaptive mixtures is used for both offline and online estimating and updating of the pdf of the measured process variables. It is a recursive nonparametric method that fits a mixture of Gaussian pdf's to the data. The changes in the process operating conditions are reflected by adding new components to the mixture.

The maximum likelihood data rectification objective function consists of two pdf's. One represents the likely process states and the other characterizes the likely adjustments to the measured values. While the first pdf, given the measured data condition, is estimated by the EM algorithm or adaptive mixtures, the second pdf is modeled by the product of bimodal Gaussian distributions each representing a process sensor. The resultant complex objective function is then maximized by an iterative EM algorithm and the most likely values of the measured process variables are found. The changes in process operation are reflected in the objective function by updating its terms. The pdf of the measured process states is updated using the rectified points and a robust approach is developed to update the pdf's representing the plant sensors operation.

These new approaches are implemented for numerical and chemical process examples. The capabilities of these new approaches are demonstrated by producing robust estimates for covariance matrices, reliable estimates of mixture probability densities, rejecting the errors in the measurements and yielding reliable rectified values for the measured process variables in these examples.
Acknowledgments

As with any undertaking, the work which went into the completion of this dissertation could not have been done without the support and interest of a number of people.

My supervisor, Dr. W. Y. Svrcek, must be especially acknowledged for his scholarly advice, sacrifice of time and thoughtful editing of this manuscript. His practical experience put my research onto the correct path to achieve a solution to real-world problems. His generous financial support ensured my peace of mind during the whole course of study.

My special thanks to Dr. R. G. Moore for always being helpful when the help was most needed. Mr. I. McKay from AEA Technologies, Hyprotech Ltd, was an invaluable resource. I am deeply indebted to his insights on this project. I offer warm thanks to our graduate secretaries, Ms. S. Hutton, Ms. D. Parker and Ms. A. Spence for their swift replies to my numerous inquiries.

I am deeply grateful to all my dear friends who helped me in many ways during this endeavor. I would especially like to thank F. Akbaryan for his great ideas and kind support and R. Babayi for many years of collaboration in various ways.

The Department of Chemical and Petroleum Engineering provided me with the financial support without which I might not even have been able to pursue my studies.

My parents have always given me the encouragement to learn more and employ my knowledge in practical ways. I feel eternal gratitude to them for the part of their lives that they have sacrificed to me.

My great thanks must go to my wife Shervin who allowed me to meet the challenge of this research. She was always a full partner in this endeavor.
To

My Wife and My Parents
Contents

Approval Page ............................................. ii
Abstract ................................................... iii
Acknowledgments ........................................... v
Dedication .................................................. vi
Contents ..................................................... vii
List of Figures ............................................. xi
List of Tables .............................................. xiv

1 Introduction .............................................. 1
  1.1 Introduction .......................................... 1
  1.2 Motives ............................................... 2
    1.2.1 Plant safety ..................................... 2
    1.2.2 Efficient plant operation ....................... 4
    1.2.3 Other issues .................................... 5
  1.3 A broader look at data reconciliation ................. 6
  1.4 Outline of approach ................................ 9
  1.5 Research objective ................................ 11
    1.5.1 Capturing all process knowledge in a data rectification framework ........................................... 11
    1.5.2 Measurement error covariance estimation ........ 12
    1.5.3 Probability density estimation using incomplete data .......................................................... 12
    1.5.4 Density estimation using adaptive mixtures .... 13
    1.5.5 Sensor model .................................... 13
  1.6 Summary of thesis ................................... 14
2 Data Reconciliation Background
   2.1 Introduction ........................................ 16
   2.2 Linear data reconciliation .......................... 20
   2.3 Nonlinear data reconciliation ....................... 28
   2.4 Knowledge based data reconciliation ............... 33
   2.5 Summary ............................................ 35

3 Estimation of Measurement Error Covariance Matrix 37
   3.1 Introduction ........................................ 37
   3.2 Overview of existing methods ....................... 39
   3.3 Outline of the new approach ....................... 44
      3.3.1 The maximum likelihood estimator ............ 44
      3.3.2 Mean and covariance estimation algorithm ...... 45
   3.4 Examples ........................................... 49
      3.4.1 Diagonal case .................................. 49
      3.4.2 Nondiagonal case .............................. 52
   3.5 Summary ............................................ 58

4 Probability Density Estimation Using Incomplete Data 60
   4.1 Introduction ....................................... 60
   4.2 Mixture density estimation ........................ 64
      4.2.1 Mixture density model ....................... 64
      4.2.2 Mixture density estimation using the method of maximum likelihood .................... 65
   4.3 The Expectation-Maximization (EM) algorithm ..... 69
      4.3.1 EM algorithm for mixture densities .......... 71
      4.3.2 Properties of EM ............................. 73
   4.4 Learning from incomplete data using the EM approach .... 75
      4.4.1 Density estimation for complete data using the EM approach: mixture of Gaussians .......... 76
      4.4.2 Density estimation from incomplete data: mixture of Gaussians ....................... 78
   4.5 Examples ........................................... 81
      4.5.1 Steady state data .............................. 81
6.6.2 Probability distribution function for error ............... 139
6.6.3 The expectation step .................................................. 142
6.6.4 The maximization step ................................................ 143
6.6.5 Initialization of the EM algorithm .................................. 144
6.6.6 On obtaining the global maximum ............................... 146
6.6.7 The overall data rectification scheme ........................... 149
6.7 Examples ........................................................................... 150
6.7.1 The heat exchanger network example .............................. 151
6.7.2 The ammonia synthesis example .................................... 157
6.8 Summary ........................................................................... 161

7 Conclusions and Recommendations for Future Research 166
7.1 Conclusions ................................................................. 166
7.2 Future studies ............................................................ 169
7.2.1 Dimensionality; what is the best approach? ................. 169
7.2.2 Data rectification for dynamic systems ....................... 172
7.2.3 Statistical process control ........................................... 173
7.2.4 Dynamic gross error detection .................................... 173

A Expectation-Maximization .............................................. 188
List of Figures

1.1 Application of plant data rectification [6]. 6

2.1 Process network for a simple example of data rectification. 17

3.1 Process network for the nondiagonal case 52

4.1 EM iteration for a mixture density 72

4.2 Density function for labeled and unlabeled data with various amounts of missing values. One-dimensional case. 87

4.3 PDF for labeled data. Two-dimensional case. 88

4.4 PDF for unlabeled data. Two-dimensional case. 89

4.5 PDF for unlabeled data with 10% missing data. Two-dimensional case. 90

4.6 PDF for unlabeled data with 30% missing data. Two-dimensional case. 91

4.7 Change in log-likelihood as the % missing data increases. One-dimensional case. 92

4.8 Change in log-likelihood as the % missing data increases. Two-dimensional case. 93

4.9 Monotonic increase in the log-likelihood function during the EM iteration. 94

4.10 The heat exchanger network for the data rectification example, Tjoa and Biegler [50] 96

4.11 Measured and rectified values for TA5 97
5.1 Training procedure for estimating probability density estimation using Adaptive Mixtures .................................. 115

6.1 Process measurements, no outlier present. ........................................ 127
6.2 Recursive mean update, no outlier present. ........................................ 128
6.3 Recursive variance update, no outlier present .................................... 129
6.4 Process measurements, outliers present. ......................................... 130
6.5 Recursive mean update, outliers present. ....................................... 131
6.6 Recursive variance update, outliers present ..................................... 132
6.7 Trade off between the likely adjustment of measurements and the likely process rectified states. ....................... 136
6.8 Recursive parameter update for sensor errors, and process states, probability density functions in data rectification. .. 149
6.9 The heat exchanger network for the first data rectification example, Tjoa and Biegler [50]. ................................. 152
6.10 Rectified and measured value of $T_{A8}$, no gross error, the heat exchanger network. ........................................... 154
6.11 The MLR objective function value in the EM iteration. ................. 155
6.12 Rectified and measured value of $F_{C1}$, with a bias of -50 in its measured values, the heat exchanger network. .............. 156
6.13 Rectified and measured value of $F_{B1}$, with a sensor failure, drawn from $N(350,5)$, the heat exchanger network. .......... 157
6.14 Rectified and measured value of process variables $F_{B1}$ and $T_{B1}$, with sensor failures, drawn from $N(350,5)$ and $N(665,5)$ respectively, the heat exchanger network. ............................ 158
6.15 The simplified ammonia synthesis plant, Mah [23]. ....................... 161
6.16 Rectified and measured values of F1, no gross error error present, the ammonia synthesis example. ....................... 162
6.17 Rectified and measured values of F3, no gross error error present, the ammonia synthesis example. ....................... 163
6.18 Rectified and measured values of F5, no gross error error present, the ammonia synthesis example. ....................... 164
6.19 Rectified and measured values of F1, sensor failure for F3, the ammonia synthesis example. ............................. 164
6.20 Rectified and measured values of F3, sensor failure for F3, the ammonia synthesis example. 165
6.21 Rectified and measured values of F5, sensor failure for F3, the ammonia synthesis example. 165
List of Tables

2.1 Measured and Estimated values for the process of figure 2.1 . 17
3.1 Comparison of norms of covariance matrices. Norm . . . . . . . 56
4.1 Estimated Mixture Component Parameters. One-dimensional
    Case .................................................. 85
4.2 Estimated Mixture Density Component Parameters. Two-
    dimensional Case ................................... 86
4.3 Probability distribution properties for process variables . . 95
4.4 Properties of the rectified variables .......................... 95
6.1 Probability distribution properties for process variables of the
    heat exchanger network .................................. 153
6.2 MSE values for measured and rectified process variables for
    the heat exchanger network ............................ 155
6.3 Effect of 'b' and 'p' on process data rectification results, no
    gross error, for the heat exchanger network ................. 156
6.4 Effect of 'b' and 'p' on process data rectification results, one
    failed sensor, for the heat exchanger network ............... 157
6.5 MSE values for measured and rectified process variables for
    ammonia synthesis example ............................ 161
Chapter 1

Introduction

1.1 Introduction

To effectively monitor a chemical processing plant, control product quality and to ensure safe operation, the state of the plant should be known at any given time. To that purpose, a large number of variables are measured in any chemical processing plant. Since all the measurements are prone to random noise and gross errors, the collected measured process data is invariably flawed.

When flawed data is used for state estimation, making operating decisions and process control, the true state of the system is misrepresented. The resulting control performance may lead to suboptimal operation, unsafe process operation, production of off-specification products, financial losses or higher operation costs.

The existence of measurement errors in measured plant values is a well-known problem and many methods have been developed during the past
four decades to resolve this measurement issue. The techniques developed to address this issue constitute what is generally referred to as data reconciliation [1].

The problem of data reconciliation is usually modeled under very restrictive assumptions. The imposed restrictive assumptions facilitate obtaining a solution to the data reconciliation problem. In many cases however, they do not reflect the true nature of the chemical processing plant and as a necessary measure, many auxiliary techniques such as gross error detection schemes are required to treat the measured data, before it can be reconciled to conform to the assumptions that allow for a solution.

This research considers the problem of data reconciliation in a statistical framework such that many of the restrictive assumptions are relaxed and consequently additional data treatment schemes are not required. Posing the data reconciliation problem as such leads to a more general framework and a more reliable and consistent solution.

1.2 Motives

1.2.1 Plant safety

In a typical chemical processing plant a set of operating procedures are designed to ensure the safe operation of the plant. This set of operating instructions outlines the steps that must be taken to prevent hazardous situations. The required safe plant operation is based on the measured plant data that reflects the current state of the plant. Erroneous measurements lead to a
faulty picture of the plant operating state and drawing any conclusion from that state is, therefore, subject to error and may lead to dangerous situations.

For a realistic set of data, all data points must have a believable sign and value [2]. For example, all flows should be positive in the correct direction, or a stream passing through a preheater should not decrease in temperature. For compositions, all fractions should be positive and their sum equal to (almost) one. Unrealistic results may indicate the existence of gross errors in the measured data. It should be noted that the existence of unrealistic data does not necessarily mean that they are the result of gross errors.

The following are typical examples of plant situations that happened as a result of decision-making based on erroneous measured data:

A thermocouple on a reactor, in which an exothermic reaction was occurring, failed by sticking at a low value. Thus the operator decided not to increase the cooling-water flow to the reactor even though the actual temperature in the reactor was steadily increasing. The reaction ran away causing the reactor to explode [3]. In this case if the true temperature had been accurately estimated the operator would have been able to take the correct course of action, i.e. increasing the cooling water, and the explosion would have been prevented.

The accident at Three Mile Island in 1979 was also an example of poor operating decisions being made as a result of faulty process measurements. There were several indications that the level in the primary water circuit was low but two instruments erroneously indicated a high level. The operators believed the high readings and ignored the other indications that implied
that the level was low. The safety procedures were initiated for a situation where the water level was high, but because the actual level was low, the situation deteriorated [3].

Data reconciliation has a significant role to play in preventing decision-making based on erroneous measurements. It can incorporate process information obtained from different sources and weigh the credibility and reliability of the sources and measurements thus providing realistic data on which to base operating decisions.

1.2.2 Efficient plant operation

Product quality and operating costs of a chemical processing plant are a function of operating decisions. Since plant measurements are used to make many operating decisions, decision-making based on erroneous measurements leads to poor process control, decrease in product quality and increase in the plant operation costs.

Sanders (1995) [4] studied key chemical processes at Monsanto Chemical Company to determine the primary cause of upsets. Upsets were defined as excursions that were severe enough to result in restrictions or downgrading of the plant product. It was found that nearly two thirds of the process upsets could be traced to sensor errors. Many of these upsets were caused by fairly large errors due to various types of sensor failures, but errors as small as 1.5% in a flow signal caused a process upset. Often the sensor fault by itself does not cause the upset. It is the operation decision made on these flawed measurements that upsets the process. Consequently it is
of great importance for safe process operation that measured plant data be cast in a framework that removes both the large and the small errors from the measurements so that operating decisions can be made based on a more reliable estimate of the true plant state.

1.2.3 Other issues

Generally, data reconciliation is useful in any aspect of the process that requires measured data and decision making that is based on this measured data.

Other than the above mentioned areas, data reconciliation can also be used for [5];

- Overall mass accounting
- Production planning
- Process monitoring and analysis
- Process control and optimization
- Detection of faulty instrumentation
- Detection of leaks

Figure 1.1 depicts data rectification [6] as a bridge between plant data and many applications in which it can be used.
1.3 A broader look at data reconciliation

The terms data reconciliation and data rectification have been used interchangeably to refer to the same problem. In practice however, data reconciliation is the process of removing the measurement error and adjusting the measured process variables such that process constraints are met.

Data Rectification literally means making the data right. In practice, the objective of data rectification is to estimate the true state of the process from
error-containing measurements, such that they maximize the probability of the process states, $x$, given the plant measurements, $y$. This means by using what is known from the process, i.e. measured data, the probability of what is desired to be known, i.e. process states, is maximized. To that end, using Bayes’ theorem, the data rectification problem is posed as:

$$\max_x p(x|y) = \max_x \frac{p(y|x)p(x)}{p(y)} \quad (1.1)$$

$$= \max_x p(y|x)p(x).$$

where $p(x|y)$ is the probability density of states given the measurements, $p(x)$ is the prior probability density of states and $p(y)$ is the probability density of measurements. $p(y|x)$ is the probability of measurements given the states. This is the probability density of the difference between the plant states and the measurements of these states, hence the probability density of measurement error. The second equality in equation 1.1 holds because $p(y)$ does not depend on $x$ and as a result appears in equation 1.1 as a normalizing factor.

Many of the procedures developed for data rectification can also be classified as data reconciliation. Equation 1.1 is the general equation that describes the data rectification problem. Making the following assumptions leads to a more specific problem of data rectification which is referred to as data reconciliation:

- Measurement error has a zero-centered Gaussian distribution.

- The probability of the states that are represented by a particular set
of equations has a uniform distribution and all the states that satisfy the process constraints are equally likely.

Consequently, data reconciliation does not always provide the most likely estimate of the process states. By applying these two assumptions, the problem posed in equation 1.1 reduces to a constrained least-squares problem. In the case of linear constraints, an analytical solution to the problem exists. The solution, however, is valid only when the assumptions are satisfied. One main issue here is the existence of gross errors in the measured variables and hence violation of the first assumption. To address this issue, many gross error detection techniques have been suggested. The purpose of introducing this extra step is to pre-process the measured data and filter out the gross errors that may exist in the measurements so that the resulting set of data has a zero-mean Gaussian distribution.

Data rectification provides a broader look at the problem posed in equation 1.1 by relaxing the above mentioned assumptions, while data reconciliation represents a special case of data rectification in which the measurement errors are reduced and the measured plant variables are reconciled with the plant model. Data reconciliation provides the most likely estimates of the plant state only if the assumptions are met.

In the data rectification method developed in this thesis the problem of equation 1.1 is solved while the two restrictive assumptions are relaxed and more information is incorporated into the model so that a better estimation of the process states is achieved.
1.4 Outline of approach

In this thesis, a probability based structure is developed to avoid the restrictive assumption that all the states that satisfy the process constraints are equally likely. By incorporating this assumption into the process model, one loses information in two ways:

- It ignores all the information that can be extracted from the process operation and its history and assumes that all the process knowledge is only contained in the process model. In many cases only a partial process model is available or it is too complex to derive a model that describes the correlation among different process variables. A good example for this situation is a distillation column in which the tray temperatures are definitely correlated but it is difficult to model this correlation.

- It is unable to describe the statistical relationships that exist among variables in a deterministic model. The probability density function of the true plant states, however provides the physical and statistical relationships. The information captured in this probability density function, $p(x)$, is then used in equation 1.1 to produce a better estimate of the plant states.

It is proposed that by posing data rectification as the optimization problem of equation 1.1 and using the existing trade off between its terms, the need for auxiliary gross error detection schemes is removed in the sense that measurements containing gross errors are rectified and the error does not
smear the estimation process. As a result the assumption of measurement errors being zero-mean Gaussian distribution is relaxed. A sensor can operate in many modes of operation with the most common one being the normal mode of operation in which the sensor is operating as designed and the measured data contains only random Gaussian noise. The sensor may also operate in failure modes, for example failure due to miscalibration or failure to a fixed value. A probability density for the sensor operation can reflect its normal mode of operation as well as failure modes.

An essential requirement for obtaining a well-structured and accurate probability density is to produce a reliable estimation of the density parameters, namely mean and covariance. Reliability of the covariance matrix is also of great importance in Principal Component Analysis (PCA) [7]. PCA is a size-reduction technique which reduces the dimensionality of a process data set that contains many interrelated variables while maintaining as much of the present variation in the data set as possible. Most of the previous approaches to covariance estimation are either non-robust or depend on the process topology and may fail under some circumstances. In this thesis, a robust direct approach to covariance matrix estimation and estimation of mean as a by-product is developed. It is shown that the proposed covariance estimation method is capable of producing reliable estimates where other methods have failed.

Due to sensor failures, many of the measurement sets contain an incomplete vector of measurements, in the sense that they are missing some measured values. One may simply discard the incomplete vector of measurements
or replace the missing values by their corresponding means. While the first approach simply wastes the collected process data, the latter leads to biased estimates of mean and covariance. Using such a set of measurements in estimating probability densities, will in turn result in an inaccurate density and such a density in equation 1.1 will lead to faulty data rectification. In this thesis, the iterative approach of Expectation-Maximization (EM) [8] is used to replace the missing values such that the resulting mean and covariance are the most likely ones and the limits to the incorporation of incomplete data are defined.

1.5 Research objective

1.5.1 Capturing all process knowledge in a data rectification framework

In most of the data reconciliation methods, the physical process model has been the primary source of process a priori knowledge. For many chemical processing plants, the physical model is either too complicated or has been too simplified to be useful in data rectification schemes. In some cases the model does not exist at all.

By formulating the data rectification problem in the form of equation 1.1, the prior knowledge of the process, not available through the process model, is explicitly incorporated into the data rectification objective function. The formulation of equation 1.1 is also capable of incorporating the process physical model should it become available. In this formulation the primary source
of information is the historic collected plant data that is used to empirically show the relationships that exist among measured process variables.

1.5.2 Measurement error covariance estimation

Calculation of the measurement error covariance matrix is an essential requirement in data reconciliation methods. It is common practice to assume that the measurement errors are normal and have a known covariance matrix. The proposed robust method of measurement error covariance estimation treats the measured process variables directly, as opposed to using residuals of the process constraints (indirect methods). Direct treatment of measured process variables is prone to outliers and consequently there is a dire need for a gross error rejection technique. The proposed method utilizes a generalized maximum likelihood estimator [9] to reject the outliers and hence produces a robust estimate of mean and covariance.

1.5.3 Probability density estimation using incomplete data

An unsupervised method of learning [8] probability density function parameters in the framework of mixture densities from incomplete data is developed. Unsupervised learning can be considered as learning from observations as they are received. In real-world processes, there are many imperfections in these observations. Consequently, there are normally many missing values in any given measurement set.

The EM algorithm is used to iteratively find the maximum-likelihood esti-
mate [10] of the missing values and the parameters of the probability density function. Reliability of the EM algorithm and its convergence properties during the learning process from incomplete data sets is also demonstrated.

1.5.4 Density estimation using adaptive mixtures

In many chemical processing plants, as the process progresses, the operating conditions may change. As such, the new conditions for the plant need to be recognized and the effect of the observed changes has to be considered. Since the learning process for the probability density estimation is an offline process, an efficient on-line technique of density estimation is also needed to incorporate the changes that are reflected in the newly arriving process data.

The method of adaptive mixtures [11] is a probability density estimation technique capable of providing both offline and on-line estimates. It is incorporated into the data rectification scheme so that it is capable of recognizing new trends and updating its parameters accordingly.

1.5.5 Sensor model

In order to capture different types of sensor behavior, such as normal operation and operation in failed modes, with models such that adjustments to the measurements are made to correct errors that are likely, sensor models were also developed. A sensor can fail in a number of ways, including miscalibration, failure to a fixed value, failure by an additive bias and change in sensor parameters.

Defining different modes of sensor operation allow for a variety of different
types of sensor errors to be considered in the rectification process. This aids the generalization of the data rectification scheme and enables it to correct many types of sensor errors.

1.6 Summary of thesis

This thesis is constructed based on four essential elements:

- Maximum likelihood data rectification
- Robust estimation of covariance matrix
- Incorporation of incomplete data
- Estimation of probability density function using the method of adaptive mixtures

The robust direct method of covariance estimation is developed in chapter three. The need for the robust covariance estimation is seen throughout the data rectification process specifically when the measured process data may contain gross errors. It also provides a reliable starting point for PCA analysis.

This thesis casts the data rectification problem in a probabilistic framework and maximizes the probability density of the states given the measurements. The Bayes theorem [12] is used in the maximum likelihood data rectification problem of equation 1.1. The objective function is defined as the product of the probability of the rectified states, based on the prior probability of the plant states, $p(x)$, and the probability of the adjustment made
to the measurements to achieve the rectified states, based on the probability density of the sensor model, $p(y|x)$.

Background and review of data rectification techniques are provided in chapter two. Prior work and relevant literature on other issues explored in this thesis are provided in relevant chapters.

In Chapter three a robust direct approach for estimation of the measurement error covariance matrix is developed and its ability to reject outliers and provide reliable estimate of measurement error covariance matrix is presented.

Chapter four focuses on the estimation of $p(x)$ from an incomplete set of data by using EM algorithm. The missing values are replaced such that the resulting probability density functions are the most likely ones.

Chapter five is devoted to the method of adaptive mixtures. This method is used to estimate and update the probability density of the plant states, as they are received. This recursive non-parametric method of probability density estimation and its merits are thoroughly explored in chapter five.

Chapter six presents an explanation of the sensor model and the method of solving the maximum likelihood objective function. Two benchmark examples are provided and the results of the data rectification process are discussed.
Chapter 2

Data Reconciliation

Background

2.1 Introduction

In a modern chemical plant, process measurements are used in a variety of aspects such as plant safety, process control, process optimization, process monitoring and performance evaluation. The presence of errors in measurements leads to inaccurate process data, which fail to satisfy plant material and energy balances. Low quality data can easily lead to poor decision making and control, which in turn, may affect other plant operating decisions.

Consider a simple process consisting of a single vessel with one input and two outputs [13] as depicted in figure 2.1. Assume that the feed rate is 100 ton/hour. Based on the conservation law of mass it can be concluded that:

\[ F_2 + F_3 = F_1 = 100. \]  

(2.1)
Also assume that the concentration of one of the components in the process

![Diagram of process network](image)

**Figure 2.1:** Process network for a simple example of data rectification.

is measured and two sets of measurements for that variable are provided as in table 2.1.

<table>
<thead>
<tr>
<th>Stream</th>
<th>Measured Concentration</th>
<th>Estimated Concentration</th>
<th>Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>0.238</td>
<td>0.521</td>
<td></td>
</tr>
<tr>
<td>$F_2$</td>
<td>0.053</td>
<td>0.407</td>
<td>61.93</td>
</tr>
<tr>
<td>$F_3$</td>
<td>0.539</td>
<td>0.634</td>
<td>38.07</td>
</tr>
</tbody>
</table>

Mass conservation also holds for the component balance. The corresponding balances for each set of measurements are as in equations 2.2a and b:

\[
0.053 F_2 + 0.539 F_3 = 0.238 F_1 = 23.8 \quad (2.2a)
\]

\[
0.407 F_2 + 0.634 F_3 = 0.521 F_1 = 52.1 \quad (2.2b)
\]

Solving the system for two sets of equations leads to the estimated values for $F_2$ and $F_3$ as are presented in table 2.1. Since the problem is overdefined and due to errors in the measurements, the resulting estimated values are inconsistent. The objective of data reconciliation is to utilize the redundancies
available through measurements and find a set of estimated values for process variables in a way that the conservation laws are met and the existing error is minimized.

The first and most common source of error is random noise associated with the measurement signal. Noise is typically assumed to be zero-mean and normally distributed. This type of error is usually attributed to the irreproducibility of the measurement devices.

The second type of error is a systematic bias. These occur when measurement devices produce consistently erroneous values, either high or low. In this case the expected value of error is not zero. Bias may arise from sources such as incorrect installation, improper calibration of measurement devices or long term drifts.

The third type of measurement error is the gross error and is usually caused by non-random events. In the case of a gross error occurring, the measurement values have little or no relation to the true values. Measurement-related errors such as malfunctioning sensors and process-related errors such as leaks are examples of gross errors.

Many process control and optimization activities are also based on small improvements in process performance. Existence of errors in process data can easily exceed and mask actual changes in process performance.

The task of improving the accuracy of process data such that it is consistent with mass and energy balances for the process is known as “data reconciliation” [14]. In a sense, the problem of data reconciliation is as old as the process industries themselves. In the absence of automatic process
measurement and data acquisition, plant operators and engineers collect the process data and check the accuracy and consistency of the collected measurements based on their judgement which in turn is based on their knowledge and experience with the plant.

Automating process data collection produced far greater volumes of collected data. The requirement to manage this large volume of data was also dramatically increased. The down side of voluminous data collection was the inability to maintain human vigilance and intervention. The volume of data is now so large it makes human vigilance infeasible if not impossible. Therefore measured process data must be treated in a systematic and rational way. Posing the problem of data reconciliation in such a reliable framework, ensures obtaining a reliable set of data that can approximate the true process values.

Data rectification refers to the class of problems aimed at "making the data correct" [15]. Data rectification is the general framework for problems of this type but it is data reconciliation that has received most of the research interest. Data reconciliation constitutes a special case of data rectification in which adjustments are made to data such that they are fitted to the model that describes the process and its constraints.

The research on data reconciliation can be divided into two main categories, linear and nonlinear, both have been conducted on steady state and dynamic systems. Steady state linear data reconciliation problems have attracted the major part of the research effort and published articles abound in the literature [15]. Mass balance networks are the type of real processes
that can be easily adapted into the specifications of a linear steady state data reconciliation problem. Although there has been a considerable increase in research aimed at nonlinear data reconciliation, the extent is small when compared to the literature on linear systems [15]. In many cases, the techniques developed for nonlinear systems can also be implemented with linear versions of those systems.

The application of artificial intelligence in data reconciliation has recently become a topic receiving research time and some studies have been published in this area as well [15]. Another trend in data rectification is posing the problem in a statistical framework. Given the measurements, the probability of the true plant states are maximized and the set of true plant states that are most likely are determined.

2.2 Linear data reconciliation

The subject of data reconciliation was first introduced in Chemical Engineering processes by Kuehn and Davidson [16]. In their seminal paper they proposed a minimum weighted least squares approach in which flow and enthalpy measurements were adjusted to satisfy heat and material balances. The objective function composed of the sum of squares of adjustments, each weighted by the inverse of its measurement noise variance, is minimized so that the smallest adjustment that satisfies the constraints is found. The
problem is mathematically stated as follows:

\[
\min_{\hat{x}} (y - x)^T Q^{-1} (y - x).
\]

\[\text{s.t. } Ax = c.\] (2.3)

where \(x\) is the vector of adjusted values, \(y\) is the vector of measurements, \(Q\) is the covariance matrix of the measurement noise, \(A\) represents the linear constraints and \(c\) is a vector of constants. Using the method of Lagrange multipliers, an analytical solution to the problem posed in equation 2.3 is found as:

\[
x = y - QA^T (AQAT)^{-1} (Ay - c).
\]

With the availability of an analytical solution, this data reconciliation method is fast and suitable for on-line implementation.

A basic assumption in this method is that the measured data contain only random errors, \textit{i.e.} zero-mean normally distributed errors. That assumption also holds true for an error that is a culmination of many errors, since its probability density tends toward a Gaussian distribution as the number of error components increase, regardless of the component distribution [12]. When the normality assumption is true (which is the case most of the time), the weighted least squares estimation is the same as the maximum likelihood and minimum variance estimation, and gives rise to unbiased estimates [17].

The weighted least squares method for solving linear data reconciliation problems has difficulty if all of the process variables that appear in the constraints are not measured. Crowe et al. [18] used a projection matrix to address this issue. The constraint matrix, \(A\), is partitioned into two matri-
ces, $A_1$ for the measured variable (denoted as $x$) and $A_2$ for the unmeasured variables (denoted as $u$). The process constraints of equation 2.3 can then be rewritten as:

$$A_1x + A_2u = c.$$  \hspace{1cm} (2.5)

where $c$ is a constant vector that represents any measurements that are known exactly (with 0 as its special case). The projection matrix, $P$, is defined as a matrix whose rows are orthogonal to column space of $A_2$ and as a result:

$$PA_2 = 0.$$  \hspace{1cm} (2.6)

Premultiplying equation 2.5 by $P$, reduces the general linear reconciliation problem to one that involves no unmeasured variables:

$$PA_1x = Pc$$  \hspace{1cm} (2.7)

The least squares solution can then be found for the problem stated by equation 2.7.

In the presence of gross errors the weighted least squares method results in biased estimates, as measurements containing gross errors are considered to be true values that contain some noise. As a result, it is necessary to detect the existing gross errors and eliminate their effects before reconciling the measured data. Ripps [19] realized that gross errors would invalidate the weighted least squares procedure giving poor reconciled values and thus should be removed before the data reconciliation was performed. Ripps re-
2.2 Linear data reconciliation

Results showed that if a measurement containing gross error was allowed into the reconciliation scheme, the resulting variable estimates would contain significant errors, with the entering gross error affecting some or perhaps all of the estimates. This effect is appropriately referred to as smearing [19]. When data containing gross errors is subjected to constrained least squares reconciliation, a large adjustment is needed for making a single large correction, because the corrections are squared when they appear in the objective function. Thus the presence of a gross error in one measurement causes a series of small corrections in other measured variables hence giving rise to bad estimates of the true values. Nogita [20] developed the serial elimination method for removing data containing gross errors. The procedure calculates a test statistic\(^1\) (a weighted least squares objective function resulting from serial elimination of suspected measurements one at a time) based on the adjustments and eliminates the erroneous measurements one at a time, each time calculating the test statistic for the reduced set of measurements. The set of measurements that pass the no gross error test are considered to be gross error free. Serial elimination is most effective if a single gross error exists in the measured data. When multiple gross errors exist in a measurement set, the possible combination of suspected measurements quickly increases and consequently the speed and computational efficiency of serial gross error elimination decreases.

Essentially all gross error detection schemes for linear steady state systems follow the same pattern. A test statistic is calculated and compared

---
\(^1\)A test statistic is a quantity calculated from sample data that is used as a basis for reaching a decision in a hypothesis testing.
against a critical value given a certain level of significance (for example, comparison to a $z$-value for a test statistic following a normal distribution or a $\chi^2$-value for chi-square distribution). Almasy and Sztano [21] classified an error to be regular (zero-mean Gaussian) or extreme by means of a chi-square test. A heuristic method developed by Vaclavek as well as other early gross error detection schemes are reviewed by Hlavacek [22]. Gross error detection techniques are generally categorized into those that directly detect the measurements containing gross errors and those that are tools for further identification of gross errors such as the serial elimination method.

The most common methods of direct gross error detection are global test, nodal test and measurement test [23]. Global and nodal tests investigate how well the measurements satisfy the constraints and utilize the constraint residuals to detect the existence of gross errors. These procedures do not need the data to be reconciled first [24].

The global test was first published by Almasy and Sztano [21]. This test combines all of constraint residuals, weighted by their covariance, into a $\chi^2$ variable and compares it to a critical value (obtained from the table of cutoff points for the chi-square distribution) to decide whether a gross error exists. For the constrained linear data reconciliation problem formulated as per equation 2.3, the vector of residuals, $r$, is defined as:

$$ r = Ay - c. $$ (2.8)
and residuals covariance matrix is calculated as:

\[ J = \text{cov}(r) = AQA^T. \] (2.9)

The test statistic for the global test is then defined as \( r^T J^{-1} r \).

The nodal test was first published by Mah et al. [25]. The nodal test examines each constraint equation separately. The test statistic for each constraint \( j \), is defined as \( |r_j|/\sqrt{J_{jj}} \), i.e. the constraint’s residual over its standard deviation. This test statistic follows a normal distribution and is compared to a critical \( z \)-value to determine if a gross error exists.

The main drawback for both the global test and the nodal test is that although they can determine whether gross errors exist in the data set, they are unable to pinpoint the measurements that contain the gross errors. Hence, additional gross error detection schemes are required to find the erroneous measurements. Mah and Tamhane [24] [26] proposed the measurement test which checks at the size of the adjustment made to a measurement during data reconciliation and on that basis, decides if a gross error exists in that measurement. Its test statistic is defined as:

\[ |z_i| = \frac{|d_i|}{\sqrt{W_{ii}}}. \] (2.10)

for the \( i^{th} \) measurement, where \( d_i \) is the \( i^{th} \) element of

\[ d = Q^{-1} \mathbf{a}. \] (2.11)
2.2 Linear data reconciliation

$a$ is the vector of adjustments and is defined as:

\[ a = y - x. \]  \hspace{1cm} (2.12)

and $W$ is the covariance matrix of $d$, $W = \text{cov}(d)$. The measurement test addresses the shortcoming of the global and nodal tests in that it can locate the erroneous measurements. This test, however, requires the data to be reconciled before any gross error detection takes place. Iordache et al. [27] found that the performance of the measurement test depends both on the absolute size of the gross error and the ratio of the size of the gross error to that of the noise standard deviation of the measurement. Mah and Tamhane [26] showed that the measurement test has the maximum power, in the sense that the test has a greater probability of finding a single gross error in a particular measurement than any other test based on a linear combination of the measurements.

There are many other gross error detection tests, some are new and others expanded on previously developed tests. Narasimhan and Mah [28] used the statistical technique of generalized likelihood ratio to develop a general framework for identifying gross errors that can be modeled. Rosenberg et al. [29] recognized that it is possible for the results produced by data reconciliation schemes to be physically impossible, such as negative concentrations, and modified the measurement test to incorporate bound violations. They developed two types of measurement tests: the extended measurement test that uses a serial elimination scheme and the dynamic measurement test.
that again uses the serial elimination test but the candidate set of potential gross errors can be updated as the procedure continues. Romagnoli and Stephanopoulos [30] used structural analysis of the balance equations, along with statistical testing and sequential analysis of the identified candidate set to detect gross errors. Tamhane [31] used a Bayesian approach to incorporate past information about sensor failure rates into the gross error detection process. Mah et al. [25] also used graph theory to break down and solve problems with missing measurements. Rollins and Davis [32] developed an unbiased estimation technique to deal with any biased process measurements.

There are two common assumptions in all of these data reconciliation and gross error detection and identification techniques:

- The process model can be described by a set of linear constraints.
- The noise covariance matrix of the data is known.

It is not always possible to model the process by appropriate linear constraints. Narasimhan and Mah [33] developed a method for taking generalized steady state models, including unmeasured and indirectly measured variables, and put these into a form containing measured variables only. The algorithm is based on partitioning the constraint matrix into parts corresponding to measured, unmeasured and indirectly measured variables, premultiplying it by a projection matrix to remove the unmeasured variables (as suggested by Crowe et al. [18]), and manipulating the resulting system so that indirectly measured variables can be expressed in terms of the measured variables.
2.3 Nonlinear data reconciliation

The noise covariance matrix is generally assumed to be known. In practice, however, that is not always the case and hence it must be estimated. One way that the noise covariance can be estimated is to hold a plant at steady state for a long period of time, take many measurements and use these to estimate a noise covariance matrix. Since it is virtually impossible to maintain a plant precisely at steady state for any length of time, the estimate of the noise covariance matrix would be erroneous due to departures from steady state. A comprehensive review of the existing approaches that address this issue will be provided in chapter 3.

2.3 Nonlinear data reconciliation

The most common approach to the data reconciliation problem with nonlinear constraints is to linearize and use one of the linear techniques [17]. Linearization itself is an approximation and may introduce considerable error into the process by forcing the measurements to conform to the linearized process model. Knepper and Gorman [34] estimated measured and unmeasured variables using a Gauss-Newton type iterative algorithm for nonlinear constraints that were approximated by a Taylor series expansion. The solution procedure generally involves the successive linearization of the constraints and solution of the resulting linear data reconciliation by using Crowe's projection matrix technique [35]. It should be noted however, that the projection matrix has to be completed at every iteration and can require considerable computation time [14].
2.3 Nonlinear data reconciliation

For bilinear processes, where balance equations in general contain products of pairs of unknowns, a simplified two-step projection matrix technique has been developed by Crowe [36]. Although the method is iterative, the two projection matrices are computed only once and not in every iteration. This helps to reduce the computation time significantly [14].

Meyer et al. [37] have developed a method that eliminates all unobservable variables and simultaneously identifies those equations that are part of the reconciliation for a multicomponent process and the resulting reduced problem is solved iteratively. Since only unobservable variables are eliminated and not all unmeasured variables, the resulting reconciliation problem is larger. Furthermore, since no special technique is used to solve the reconciliation problem, the method is likely to be less efficient than Crowe's two-step projection matrix method.

A completely novel approach has been proposed by Sirnpson et al. [13] to solve the bilinear data reconciliation problems arising in mineral processing industries. Instead of obtaining a reduced set of constraints involving only measured variables, their approach completely eliminates all constraints. This is accomplished by dividing the set of variables into dependent and independent variables. The constraints are utilized to obtain an explicit relationship between the dependent and independent variables, which is used to eliminate all dependent variables from the objective function and obtain a reduced unconstrained problem in the space of independent variables.

General nonlinear data reconciliation has also been a target for many investigations. Error-in-Variables (EIV) has been one of the most successful
approaches in this field. Britt and Leucke [38] presented a general algorithm for the EIV method. The objective function was optimized using Lagrange multipliers and the constraints were successively linearized with respect to the parameters and measured variables. In this approach, both the parameter estimates and the reconciled measurements were obtained simultaneously.

In the EIV approach, measurement errors in all variables are treated in the calculation of parameter estimation. The EIV is extremely useful when the form of the process model is known but the parameters of the model are unknown. The model equations are in the form of the measured dependent variables, \( y \), being a function of measured independent variables, \( x \), and the set of unknown parameters, \( \theta \):

\[
\hat{y} = f(x, \theta).
\]  

(2.13)

Assuming that there are no errors in the independent variables, the measurements can be used to find the maximum-likelihood estimates of the model parameters [39]:

\[
\min_{\theta} e_y^T Q^{-1} e_y.
\]  

(2.14)

where \( e_y \) is defined as:

\[
e_y = y - \hat{y} = y - f(x, \theta).
\]  

(2.15)

and \( Q \) is the covariance matrix of the dependent variables. For almost all systems, there are errors in both dependent and independent measured vari-
2.3 Nonlinear data reconciliation

ables. This is reflected in the EIV approach by forming an error vector consisting of the errors in the measured variables and model residuals:

\[ e = \begin{bmatrix} e_x \\ e_y \end{bmatrix}. \]  \hspace{1cm} (2.16)

where:

\[ e_x = x - \hat{x}. \]  \hspace{1cm} (2.17)

The parameters and reconciled \( x \) values are then solved by solving the following optimization problem [39]:

\[ \min_{\hat{\mathbf{x}}, \hat{\mathbf{\theta}}} \mathbf{e}^T \mathbf{V}^{-1} \mathbf{e}. \]  \hspace{1cm} (2.18)

where the weighting matrix, \( \mathbf{V} \), is defined as:

\[ \mathbf{V} = \begin{bmatrix} \mathbb{E}(e_x e_x^T) & \mathbb{E}(e_x e_y^T) \\ \mathbb{E}(e_x^T e_y) & \mathbb{E}(e_y e_y^T) \end{bmatrix}. \]  \hspace{1cm} (2.19)

The reconciled \( y \) values are then calculated using reconciled \( x \) values and estimated values for parameters are found as in equation 2.13. Deming [40] solved the problem posed in equation 2.14 using the method of Lagrange multipliers with linearization of the model at the experimental measurements. Similar approaches have been developed by Peneloe et al. [41], Sutton and McGregor [42] and Reily and Patino-Leal [43]. Powell and Macdonald [44] used a Newton-Raphson approach. Others have focused on decomposing the problem into subproblems and solving them in a reduced
2.3 Nonlinear data reconciliation

space [39] [45] [46] [47]. The performance of EIV solution techniques has been evaluated by Ricker [48]. Almost all of the EIV solution methods experience convergence problems in certain situations. The problem can be overcome by using nonlinear programming (NLP), but the computational cost can be prohibitively expensive, even for simple models. Thus data focus in data reconciliation has been on solution techniques that are fast, hence can be applied in real time [15]. Tjoa and Biegler [49] developed a decoupled sequential quadratic programming (DSQP) method to solve the EIV problem. The solutions obtained were close to the NLP solutions, but the DSQP solutions were found in a fraction of the time compared to that of NLP.

Tjoa and Biegler [50] developed a consolidated approach to data reconciliation and gross error detection, that eliminated the gross errors in the reconciliation step and lead to unbiased estimates. This was accomplished by using a bivariate objective function. Instead of strictly minimizing the adjustments, as in the weighted least squares approaches, a probability density of the adjustment is formed, and it is maximized subject to the constraints. That is, the most probable adjustments that satisfy the constraints are used to find the reconciled values. The probability density of the adjustments contained two parts. One is a narrow Gaussian distribution centered at zero, representing normal process noise. The other is a broad distribution also centered at zero, that made large adjustments probable. The two Gaussian distributions are weighted by the probability of not having a gross error and the probability of having a gross error, respectively. The distribution for normal noise or the gross error is the distribution that the final adjustments are
most likely coming from. That distribution is then used to identify possible
gross errors [15].

2.4 Knowledge based data reconciliation

The data reconciliation methods reviewed earlier have all approached the
data reconciliation problem assuming that an adequate model of the process
is available for use in the reconciliation schemes. Often this assumption is
not true. The knowledge based approaches try to "learn" about the process
through the use of historical plant data and use that knowledge to solve the
data reconciliation problem.

Dobrzeniecki and Lindsey [51] developed a sensor reasoning and abstrac-
tion with neural networks (SRANN), to identify gross errors based on sensor
signal patterns. The SRANN is a fully connected feedforward network that
is trained with examples of normal and abnormal (presence of gross errors)
wave forms for a sensor. The SRANN then classifies the sensor wave form
as either normal or abnormal and notifies the operators of potentially faulty
sensors. Karjala and Himmelblau [52] reported limited success using recur-
rent Elman neural networks [53] for rectification of time series of data. In
the Elman net, the output of the hidden layer nodes at the current time step
are fed back to the input layer and to what are called context nodes, and
then used in the feedforward propagation of the network in the next time
step. This structure is reported to have the effect of retaining more knowl-
edge about the process, however, considerably more development is required
in this approach before definite conclusions can be drawn [15].

Kramer [54] [55] developed autoassociative neural networks (AANN) for data rectification. The AANN is a fully connected feedforward neural network with three hidden layers. The first layer is the mapping layer that maps the input space to a lower dimensional bottleneck space, represented by the middle layer. The third hidden layer is the de-mapping layer that maps the bottleneck space to the output space. The network is trained to learn the identity function, i.e. input equals output. The compression of the data through the bottleneck layer allows for the removal of noise from the data. Any neural networks performance greatly degrades outside the region of the data it was trained with and consequently AANN does not perform well in the presence of gross errors. By artificially corrupting measurements in the training set and thus increasing the size of training set, mapping from data with gross errors to their correct values can be used to train the neural network. This results in robustness of the AANN and improves its performance in the presence of gross errors.

Johnston and Kramer proposed a statistical framework for steady state data rectification [1]. Their maximum-likelihood rectification objective function is a product of the probability of the true plant states and the probability of the adjustments made to the plant states. The knowledge of the probability of the true plant states is obtained by learning the parameters of the probability density from the historical plant data. They developed Elliptical Basis Functions (EBF) networks as a nonparametric probability density estimation technique [56] and used these in the learning process for the true
2.5 Summary

Erroneous measured process values result in inconsistencies in mass and energy balances and lead to poor operating decision-making and faulty process control. Data rectification is the estimation of the most likely values of the measured process variables such that the process constraints are met and measurement error is reduced.

Data reconciliation as a specific case of data rectification and its linear steady state case have been the focal point of research in this field. Linear steady state data reconciliation is generally based on a weighted least squares objective function minimized subject to the process constraints.

A key assumption in data reconciliation is that only random errors are present in the measured data. As such, gross error elimination has become an inseparable part of data reconciliation and a point of research interest.

The most common approach for nonlinear data reconciliation has been linearization and implementation of an existing linear technique to the linearized problem. The method of Error-in-Variables provides a more general solution to nonlinear data reconciliation. In the EIV, measurement errors in all variables are treated in the calculation of parameter estimates. EIV provides both parameter estimates and reconciled data estimates.

In the absence of process models, knowledge-based data rectification approaches replace the model-based ones. In these approaches, the plant history
is used to learn the parameters of the model. The model, used to approximate the process behavior, and the plant data are then used alternatively to rectify the measured process values and update the model parameters.
Chapter 3

Estimation of Measurement

Error Covariance Matrix

3.1 Introduction

Covariance/Correlation matrices are necessary tools in multi-variate data analysis. If a multi-variate observation is taken as a point in a $d$-dimensional space, a total of $n$ observations would create a cloud [57] in that space. The objective of multi-variate data analysis is to find and describe the structure of the cloud. Covariance/Correlation matrices are a basic requirement for analyzing the reliability, dispersion and/or association of the data cloud.

Reducing the size of high dimensional data and its projection to a lower dimensional subspace is a necessary procedure in many multi-variate problems. Principal component analysis (PCA) is one of the frequently used techniques in reducing the dimensionality of the data. The basic idea of PCA is to reduce the dimensionality of a data set which contains a large
number of correlated variables, while maintaining as much variation present in the data set as possible [7]. PCA accomplishes this by creating a new set of variables, principal components, which are uncorrelated. In order to achieve this objective, PCA depends on a reliable estimate of the covariance/correlation matrix. Covariance/correlation matrices are unfortunately highly sensitive to the presence of outliers. Even the presence of a few outliers will lead to a faulty evaluation of the principal components [58] and in turn, to inappropriate size reduction.

The covariance matrix and principal components are both necessary instruments in performing classical and probabilistic data reconciliation. Data reconciliation is one application in which data reliability is of critical importance. Data reconciliation is the process of adjusting measured values of variables, such as flows, compositions and temperatures, to satisfy mass and energy balances. Hence, data reliability/quality is an important issue and a reliable method for assessing the measured data is required.

The classical methods for calculating covariance matrices are not robust [59], and as a result, highly vulnerable to the presence of outliers in the process measurements. Errors in covariance matrices can lead to incorrect data reconciliation and may result in poor control of the process.

In the following, the existing methods for covariance matrix calculation are discussed. After stressing the shortcomings of the existing methods, the proposed Robust Direct method for Covariance Estimation (RDCE) is outlined. Finally, two examples comparing the covariance matrix estimated by the existing methods and the RDCE are provided.
3.2 Overview of existing methods

Generally speaking, the methods for covariance matrix estimation can be divided into Direct and Indirect methods. In the direct method, assuming the process is truly at steady-state, the mean and covariance matrix can be estimated by:

\[ \bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}. \]  

(3.1)

\[ \text{cov}(x_i, x_j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j). \]  

(3.2)

where \( n \) is the width of the window in which the process is at steady-state. In a strict sense, the steady-state condition almost never prevails [23]. In practice, there are always deviations from steady-state and transitions from one steady-state to another. It is only in the framework of a time interval and accepted variations that steady-state finds meaning. In the presence of variations in steady-state values of process variables, the direct method would produce poor estimations of both the mean and covariance matrix. Furthermore, equations (3.1) and (3.2) are an unbiased estimation of the mean and covariance matrix, assuming that only normally distributed random errors are present. A gross error in the measurements would propagate throughout the estimation process.

These drawbacks are addressed in the indirect and robust indirect methods (Almasy, Mah, 1984; Darouach et al., 1989; Keller et al. 1992; Chen et al., 1997). Indirect methods try to mitigate these issues by incorporating
3.2 Overview of existing methods

additional process information, namely the mass balance equations.

A linear system can be stated as shown in equation (3.3);

\[ A x_k = 0. \]  \hspace{1cm} (3.3)

where \( A_{m \times d} \) is the incidence matrix and \( x_k \) is the vector of true values and of length \( d \). Assume a measurement model of the form shown in equation (3.4);

\[ \bar{x}_k = x_k + e_k. \]  \hspace{1cm} (3.4)

where \( \bar{x}_k \) is the vector of measured values and \( e_k \) is the vector of zero-mean, normally distributed errors. Defining the vector of residuals as:

\[ r_k = A \bar{x}_k. \]  \hspace{1cm} (3.5)

and incorporating equations (3.3) and (3.4) into (3.5) results in equation (3.6).

\[ r_k = A e_k. \]  \hspace{1cm} (3.6)

Since \( e_k \) is of zero-mean and normally distributed:

\[ E(r_k) = E(A e_k) = A E(e_k) = 0. \]  \hspace{1cm} (3.7)

The covariance of balance residuals, \( H_{m \times m} \), can be found as follows:
\[ H = \text{cov}(r_k) = E[(Ae_k)(Ae_k)^T] = AE(e_k e_k^T)A^T. \quad (3.8a) \]
\[ H = AVA^T. \quad (3.8b) \]

where \( V_{d \times d} \) is the covariance matrix of the measurement error. In the Indirect methods [60], the covariance matrix of residuals is then used to estimate the measurement error covariance matrix:

\[ \text{vec}(H) = (A \otimes A)\text{vec}(V) \quad (3.9) \]

\( \text{vec}(H) \) is defined as \( \text{vec}(H) = [h_1^T h_2^T \ldots h_m^T]^T \) where \( h_j \) is the \( j^{th} \) column of \( H \) [61]. Let's redefine equation (3.9) as

\[ \text{vec}(H) = GX_G. \quad (3.10) \]

For the general case where \( N \) off-diagonal elements exist in the covariance matrix, \( X_G \) is of length \( d + N \) and is defined as:

\[ X_G = \text{vec}(V) = [v_{11} \ v_{22} \ldots \ v_{dd} \mid v_{ij} \ldots v_{kl}]^T. \quad (3.11) \]

and \( G \) is of size \( m^2 \times d + N \) and is defined as:

\[ G = A \otimes A = [c_1 \ c_2 \ldots \ c_d \mid c_{ij} \ldots c_{kl}]. \quad (3.12) \]

where \( c_i = [a_{ii}A_i \ldots a_{ni}A_i]^T \) and \( c_{ij} = [a_{ji}A_i \ldots a_{ni}A_i]^T. \)
3.2 Overview of existing methods

The estimability conditions of covariance matrix may be expressed as:

\[ \text{rank}(G) = d + N. \]  \hspace{1cm} (3.13)

For steady-state cases, Almasy and Mah [61] suggested an element-wise method in which the diagonal and off-diagonal elements of the covariance matrix can be found by means of constrained minimization of the sum of squares of the off-diagonal elements of the measurement covariance matrix.

Darouach et al. [59] proposed a method for estimating a diagonal covariance matrix based on a maximum likelihood estimator, the material balance constraints and the statistical properties of their residuals. Their algorithm couples the two phases of covariance matrix estimation and data reconciliation. They solve a non-linear optimization problem and obtain an iterative solution for variance estimation, so that the data reconciliation process is hindered by both computational and convergence problems.

Keller et al. [62] have extended the work of Darouach et al. [59] to covariance matrices that are not strictly diagonal. They present an algorithm based on the relations deduced from statistical properties of the material balance constraints in order to estimate the covariance matrix of the measurement errors. They suggested, based on simulation results, that their method should be used in practical applications.

All of the above mentioned approaches are sensitive to the presence of outliers in the measurements and as a result, a few outlying observations lead to an incorrect estimation of error covariance matrix.
3.2 Overview of existing methods

Chen et al. [60] [63] proposed a robust method based on the M-Estimate approach, originally outlined by Huber [9]. In this method, each vector of measured variables, depending on its distance from the median of observed values, is given a weight between zero and one. The closer the observation to the median, the closer the weight to one and the further away the observation, the closer the weight is to zero. In this way, the effect of outliers is minimized.

Utilizing additional process information in Indirect methods is limited to linear constraints. Incorporation of nonlinear constraints would make the calculation of the covariance matrix a computational challenge. As indicated in equations (3.9) through (3.13), if matrix $G$ is not a full-rank matrix, $\hat{X}_G$ cannot be estimated from $\hat{X}_G = (G^T G)^{-1} G^T vec(H)$ and hence the covariance matrix cannot be estimated as in $\hat{V} = vec^{-1}(\hat{X}_G)$. Since the arrangement of $G$ depends on the incidence matrix (process topology) and the correlated measurements, the possibility of obtaining a rank-deficient $G$ always exists. Should $G$ become ill-conditioned or singular, it would negate the covariance matrix calculation. Another area requiring attention are the off-diagonal elements of the covariance matrix. All indirect approaches require a priori information that describes which variables are correlated.

The M-estimator in the Chen et al. [60] [63] robust method was used in such a way that once any element of a vector of measured variables is found to be out of an acceptable range, a small weight is given to the vector. This leads to loss of some valuable acceptable measurements because they are in the same vector with the faulty measurements.

In the RDCE, the M-estimator with an iterative scheme for rejecting out-
liers is applied directly to the measured process values. In this way, because of the weighting process, the need for incorporating residuals is removed. In addition, the a priori information for locating the correlation between measurements and hence, off-diagonal elements of the covariance matrix, is no longer required.

The RDCE also uses weights both, vector-wise and element-wise. First, the vector of measurements is evaluated. If no error is detected, a large weight is assigned to the whole vector. Should an outlier be detected, an element-wise search is conducted within the vector. A small weight is assigned to the outlying elements with large weights being assigned to the remaining elements. This process is further explained in the next section.

3.3 Outline of the new approach

3.3.1 The maximum likelihood estimator

The maximum likelihood estimator or M-estimator, first suggested by Huber [9], is a very robust estimator. Each vector of measurements, based on its Mahalanobis distance, is assigned a weight. These weights determine the influence of each vector so that they approach zero as the measurement error vector becomes less characteristic [63].

Let $X = [x_1 x_2 ... x_n]$ be a matrix of $n$ observations of $d$ dimensions. The maximum likelihood estimate of the vector of means, $m$, and the covariance matrix, $V$, can then be estimated by simultaneously solving the following system of equations [64] [65]:
\[ \frac{1}{n} \sum_{i=1}^{n} h_1(d_i)(x_i - m) = 0. \]  
\[ (3.14) \]

\[ \frac{1}{n-1} \sum_{i=1}^{n} h_2(d_i)(x_i - m)(x_i - m)^T = V. \]  
\[ (3.15) \]

where \( h_1 \) and \( h_2 \) are weights and

\[ d_i^2 = (x_i - m)^T V^{-1} (x_i - m). \]  
\[ (3.16) \]

is the Mahalanobis distance of the \( i^{th} \) vector of measurements that gives its square distance from the current estimation of the mean, scaled by the current estimation of the variance. If \( h_1 = h_2 = 1 \), then equations (3.14) and (3.15) will reduce to equations (3.1) and (3.2).

### 3.3.2 Mean and covariance estimation algorithm

**Initialization**

The number of observations and measurements are represented by \( n \) and \( d \) respectively, so that the matrix of observations, \( X \), is \( n \times d \). Initialization of the vector of means, \( m \), and the covariance matrix, \( V \), is as follows:

\[ m_j = \text{median}_i (x_{ij}) \quad i = 1, \ldots, n \quad j = 1, \ldots, d. \]  
\[ (3.17) \]

where \( m_j \) is the \( j^{th} \) element of the vector of means. For the covariance matrix:

\[ t_j = 1.483 \text{ median}_i (|x_{ij} - m_j|). \]  
\[ (3.18) \]
where \( t_j \)'s are called scales and 1.483 is a correction factor to make the estimator consistent with the ordinary parameters of the Gaussian distribution [57].

The strictly diagonal initialization of covariance matrix is then defined as equation (3.19):

\[
V = \begin{bmatrix}
    t_1^2 & & \\
    & t_2^2 & \\
    & & \ddots \\
    & & & t_d^2
\end{bmatrix}.
\]  

(3.19)

The Huber-type weighting scheme [9] is employed so as to make \( V \) an asymptotically unbiased estimator of the covariance matrix. The first and second Huber-type weights are defined as follows:

\[
h_1(d_i) = \begin{cases} 
    1, & d_i \leq k. \\ 
    k/d_i, & d_i > k.
\end{cases} \quad (a) \\

(3.20)

\[
h_2(d_i) = (h_1(d_i))^2 / \beta. \quad (b)
\]  

(3.21)

where \( k^2 \) is the 90\% point of \( \chi^2_p \) distribution with \( p \) degrees of freedom. The degrees of freedom is determined by the dimension of the observation vector.

There will be two \( k \)'s, namely \( k_1 \) and \( k_2 \), corresponding to two different weight calculations (vector-wise and element-wise), differing from each other in the degrees of freedom used in the calculation of \( k \). \( \beta \) is defined by equation (3.22):

\[
\beta = G(k^2/2, 1.5) + 2k^2(1 - \Phi(k)).
\]  

(3.22)

where \( \Phi(k) \) is the standard normal distribution function and \( G(x, f) \) is the
3.3 Outline of the new approach

Gamma distribution function with \( f \) degrees of freedom [66]. The first Huber-type weight is defined as matrix \( H_1 \) of size \( n \times d \), initialized as a matrix of ones.

Iterative computations

In this part, the Mahalanobis distance for each vector of measurement is calculated and compared against a pre-determined value and Huber-type weights are assigned to the vector. If an outlier flags, the elements of the vector of measurement is further investigated and a second class of Huber-type weights are used to isolate the erroneous elements of that vector.

The new values for mean and covariance matrix are then calculated and the convergency is checked. These steps are explained more in the following:

- For each observation entry, calculate the Mahalanobis distance using equation (3.16). This value, when compared with \( k_1 \) (equation (3.20a)), determines whether the corresponding first Huber-type weights should be modified.

- Define \( ds_{ij} \) as Mahalanobis distance for single dimension of \( j \) for the \( i^{th} \) observation and calculate it as:

\[
ds_{ij}^2 = \left( \frac{x_{ij} - m_j}{\sigma_j} \right)^2 \quad i = 1, \ldots, n \quad j = 1, \ldots, d. \tag{3.23}
\]

where \( m_j \) and \( \sigma_j^2 \) are the current estimation of median\(^1\) and variance

---

\(^1\)For a random sample of size \( N \), arranged in increasing order, the sample median is defined as \( X_{N/2} \) if \( N \) is odd and \( \frac{X_{N/2} + X_{N/2+1}}{2} \) if \( N \) is even.
3.3 Outline of the new approach

calculated over $n$ observations for each measurement. $ds_{ij}^2$ constitutes the element-wise Huber-type weights and are then applied to the elements of the vector of observation. This, in an asymptotical sense, is equal to filling in some missing values in an observation vector with their weighted mean. This technique is known as mean amputation [67]. It ignores the covariance structure (i.e. the off-diagonal elements) and if used alone, leads to a biased estimate.

- The calculated distance, $ds_j$, is then compared with $k_2$ and the corresponding first Huber-type weights will be modified using equation (3.20b).

- Assuming $H_1$ as an $n \times d$ matrix of first Huber-type weights, $X_{n \times d}$ as the matrix of observation and $m$ and $w$ as the vector of means and the vector of ones of sizes $d$ and $n$ respectively, define the term $J$ as:

$$J = H_1 \times (X - wm).$$  \hspace{1cm} (3.24)

where $J_i$ is the $i^{th}$ row of $J$ and represents the weighted $j^{th}$ observation.

The new estimate of the mean vector and covariance matrix are then calculated as:

$$m^{new} = m + \text{sum}(J) / \text{sum}(H_1).$$ \hspace{1cm} (3.25)

$$V^{new} = V + \frac{1}{\beta} \sum_{i=1}^{n} J_i J_i^T.$$

where $\text{sum}(\cdot)$ returns a vector with the sum over each column, operator $\cdot \times$, is defined as $A \times B = \{a_{ij} \times b_{ij}\}$ and operator $\cdot /$, is defined as
$A./B = \{a_{ij}/b_{ij}\}$.

The iterative calculation continues until the difference between the largest elements of two most recently calculated covariance matrices is less than a pre-determined threshold (e.g. $1 \times 10^{-6}$).

### 3.4 Examples

To examine the efficiency of the proposed method, two examples are presented. Both uncorrelated and correlated systems have been chosen, leading to strictly diagonal and non-diagonal covariance matrices.

#### 3.4.1 Diagonal case

The first case, taken from Romagnoli and Stephanopoulos [30], is a system consisting of a chemical reactor with 4 streams, 2 entering and 2 leaving the process. All the stream flow rates are assumed to be measured and their true values are $M = \begin{bmatrix} 0.1739 & 5.0435 & 1.2175 & 4.00 \end{bmatrix}^T$. The corresponding incidence matrix $A$, and the covariance matrix, $V$, are also known and given as follows:

$$A = \begin{bmatrix} 0.1 & 0.6 & -0.2 & -0.7 \\ 0.8 & 0.1 & -0.2 & -0.1 \\ 0.1 & 0.3 & -0.6 & -0.2 \end{bmatrix}$$
3.4 Examples

\[
V = \begin{bmatrix}
0.0029 & 0 & 0 & 0 \\
0 & 0.0025 & 0 & 0 \\
0 & 0 & 0.0006 & 0 \\
0 & 0 & 0 & 0.04 \\
\end{bmatrix}
\]

Process measured values were generated about the above mean and covariance matrix, using the Monte Carlo method. The data set size used was 1000 points.

**Without outliers**

For the generated data set, the direct method, the Chen *et al.* method and the RDCE result in the following covariance matrices, respectively:

**Direct method:**

\[
V = \begin{bmatrix}
0.0026 & 0 & 0 & 0 \\
0 & 0.0027 & 0 & 0 \\
0 & 0 & 0.0006 & 0 \\
0 & 0 & 0 & 0.0393 \\
\end{bmatrix}
\]

**Chen *et al.* method:**

\[
V = \begin{bmatrix}
0.0028 & 0 & 0 & 0 \\
0 & 0.0025 & 0 & 0 \\
0 & 0 & 0.0005 & 0 \\
0 & 0 & 0 & 0.0411 \\
\end{bmatrix}
\]
3.4 Examples

The proposed RDCE method:

\[
V = \begin{bmatrix}
0.0025 & 0 & 0 & 0 \\
0 & 0.0023 & 0 & 0 \\
0 & 0 & 0.0006 & 0 \\
0 & 0 & 0 & 0.0368
\end{bmatrix}
\]

The similarity of results, confirms the accuracy of the iterative robust methods used in conjunction with the residuals, as in Chen et al., and in conjunction with covariance matrix, as in the RDCE.

With outliers

When several outliers are included in the same data set, the direct method is no longer able to predict any useful estimation of the covariance matrix.

The Chen et al. approach and the RDCE result in the following covariance matrices:

Chen et al. method:

\[
V = \begin{bmatrix}
0.0028 & 0 & 0 & 0 \\
0 & 0.0025 & 0 & 0 \\
0 & 0 & 0.0005 & 0 \\
0 & 0 & 0 & 0.0412
\end{bmatrix}
\]

The proposed method:

\[
V = \begin{bmatrix}
0.0025 & 0 & 0 & 0 \\
0 & 0.0023 & 0 & 0 \\
0 & 0 & 0.0006 & 0 \\
0 & 0 & 0 & 0.0368
\end{bmatrix}
\]
The above results show that both methods are capable of rejecting the introduced outliers and generating a very good estimate of the covariance matrix.

3.4.2 Nondiagonal case

This example is taken from Keller et al. [62]. The process consists of 7 nodes and 12 streams and is shown in Fig. 3.1. Chen et al. [63] changed the position of the correlated sensors to make certain that the target covariance conforms with the common assumption of the property of positive definiteness. Therefore, measurement errors for sensors two and five, and for sensors six and 11 are correlated. The process incidence matrix and “true” covariance matrix of the system are as follows:

\[
A = \begin{bmatrix}
1 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
\end{bmatrix}
\]
Again a data set of size 1000 was generated to simulate the plant measured variables.
Without outliers

Both the Chen et al. method and the RDCE produce very good estimates of the covariance matrix.

Chen et al. method:

\[
V = \begin{bmatrix}
29.13 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 29.81 & 0 & 0 & 4.91 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 19.56 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 19.13 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4.91 & 0 & 0 & 6.98 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 12.38 & 0 & 0 & 0 & 0 & 9.54 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 11.24 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 10.09 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 11.45 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 8.22 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 9.54 & 0 & 0 & 0 & 0 & 19.83 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 21.22 & 0 \\
\end{bmatrix}
\]

The RDCE:

\[
V = \begin{bmatrix}
29.48 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 27.68 & 0 & 0 & 5.53 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 19.60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 17.57 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 5.53 & 0 & 0 & 6.99 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 13.44 & 0 & 0 & 0 & 0 & 7.69 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 9.47 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 8.90 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 9.13 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7.60 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 7.69 & 0 & 0 & 0 & 0 & 18.21 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 20.35 & 0 \\
\end{bmatrix}
\]
3.4 Examples

With outliers

The covariance matrices in the nondiagonal case when some outliers are introduced to the data set are as follows:

Chen et al. method:

\[
V = \begin{bmatrix}
28.94 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 29.90 & 0 & 0 & 4.80 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 19.56 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 19.06 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 4.80 & 0 & 0 & 7.12 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 12.36 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 11.18 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 11.05 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 11.46 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8.21 & 0 \\
0 & 0 & 0 & 0 & 0 & 9.54 & 0 & 0 & 0 & 0 & 19.90 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 21.21
\end{bmatrix}
\]

The RDCE:

\[
V = \begin{bmatrix}
29.48 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 27.95 & 0 & 0 & 5.52 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 19.73 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 17.57 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 6.99 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 13.44 & 0 & 0 & 0 & 0 & 7.69 \\
0 & 0 & 0 & 0 & 0 & 0 & 9.48 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 8.90 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 9.13 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7.60 & 0 \\
0 & 0 & 0 & 0 & 0 & 7.69 & 0 & 0 & 0 & 0 & 18.21 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 20.35
\end{bmatrix}
\]

The obtained results show that both methods are robust and capable
3.4 Examples

of rejecting outliers and generating very good estimates of the covariance matrix. Their performance for data sets without outliers were equally satisfactory. A numerical comparison in terms of the norms of the covariance matrices for both diagonal and nondiagonal cases is presented in table 6.1.

<table>
<thead>
<tr>
<th>Norm</th>
<th>True Value</th>
<th>Robust Method</th>
<th>RDCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>diagonal, no outliers</td>
<td>0.0402</td>
<td>0.0413</td>
<td>0.0370</td>
</tr>
<tr>
<td>diagonal, outliers present</td>
<td>0.0402</td>
<td>0.0414</td>
<td>0.0370</td>
</tr>
<tr>
<td>nondiagonal, no outliers</td>
<td>66.1881</td>
<td>64.7171</td>
<td>61.5855</td>
</tr>
<tr>
<td>nondiagonal, outliers present</td>
<td>66.1881</td>
<td>61.9467</td>
<td>61.7485</td>
</tr>
</tbody>
</table>

It is worth noting that contrary to the Chen et al. method, the proposed RDCE method requires only the data obtained from plant measurements. Unlike the indirect method, the RDCE does not rely on the (linear) constraints. That, in turn, allows it to avoid the matrix manipulation and inversion of the derivations of the incidence matrix. This point becomes increasingly important when one faces derivations of the incidence matrix that are not full-ranked and cannot be inverted. For example, if measurement errors for sensors 2 and 8 are correlated

\[ V(2, 8) = V(8, 2) = 6. \]

the true covariance matrix for the example taken from Keller et al. [62] would be:
3.4 Examples

\[
\begin{bmatrix}
30 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 30 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 \\
0 & 0 & 20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 20 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 7.5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 15 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8.1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 20 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 20
\end{bmatrix}
\]

\( \nu = \)

Then none of the indirect methods or the Chen et al. method would be able to provide an estimate on the covariance matrix because of the rank deficiency that occurs in the calculation. The RDCE however, produces the following estimate:

\[
\begin{bmatrix}
26.26 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 27.93 & 0 & 0 & 0 & 0 & 0 & 0 & 5.77 & 0 \\
0 & 0 & 19.00 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 20.87 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 7.03 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 14.15 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 9.43 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 9.52 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 9.55 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 7.51 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 18.69 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 19.93
\end{bmatrix}
\]

Furthermore, the RDCE does not require the correlated measurement errors a priori. It does, however, employ a robust iterative method to reject
3.5 Summary

The covariance matrix of the measurement error is a necessary statistical tool in treatment of measured process data. It provides an insight into process measurements and produces helpful outputs that can be used in other process data analyses.

The Direct Method of covariance matrix estimation, when there are deviations from steady-state and in the presence of outliers, due to ramification of errors, fails to provide an acceptable estimate of the covariance matrix. Indirect methods can deal with deviations from steady-state, but are very sensitive to outliers. These methods are unable to provide a reasonable estimate of the covariance matrix when outliers are present in the measured process data. Chen et al. [63] eliminated this drawback by using a robust approach that incorporates the M-estimator.

Both indirect methods and the Chen et al. method, however, are based on linear constraints, require a knowledge of the correlated measurements and are vulnerable to singularity and rank deficiency in the estimation process. The proposed RDCE method combines the direct use of plant measured data with the robust M-estimator. In this way avoiding the potential numerical pitfalls, while extracting information directly from measured data. In the provided examples, it was shown that the RDCE produced very good estimates of the covariance matrix, without the a priori knowledge of process or
the computational load that is required by other methods.
Chapter 4

Probability Density Estimation Using Incomplete Data

4.1 Introduction

Many process monitoring and control tasks require statistical characterization of the process. One of the most useful statistics is the joint probability density function (pdf) of the process variables. The joint pdf can be utilized in different process monitoring and control tasks. The pdf can be used to describe the normal operating condition of a process, or the parametric region capable of producing product within acceptable quality specification and to evaluate the probability that current process condition will produce product within specification [56]. A low probability would signal the need for making adjustments to the process operating conditions such that the probability of producing products within acceptable range increases and moves toward its maximum. The joint pdf can also be used in gross error detection and sensor
validation. Assuming a plant operating under normal conditions, a low joint pdf of a set of measurements is an indication of one or more gross errors in the measurement set.

Estimation of the joint pdf is an important part of the data rectification problem when studied in a probabilistic framework. In this framework, the probability density of the process states, given the process measurements, is maximized to find the rectified states. By Bayes’ theorem, the conditional probability of states given the measurements may be written in terms of the conditional probability of the measurements given the states and the probability of the states.

Several pdf estimation techniques, both parametric and nonparametric exist. In the parametric techniques, a probability density model is assumed and the plant data is used to find the parameters of the model [12]. The parametric pdf estimation technique can be used when prior knowledge about the underlying distribution is available and the pdf of the plant data is known to be of a certain type. The main drawback of parametric techniques is that in most cases the prior knowledge of the underlying pdf does not exist and as a result, an assumed model fails to provide satisfying performance. While parametric approaches yield impressive results, nonparametric techniques [8] [68] free of a priori assumptions are more powerful and because of their generality show a broader range of applicability.

Generally speaking, any pdf estimation approach that does not assume a functional form of the density parametrized by a fixed set of parameters can be categorized as a nonparametric approach. A pdf with a Gaussian
distribution model is an example of a parametric technique. Histograms, splines, orthogonal series estimators and mixture densities are a few examples of nonparametric techniques [12].

Among the different nonparametric methods of pdf estimation, considerable interest has been shown in mixture density models [69]. The method of mixtures models the probability densities as a mixture of a given fixed probability density functions. A mixture density is composed of a number of components and each is defined by its own set of parameters. In this way, mixture models combine much of the flexibility of nonparametric methods with the analytic advantages of parametric methods.

The construction of the mixture density model or estimating the parameters of a mixture model from the plant data set is called “Learning”. The most commonly used method for learning the parameters of a mixture is the Expectation-Maximization (EM) algorithm [70]. EM is an iterative method, which starts with an initial estimate and proceeds to construct a sequence of estimates by first estimating the expectation of the log-likelihood (c.f. Subsection 4.4.1) of the current estimate, then constructing the new estimate by maximizing this expectation.

The plant data available for the learning process is typically incomplete. A real process often operates in an environment that provides only partial data. Let’s assume a data set \( X = \{(x_i, y_i)\}_{i=1,...,N} \) is an incomplete data set in the sense that some observations of \( y_i \) are missing. One way of estimating the mean vector, \( \mu = [\mu_x \mu_y]^T \), is to estimate its \( x \) component, \( \mu_x \), by averaging observed \( x_i \)'s and its \( y \) component, \( \mu_y \), by averaging observed \( y_i \)'s of the
data set $X$. This method of estimating $\mu$ may lead to an underestimated $\mu_v$ that would be far from its most likely value. That is simply a consequence of incorporating less than available information in estimating $\mu_v$. Another approach is to use the covariance matrix and fill in the missing values of $y_i$ by regressing $y_i$ as a function of $x_i$. This method results in a biased estimate of the covariance matrix since the filled in data points lie along the regression line. Both of the above methods, known as mean imputation and regression imputation, yield unsatisfactory results [67].

It is also assumed that the measured process data is obtained while the process is maintained under steady state condition. In the strict sense of the concept, however, the steady state condition almost never prevails [23]. Stanley and Mah [71] outlined the concept of quasi steady state (QSS) conditions. The QSS assumption, guarantees that the network of pdf’s for the mixture density is trained with the data obtained from one operating point and that there has been no departure from steady state operating condition at the given operating point.

In this chapter, unsupervised learning of mixture models from incomplete data is examined. Supervised learning is learning a function when the functionality $(x_i, f ((x_i)))$ is provided by a teacher. Unsupervised learning is learning without a teacher. No target information and functionality is provided and only a set of input data, $x_i$, called the training set is provided and the objective is to associate useful characteristics to this data set.

The effect of missing data on the performance of the EM algorithm when it is used to fill in the missing values and learn the mixture parameters, is
delineated. The weights of the mixture components are viewed as mixing parameters and are estimated through the same learning process as the rest of the parameters of the density mixture.

In Section 4.2 the mixture density models, their methods of estimation and mixture density estimation using the principle of maximum likelihood are presented. Within Section 4.3, an iterative procedure for numerically approximating maximum likelihood estimates of the parameters of a mixture density formulated by Dempster, Laird and Rubin [8] and termed the EM algorithm is briefly examined. Using the framework presented in Section 4.3, learning from incomplete data is presented in Section 4.4. In section 4.5, three examples of pdf estimation are examined and discussed.

4.2 Mixture density estimation

4.2.1 Mixture density model

A parametric family of finite mixture densities, i.e. a family of probability density functions is of the form

\[ p(x|\Phi) = \sum_{j=1}^{M} \alpha_j p_j(x|\phi_j) \quad x = [x_1, ..., x_n]^T \in \mathbb{R}^n. \quad (4.1) \]

where \( x \) is a \( n \)-dimensional vector of measurements, each \( \alpha_j \) is a non-negative weight and \( \sum_{j=1}^{M} \alpha_j = 1 \), and each \( p_j \) is itself a density function parameterized
by \( \phi_j \in \Omega_j \subseteq R^n \). Also denote

\[
\Phi = \{ (\alpha_1, \ldots, \alpha_M), (\phi_1, \ldots, \phi_M) \}.
\]

(4.2)

and

\[
\Omega = \left\{ (\alpha_1, \ldots, \alpha_M, \phi_1, \ldots, \phi_M) : \sum_{j=1}^{M} \alpha_j = 1 \text{ and } \alpha_j \geq 0, \phi_j \in \Omega_j \text{ for } j = 1, \ldots, M \right\}.
\]

(4.3)

Finite mixture densities arise naturally as densities associated with a statistical population which is a mixture of \( M \) component populations with associated component densities \( \{p_j\}_{j=1, \ldots, M} \) and mixing proportions \( \{\alpha_j\}_{j=1, \ldots, M} \).

A sample observation on a mixture is labeled if its original component population is known with certainty, otherwise it is termed an unlabeled observation. If a large sample of measurements which are labeled is available, then it would be an easy and straightforward task to obtain a satisfactory estimate of \( \Phi \). Unfortunately, this is not usually the case and \( \Phi \) may have to be estimated by using a sample in which some or all observations are unlabeled. This problem is referred to as the Mixture Density Estimation problem.

### 4.2.2 Mixture density estimation using the method of maximum likelihood

The method of moments is perhaps the oldest method of estimation dating back to Karl Pearson in the late 1800s [12]. It has the virtue of being quite
simple to use and almost always yields estimates. The method of moments consists generally of equating a set of sample moments to their expected values and thereby obtaining a system of nonlinear equations for the parameters in the mixture density. The method of moments in many cases does not lead to the best estimate of the mixture parameters and consequently the resulting estimates still need to be improved. In addition, the computational load increases drastically with an increase in the size of the system and the number of parameters. Given these facts [12], the method of moments becomes less and less applicable as the dimensionality and number of components of a mixture increases.

Another method of mixture parameter estimation is the method of maximum-likelihood. The method of maximum likelihood is by far the most popular technique for estimating parameters [8]. If \( x_1, ..., x_n \) are an independent and identically distributed sample from a population with probability density function \( p(x|\Phi) \), the likelihood function is defined by:

\[
L(\Phi|x) = L(\phi_1, ..., \phi_M|x_1, ..., x_n) = \prod_{i=1}^{N} p(x_j|\Phi). \tag{4.4}
\]

Assume \( \hat{\Phi} \) is the parameter set for which \( L(\hat{\Phi}|x) \) attains its maximum as a function of \( \Phi \) when \( x \) is being kept constant. Then, the maximum likelihood estimate (MLE) of the parameter set \( \Phi \) based on a sample \( X \) is \( \hat{\Phi}(X) \). In other words, an MLE associated with a sample is a choice of parameters which maximizes the probability density function of the sample, called the likelihood function.
Therefore, intuitively, the MLE is a reasonable choice for an estimator. The MLE leads to the parameter set for which the observed sample is most likely.

It is assumed that a parametric family of mixture densities of equation 4.1 is specified and that a particular $\Phi^* = \{ (\alpha^*_1, ..., \alpha^*_M), (\phi^*_1, ..., \phi^*_M) \} \in \Omega$ is the true parameter value to be estimated.

It is both natural and convenient to regard $p(x|\Phi)$ in equation 4.4 as modeling a statistical population which is a mixture of $M$ component populations with associated component densities $\{p_j\}_{j=1,...,M}$ and mixing proportions $\{\alpha_j\}_{j=1,...,M}$. The likelihood function of a sample of observations is the probability density function of the random sample evaluated at the current observations. When maximum likelihood estimates are of interest, it is usually convenient to deal with the logarithm of the likelihood function, called the log-likelihood function, rather than the likelihood function itself.

Assuming $\{x_i\}_{i=1,...,N}$ is an independent sample of $N$ unlabeled observations of the mixture, i.e., a set of $N$ observations of independent, identically distributed random variables with density $p(x|\Phi^*)$, the log-likelihood function is defined as:

$$L(\Phi) = \sum_{i=1}^{N} \log p(x_i|\Phi). \quad (4.5)$$

There are two inherent drawbacks associated with the general problem of finding the maximum of a function and hence of maximum likelihood estimation.

The first problem is that of finding the global maximum and verifying that a global maximum has actually been found. What is meant by a max-
4.2 Mixture density estimation

imum likelihood estimate of \( \Phi^* \), is that any choice of \( \Phi \) in \( \Omega \) at which the log-likelihood of the sample, denoted by \( L(\Phi) \), attains its largest local maximum in \( \Omega \). In defining a maximum likelihood estimator in this way, two practical difficulties associated with maximum likelihood estimation for mixture densities are taken into account.

The first difficulty is that one cannot always take \( \Omega \) to be a set in which the log-likelihood function is bounded above, and so there are not always points in \( \Omega \) at which \( L \) attains a global maximum over \( \Omega \) [8]. The second difficulty is multiplicity of choices of \( \Omega \) at which the log-likelihood function attains its largest local maxima. If \( p_i \) and \( p_j \) are of the same parametric family for some components \( i \) and \( j \), then the value of \( L(\Phi) \) will not change if the component pairs \((\alpha_i, \phi_i) \) and \((\alpha_j, \phi_j) \) are interchanged.

The main concern is that the log-likelihood function often has local maxima which are not the largest local maxima. From a computational point of view, there is little that can be done about this problem but one must decide whether to accept a given local maxima as the largest or to go on to search for others.

The second problem is that of numerical sensitivity. That is, how sensitive\(^1 \) is the estimate to small changes in data. Sometimes a slightly different sample produces an entirely different set of maximum likelihood estimates.

The traditional, general approach of determining a maximum likelihood estimate is first to arrange a system of likelihood equations satisfied by the

\(^1\)Strictly speaking, this is a mathematical rather than statistical problem associated with any maximization procedure including maximum likelihood estimation which is obtained using a maximization scheme.
maximum likelihood estimate and then try to obtain a maximum likelihood estimate by solving the likelihood equations. The likelihood equations are found by considering the partial derivatives of the log-likelihood function with respect to the components of $\Phi$.

If $\hat{\Phi} = \{(\hat{\alpha}_1, \ldots, \hat{\alpha}_M), (\hat{\phi}_1, \ldots, \hat{\phi}_M)\}$ is a maximum likelihood estimate, then the likelihood equations are determined by:

$$\nabla_{\phi_j} L = 0, \quad j = 1, \ldots, M. \quad (4.6)$$

To obtain likelihood equations determined by the proportions, which are constrained to be non-negative and sum to one, the approach of Redner and Walker [10] is followed. They showed that:

$$\hat{\alpha}_j = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{\alpha}_j p_j(x_i|\hat{\phi}_j)}{p(x_i|\hat{\Phi})}, \quad j = 1, \ldots, M. \quad (4.7)$$

### 4.3 The Expectation-Maximization (EM) algorithm

Expectation-Maximization (EM) is an iterative technique originally proposed by Dempster et al. [8] for numerically approximating the parameters of a probability density function when it is provided in the form of a mixture density. EM is typically used to calculate the minimum likelihood estimates given an incomplete set of data. EM procedures have been proposed in literature for many years [8] [72] [73] [74] but advances in computation have
now made them considerably easier to implement.

Let \( x \) be a set of observable variables that constructs the known part of a more general incomplete observation \( y \). Let \( g(x|\Phi) \) specify a family of functions from which the complete samples are drawn and \( f(y|\Phi) \) specify a family of functions from which the incomplete samples are drawn.

The EM algorithm attempts to find a value for \( \Phi \) which maximizes \( g(x|\Phi) \) given an observation \( x \), by using the relationship that exists between \( f \) and \( g \) [8]. The conditional density \( k(y|x, \Phi) \) is defined as:

\[
f(y|\Phi) = k(y|x, \Phi) g(x|\Phi).
\]  

(4.8)

For \( \Phi \) and \( \Phi' \in \Omega \) and by taking expectation from equation 4.8 with respect to the probability density \( f(y|\Phi) \) we obtain:

\[
L(\Phi) = Q(\Phi|\Phi') - H(\Phi|\Phi').
\]  

(4.9)

where

\[
Q(\Phi|\Phi') = E \left( \log f(y|\Phi) | x, \Phi' \right).
\]  

(4.10a)

\[
H(\Phi|\Phi') = E \left( \log k(y|x, \Phi) | x, \Phi' \right).
\]  

(4.10b)

and \( L(\Phi) \) is the maximum-likelihood objective function.

The EM algorithm procedure is given as follows:

1. Set \( i = 0 \) and initialize \( \Phi \) with an arbitrary \( \Phi^0 \).
2. Calculate $Q (\Phi^{i+1} | \Phi^i)$.

3. Find $\Phi^{i+1}$ such that $Q (\Phi | \Phi^i)$ is maximized.

4. If $\Phi^i = \Phi^{i+1}$ increment $i$ and return to step 2.

In step 2, the Expectation step, the current estimate of model parameters is used to estimate the missing data. The resulting complete data set is then used in step 3, the Maximization step, to produce a new set of parameters that maximizes the log-likelihood objective function. The iterative process continues until convergence occurs.

The importance of the EM algorithm lies in the fact that at any EM iteration $L (\Phi^{i+1}) \geq L (\Phi^i)$. In fact the way in which $\Phi^{i+1}$ is determined guarantees that $Q (\Phi^{i+1} | \Phi^i) \geq Q (\Phi^{i+1} | \Phi^i)$ and $H (\Phi^{i+1} | \Phi^i) \leq H (\Phi^{i+1} | \Phi^i)$ [8] and that in turn secures a monotonic increase in log-likelihood during any iteration sequence of EM.

### 4.3.1 EM algorithm for mixture densities

The EM algorithm for a mixture density estimation problem is derived by first casting the problem as one involving incomplete data then obtaining the algorithm from its general formulation. The EM steps, as depicted in figure 4.1, are defined as:

- **E-Step:** Assuming the model (cluster parameters) to be correct, find the most likely distribution of the data with respect to the model.

- **M-Step:** Assuming the distribution to be correct, maximize the likelihood of the model (cluster parameters) with respect to the data.
4.3 The Expectation-Maximization (EM) algorithm

Figure 4.1: EM iteration for a mixture density

For a mixture of normal distributions, the multivariate probability density for any cluster $j$ is defined as:

$$
p_j(x_i | \phi_j) = \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)\right). \quad (4.11)$$

where $\phi_j$ is the set of parameters for the normal distribution representing cluster $j$ and is defined as:

$$\phi_j = (\mu_j, \Sigma_j). \quad (4.12)$$

In the Expectation step for this case, assuming the current maximum-likelihood estimate of the model $\Phi^c$ as $\Phi^c = ((\alpha^c_1, ..., \alpha^c_M), (\phi^c_1, ..., \phi^c_M))$, the
4.3 The Expectation-Maximization (EM) algorithm

posterior probability for each data point \( x_i \) is calculated as \( \alpha_j^n p_j(x_i|\phi_j^n) / p(x_i|\Phi^c) \) and the network is re-structured. In the maximization step, given the current structure of the density mixture, the maximum likelihood estimate of the density parameters are found and the model is re-parameterized. The new set of parameters are estimated as:

\[
\alpha_j^n = \frac{1}{N} \sum_{i=1}^{N} \frac{\alpha_j^n p_j(x_i|\phi_j^n)}{p(x_i|\Phi^c)} \quad j = 1, ..., M. \tag{4.13}
\]

\[
\mu_j^n = \frac{\sum_{i=1}^{N} x_i \alpha_j^n p_j(x_i|\phi_j^n)}{\sum_{i=1}^{N} \alpha_j^n p_j(x_i|\phi_j^n) / p(x_i|\Phi^c)}. \tag{4.14}
\]

\[
\Sigma_j^n = \frac{\sum_{i=1}^{N} (x_i - \mu_j^n) (x_i - \mu_j^n)^T \alpha_j^n p_j(x_i|\phi_j^n)}{\sum_{i=1}^{N} \alpha_j^n p_j(x_i|\phi_j^n) / p(x_i|\Phi^c)}. \tag{4.15}
\]

\( \Sigma_j^n \) would be a positive-definite symmetric matrix, with possibility of one, if \( N \), the number of data points (or observations) remains larger than \( n \), the dimensionality of the system.

Details of the EM algorithm and its application to mixture densities are provided in Appendix A.

4.3.2 Properties of EM

1. **Heuristic Appeal.** The most appealing general property of EM is that it produces sequences of iterations in which the log-likelihood function increases monotonically.
2. **Reliable Global Convergence.** The monotonic increase in log-likelihood is the basis of the general convergence theories for EM. The desirable theoretical properties automatically enjoyed by the EM algorithm suggest the good global convergence behavior of the algorithm which has been observed in practice by many investigators [10]. Theorems which essentially confirm this suggested behavior have been published by Vardi [75] Boyles [76] and Wu [77]. In general, one can expect an EM iteration sequence to converge to a local maximum of the log-likelihood function. EM does not necessarily converge to the global maximum for each and any application but this same problem exists for most other optimization schemes. There are no general purpose algorithm which can reliably and efficiently find a global maximum.

3. **Ease of Programming and Economy of Storage.** The EM algorithm is easy to program and needs little storage. Any constraints on $\Phi$ are likely to be satisfied, or at least nearly satisfied for large samples [10]. The sequence of $\Sigma^k_j$ generated by EM algorithm for a mixture of multivariate normal densities is symmetric and, with probability one, positive definite when $N > n$. The mixing proportions are always non-negative and sum to one.

4. **Low Computational Cost.** In the case of a mixture of multivariate normal densities, the EM algorithm requires $O(mn^2N)$ arithmetic operations per iteration, compared to $[O_1(M^2n^4N) + O_2(M^3n^6)]$ for Newton's method and $[O_1(mn^2N) + O_2(M^2n^4)]$ for a quasi-Newton method
Using a secant update [10].

In spite of these attractive features, the EM algorithm may encounter problems in practice. The source of most serious practical problems associated with the algorithm is the speed of convergence of sequences of iterations generated by EM, which can be slow [10]. In the case of a mixture of normal densities, the convergence of EM iteration sequence is linear, as opposed to the quadratic convergence rate of Newton's method and superlinear convergence rate of quasi-Newton method using a secant update.

4.4 Learning from incomplete data using the EM approach

In performing a parametric estimation for mixture densities and obtaining an explicit probabilistic model, one generally has to deal with imperfection of information. One form of imperfection is incompleteness in sensor data which can arise extrinsically from the data generation process and intrinsically from failures of the system's sensors. For example, a vision system may encounter many partially occluded examples of an object and yet has to recover a model for the occluded object. Similarly, an adaptive controller may be required to learn a mapping from sensor readings to actions even if the sensors produce unreliable readings or sometimes fail to produce readings.
4.4 Learning from incomplete data using the EM approach

4.4.1 Density estimation for complete data using the EM approach: mixture of Gaussians

The principles of maximum likelihood parameter estimation for a mixture as described by Duda and Hart [68] is followed. Modifying equation 4.1 to include the probability of components of a mixture and assuming a sample \( X = \{x_1, \ldots, x_N\} \), the probability that \( x_i \) is generated independently from a mixture density is given by:

\[
p(x_i) = \sum_{j=1}^{M} p(x_i | \omega_j, \phi_j) p(\omega_j). \tag{4.16}
\]

where each component of the mixture is denoted by \( \omega_j \) and parameterized by \( \phi_j \).

The log-likelihood function can then be evaluated as:

\[
L(\Phi | X) = \sum_{i=1}^{N} \log \sum_{j=1}^{M} p(x_i | \omega_j, \phi_j) p(\omega_j). \tag{4.17}
\]

Maximization of this function is not an easy task because it involves the logarithm of a summation. Another problem is that it is not clear which component of the mixture generated a given data point [78] and therefore which parameters to adjust to fit the observation. The iterative scheme of EM is used to solve this problem. If a set of new binary variables, \( Z \), are defined such that \( z_{ij} = 0 \) if the observation \( x_i \) does not belong to component \( j \) and \( z_{ij} = 1 \) otherwise, then the maximization problem of equation 4.17 would decouple into a set of simple maximization problems [78]. The
new maximization problem then can be formulated as a complete-data log-likelihood.

\[ L(\Phi | X, Z) = \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \log \left( p(x_i | z_i, \phi_j) p(z_i | \phi_j) \right). \] (4.18)

Since the values of \( z \) are unknown, equation 4.18 cannot be directly used for the maximization of the log-likelihood function. As discussed in section 4.3, the EM algorithm can be used for the maximization of a log-likelihood function such that:

1. E-step: \( Q(\Phi|\Phi^k) = E \left( L(\Phi | X, Z) \mid X, \Phi^k \right) \).

2. M-step: \( \Phi^{k+1} = \max_{\Phi} Q \left( \Phi | \Phi^k \right) \).

where \( k \) denotes the number of iterations.

**E-Step:** For a mixture of Gaussians, the expectation step is simplified to the calculation of \( E \left( z_{ij} | x_i, \Phi^k \right) \), that is, the probability of observation \( x_i \) being generated by (Gaussian) component \( j \) given the estimated component parameter at step \( k \):

\[ h_{ij} = \frac{|\Sigma_j^k|^{-1/2} \exp \left( -\frac{1}{2} (x_i - \mu_j^k)^T \Sigma_j^{-1} (x_i - \mu_j^k) \right)}{\sum_{w=1}^{M} |\Sigma_w^k|^{-1/2} \exp \left( -\frac{1}{2} (x_i - \mu_w^k)^T \Sigma_w^{-1} (x_i - \mu_w^k) \right)}. \] (4.19)

**M-Step:** The maximization step, as described in Appendix A, equations 4.14 and 4.15, evaluates and updates the parameters of the Gaussian
components such that the log-likelihood function is maximized:

\[
\mu_{j}^{k+1} = \frac{\sum_{i=1}^{N} h_{ij} x_{i}}{\sum_{i=1}^{N} h_{ij}}. \quad (4.20a)
\]

\[
\Sigma_{j}^{k+1} = \frac{\sum_{i=1}^{N} h_{ij} (x_{i} - \mu_{j}^{k+1}) (x_{i} - \mu_{j}^{k+1})^{T}}{\sum_{i=1}^{N} h_{ij}}. \quad (4.20b)
\]

### 4.4.2 Density estimation from incomplete data: mixture of Gaussians

In the previous section, EM is used for learning parameters of a mixture density. The same principle is used in this section to learn the parameters for a mixture density from data sets with missing values \[72\] \[8\]. This application of EM has been studied for non-mixture density estimation problems. Ghahramani and Jordan \[78\] extended this application of EM to mixture densities.

Assume a sample of observations \(X = \{x_1, ..., x_N\}\) which is divided into an observed component \(X^o\) and a missing component \(X^m\). Each vector of observations, \(x_i\), can be divided into observed and missing parts so that \(x_i = \left[ \begin{array}{c} x_i^o \\ x_i^m \end{array} \right]^{T}\).

The EM algorithm in this case is stated as follows:

1. E-step: \(Q(\Phi|\Phi^k) = E \left( L(\Phi|X^o, Z) | X^o, X^m, \Phi^k \right)\).

2. M-step: \(\Phi^{k+1} = \max_{\Phi} Q(\Phi|\Phi^k)\).
4.4 Learning from incomplete data using the EM approach

In addition to $Z$, the EM algorithm also include $X^m$ as a set of unknown variables and the E-step estimates both of these missing values. The E-step uses the current estimate of parameters of mixture components to produce an estimate for the missing data. The log-likelihood function can be stated as:

$$L(\Phi|X^m, X^o, Z) = \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \log p(x_i|z_i, \phi_j) + \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \log p(z_i|\phi_j).$$

(4.21)

Since we are only concerned with finding those set of parameters that maximizes the log-likelihood function given the probability density function of measured values, $p(x_i|z_i, \phi_j)$, the second term is ignored.

Rewriting equation 4.21 in more detail to express subvectors and submatrices of unknown and missing variables results in:

$$L(\Phi|X^m, X^o, Z) = \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \left\{ -\frac{n}{2} \log 2\pi - \frac{1}{2} \log \Sigma_j + \frac{1}{2} (x_i^o - \mu_j^o)^T \Sigma_j^{-1, oo} (x_i^o - \mu_j^o) + \frac{1}{2} (x_i^m - \mu_j^m)^T \Sigma_j^{-1, mm} (x_i^m - \mu_j^m) \right\}.$$  

(4.22)

where $x_i^o$, $x_i^m$, $\mu_j^o$ and $\mu_j^m$ are the observed and missing part of the observation and mean vector respectively such that $x_i = \begin{bmatrix} x_i^o & x_i^m \end{bmatrix}^T$ and $\mu_j = \begin{bmatrix} \mu_j^o & \mu_j^m \end{bmatrix}^T$ and $\Sigma_j^{-1, oo}, \Sigma_j^{-1, om}$ and $\Sigma_j^{-1, mm}$ are the corresponding parts of $\Sigma_j^{-1}$ according to indices of observed and missing values.

**E-Step:** In performing the E-step, the sufficient statistics for obtaining mixture parameters would include three unknowns: $z_{ij}$, $z_{ij}x_i^m$ and $z_{ij}x_i^m^Tx_i^m^T$. As a result, in order to estimate mixture density parameters $E(z_{ij}|x_i^o, \phi_j)$,
4.4 Learning from incomplete data using the EM approach

\[ E(\mathbf{z}_i^m, \mathbf{x}_i^m | \phi^k_j) \text{ and } E(\mathbf{z}_i^m \mathbf{x}_i^m \mathbf{m}^T | \mathbf{x}_i^o, \phi^k_j) \text{ must be estimated.} \]

If the estimate of missing data required in the E-Step is calculated by using the current estimate of mixture density and then the resulting completed data set is used in the M-Step to update the mixture parameters, the expectation of the missing data always lies along a line that biases the estimate of the covariance.

The approach arising from the application of the EM algorithm specifies that one should use the current density estimate to calculate the expectation of the incomplete terms in the log-likelihood function. For the mixture of Gaussians, the incomplete terms involve interactions between the \(z_{ij}\) variables and the first and the second moment of \(x_i^m\). This point leads to the conclusion that calculating \(z_{ij}\) and \(x_i^m\) and substituting them into M-step is not sufficient to guarantee an increase in likelihood of the parameters. Ghahramani and Jordan [78] proposed the following relations for calculation of the expectation of the missing variables:

\[ E(\mathbf{z}_i | \phi^k_j) = h_{ij} \quad (4.23) \]

\(h_{ij}\) is defined in equation 4.19 and is calculated using measured variables only,

\[ E(\mathbf{z}_i, \mathbf{x}_i^m | \phi^k_j) = h_{ij} E(\mathbf{x}_i^m | z_{ij} = 1, x_i^o, \phi^k_j) \quad (4.24) \]

\[ = h_{ij} (\mu_j^m + \Sigma_j^m \Sigma_j^o, -1 (x_i^o - \mu_j^o)). \]

\[ E(\mathbf{z}_i, \mathbf{x}_i^m \mathbf{m}^T | x_i^o, \phi^k_j) = h_{ij} \left( \Sigma_j^{mm} - \Sigma_j^o, -1 \Sigma_j^m \mathbf{m}^T + x_{ij}^m x_{ij}^m \mathbf{m}^T \right). \quad (4.25) \]
where \( x_{ij}^m \) is defined as:

\[
x_{ij}^m = E (x_i^m | z_{ij} = 1, x_i^o, \phi_j^k) = \mu_j^m + \Sigma_j^m \Sigma_j^{co, -1} (x_i^o - \mu_j^o).
\]  

(4.26)

and is the regression of \( x_i^m \) on \( x_i^o \).

**M-Step**: These equations are then substituted into equations 4.20a and 4.20b to perform the M-step and produce the new estimate of the mixture density parameters.

### 4.5 Examples

#### 4.5.1 Steady state data

The estimation of probability density investigated here is based on measured process data obtained from a process operating at steady state condition. This makes the process simpler and more straightforward. But what is steady state and what are indications of operating at steady state?

In a strict sense, steady state conditions almost never prevail in practice [23]. Many continuous processes operate with steady state conditions prevailing for long periods of time with relatively rapid transitions between steady states. This type of process which is essentially at steady state except for small drifts and occasional sharp transitions between steady states is referred to as quasi steady state (QSS) [71]. The simplicity of the steady state comes at a price. As is obvious from the definition of QSS, if departures from steady state are not detected, the deviations will be reflected in the
measurement errors.

There are some statistical tests developed to detect changes in steady state [79] [80]. There are several other control charts, in each of these a test statistic is being plotted to check the quality of the data. The Shewhart control chart (for the mean of the measured variable) is one of the initial tools of statistical quality control [81].

In the investigation of the pdf estimation technique here, the QSS condition is assumed. Since multi-dimensional data with a non-diagonal covariance matrix is used, the previously mentioned methods of checking steady state become improper or too computationally expensive to use. Instead, a robust method of outlier rejection is implemented. In this method, the Mahalanobis distance of an observation (from the median) is computed. If the distance is larger than a predetermined threshold, a very small weight is assigned to that observation. The assigned weight would be very close to one otherwise [82]. As a result, the effect of outliers is minimized.

4.5.2 Examples of probability density function estimation

Two examples are presented that use the EM technique for estimating the pdf using incomplete data. The first case is a one-dimensional bimodal distribution with a 0.3 probability of being drawn from $N(0, 1)$ and a 0.7 probability being drawn from $N(4, 2)$, that is, $p(x) = 0.3N(0, 1) + 0.7N(4, 2)$. The second case is a two-dimensional bimodal distribution. It has a 0.3 probability of being drawn from $N(0, I)$ and a 0.7 probability being drawn from
$N(\mu, \Sigma)$ where $\mu = \begin{bmatrix} 4 & 3 \end{bmatrix}^T$ and $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$.

Although all of the examples have diagonal covariance, they are equivalent to more general problems that do not have strictly diagonal covariance. Translation, rotation and linear scaling can be used to move one of the distributions to the origin and make it spherical. The distributions can then be rotated about the origin until the other distribution has its axes aligned with the principal axes, having a diagonal covariance matrix [83].

For each case a training set of 100 points is generated. A test set of 1000 points is also generated and used for calculating the log-likelihood of the estimated pdf's.

In estimating a pdf based on labeled data, it is assumed that the population of origin for all the data points is known. This assumption is relaxed in the unlabeled estimation of pdf. The K-Means clustering technique is used to cluster the data points.

It is always useful, but not necessarily practical, if a classifier can be trained by a set of labeled data and then the larger set of data be treated as unsupervised. However, to study the effect of missing data in a more general manner, missing data and the estimation of pdf parameters are combined into one process. To investigate the effect of missing data, elements of the training data set were randomly replaced by NaN (not a number), representing a missing value. Different percentages of missing values were examined.

Figure 4.2, for the one-dimensional case, shows plots of pdf's for labeled and unlabeled data with different percentages of missing values. Component
parameters for each pdf are presented in table 6.1. As can be seen in figure 4.2 and table 6.1, replacing missing values with their maximum likelihood estimates would result in an estimate of the pdf that closely follows the original pdf.

The results obtained for the two dimensional case for labeled data, unlabeled data with no missing values and unlabeled data with 10% missing values and unlabeled data with 30% missing values are shown in figures 4.3, 4.4, 4.5 and 4.6 and table 6.2.

Increases in the percentage of missing data would result in a less efficient clustering of data points and a poor initial estimates of parameters of the pdf. Further increases in the percentage of missing data leads to non-identifiability. The study of unsupervised learning is greatly simplified if the assumption of identifiability for the pdf mixture holds. For a mixture of Gaussians, it can be shown that mixture of normal densities are usually identifiable. A mixture of Gaussians cannot be uniquely identified when components weights are equal but component parameters can be interchanged without affecting the estimated pdf [68].

As the percentage of missing data is increased, the estimated parameters of the mixture density and its components depart from the original structure of the mixture pdf and deviate drastically from the original values. This result can be tracked in the log-likelihood versus percentage of missing data curves.

Using the test data set, the log-likelihood for each pdf was examined, estimated at different percentages of missing data. Typical curves for the
one-dimensional and two-dimensional cases are depicted in figures 4.7 and 4.8, respectively. In both cases, the extremum point of the curve shows the proximity of percentage of missing data up to the point the estimated values for the pdf are in agreement with the original pdf. Any further increase in missing data reduces the break down value of the training data set and as a result, the estimated parameters become very sensitive to the quality of data and values of their initial estimates.

Figure 4.9 displays a typical variation of the value of log-likelihood (divided by its constants) as the EM iteration proceeds. As expected, it shows that the log-likelihood monotonically increases with the EM iteration. The value of the log-likelihood at the start cannot be calculated since the missing data are not yet estimated and hence it starts from the first iteration. At a constant percentage of missing data, the increase of the log-likelihood at each iteration shows that a more accurate estimate of missing values and mixture pdf parameters are obtained after each EM iteration.

Table 4.1: Estimated Mixture Component Parameters. One-dimensional Case

<table>
<thead>
<tr>
<th>pdf Parameters</th>
<th>1st Component</th>
<th>2nd Components</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Variance</td>
<td>Weight</td>
<td>Mean</td>
<td>Variance</td>
</tr>
<tr>
<td>Labeled</td>
<td>0</td>
<td>1</td>
<td>0.30</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>0.275</td>
<td>1.283</td>
<td>0.37</td>
<td>4.289</td>
<td>1.255</td>
</tr>
<tr>
<td>Unlabeled 10% missing</td>
<td>0.191</td>
<td>1.104</td>
<td>0.33</td>
<td>4.295</td>
<td>1.374</td>
</tr>
<tr>
<td>Unlabeled 20% missing</td>
<td>0.128</td>
<td>1.096</td>
<td>0.30</td>
<td>4.314</td>
<td>1.526</td>
</tr>
<tr>
<td>Unlabeled 30% missing</td>
<td>0.247</td>
<td>1.428</td>
<td>0.56</td>
<td>4.318</td>
<td>1.608</td>
</tr>
</tbody>
</table>
Table 4.2: Estimated Mixture Density Component Parameters. Two-dimensional Case

<table>
<thead>
<tr>
<th>pdf Parameters</th>
<th>1&lt;sup&gt;st&lt;/sup&gt; Component</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt; Component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Covariance</td>
</tr>
<tr>
<td>Labeled</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>0.219</td>
<td>2.374</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>1.437</td>
<td>3.190</td>
</tr>
<tr>
<td>10% missing</td>
<td>0.356</td>
<td>1.511</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>1.750</td>
<td>3.934</td>
</tr>
<tr>
<td>20% missing</td>
<td>0.808</td>
<td>2.717</td>
</tr>
<tr>
<td>Unlabeled</td>
<td>2.089</td>
<td>4.593</td>
</tr>
<tr>
<td>30% missing</td>
<td>0.936</td>
<td>2.807</td>
</tr>
</tbody>
</table>

4.5.3 The data rectification case

Data rectification is defined as the process of adjusting the measured process variables such that the error resulting from the normal noise is minimized and the effects of outliers existing in the data set are eliminated. The presence of errors in the measured process variables leads to process control difficulties that in turn could cause the process to operate in an infeasible or even haz-
Figure 4.2: Density function for labeled and unlabeled data with various amounts of missing values. One-dimensional case.

ardous region. As such, data rectification is important not only for efficient plant operation but also for plant safety.

This example demonstrates the use of probability density estimation as a part of the probabilistic data rectification process. The probability density
functions used in the data rectification scheme are estimated through the use of incomplete data and the effect of missing measurements on the rectified values of the measured process variables is investigated.

The heat exchanger network of Tjoa and Biegler [50] with modifications is selected as a chemical process example to show the effect of incomplete
data in estimating probability density functions. The resulting probability density mixture is incorporated into a data rectification scheme to estimate the most likely values of the measured process variables.

The heat exchanger network as depicted in figure 4.10 consists of four heat exchangers. Fluid A is heated by fluid B in two of the heat exchangers (HX1
Figure 4.5: PDF for unlabeled data with 10% missing data. Two-dimensional case.

and HX2) and by fluid C and fluid D in two other heat exchangers (HX3 and HX4). The flow of stream D is manipulated to maintain the temperature of the stream A8 that leaves the heat exchanger network. The input flows and temperatures were drawn from Gaussian distributions as shown in table 6.3. Stream A2 is split into streams A3 and A6 such that $F_{A3} = 0.42F_{A2}$.
Figure 4.6: PDF for unlabeled data with 30% missing data. Two-
dimensional case.

The process was simulated in HYSYS™ using hypothetical components
and the values of the measured variables were recorded given the input condi-
tions. The HYSYS™ [84] generated steady state data were collected and the
missing values were ensured in the data set by randomly replacing measured
values with NaN, given the required level of missing values.
Figure 4.7: Change in log-likelihood as the % missing data increases. One-dimensional case.

A data set of 100 points was created and after introducing different percentages of missing values, the data set was used to train the mixture of probability density estimations.

The temperature of stream A5 is used to show the effect of varying the percentage of missing values on the rectified values of process variables. Figure 4.11 depicts the variation of $T_{A5}$ calculated from the measured process data with no missing data, rectified process values and process rectified values using training data with 30% missing values. Deviation from the rectified values with no missing values in the data set is denoted by white and black
Figure 4.8: Change in log-likelihood as the % missing data increases. Two-dimensional case.

bars, depending on the positiveness or negativeness of the deviation respectively.

When the measured process variables contain only random noise, which is the case in the present example, the rectified values provide a better estimate of the process state by minimizing the error. Figure 4.11 shows that the process state represented by the rectified values (for the 30% missing case) contains more variation compared to the rectified data obtained from the data set with no missing values and even the measured data.

To elaborate on this finding, the standard deviation of the rectified data
Figure 4.9: Monotonic increase in the log-likelihood function during the EM iteration.

set has been tabulated against the percentage of the missing data in table 6.4. Note that for the lower percentages of missing data the standard deviation remains almost the same. But it does increase, and in fact surpasses that of the observed variables, when the percent of missing data passes a process dependent threshold. A similar trend can be seen when the mean square error (MSE) for different data sets is tabulated against the percentage of missing values. The jump in MSE demonstrates that the rectified values calculated
based on using 30% missing data are not reliable estimates for the measured process data.

**Table 4.3: Probability distribution properties for process variables**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Distribution</th>
<th>Noise Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{A1}$</td>
<td>N(465,5)</td>
<td>0.75</td>
</tr>
<tr>
<td>$F_{A1}$</td>
<td>N(800,5)</td>
<td>4</td>
</tr>
<tr>
<td>$T_{B1}$</td>
<td>N(620,5)</td>
<td>0.75</td>
</tr>
<tr>
<td>$F_{B1}$</td>
<td>N(200,5)</td>
<td>2</td>
</tr>
<tr>
<td>$T_{C1}$</td>
<td>N(670,5)</td>
<td>0.75</td>
</tr>
<tr>
<td>$F_{C1}$</td>
<td>N(240,5)</td>
<td>3</td>
</tr>
<tr>
<td>$T_{D1}$</td>
<td>N(688,5)</td>
<td>0.75</td>
</tr>
<tr>
<td>$F_{D1}$</td>
<td>N(514,5)</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table 4.4: Properties of the rectified variables**

<table>
<thead>
<tr>
<th></th>
<th>Std. Dev.</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Data</td>
<td>0.25</td>
<td>0.8237</td>
</tr>
<tr>
<td>10% missing</td>
<td>0.1789</td>
<td>0.4248</td>
</tr>
<tr>
<td>20% missing</td>
<td>0.1828</td>
<td>0.4405</td>
</tr>
<tr>
<td>30% missing</td>
<td>0.2707</td>
<td>1.0697</td>
</tr>
</tbody>
</table>

4.6 **Summary**

Mixture models provide a reliable framework for handling missing data. By incorporating the EM algorithm into the learning process, the maximum likelihood estimate of the mixture parameters is ensured.

An unsupervised learning process in which similarities between observations are sought, proved useful when observations including the imperfect data are not categorized by a supervisor. Incorporation of observations with missing values in this framework provides a means of preserving the "valuable" available data. As an added benefit, the most likely estimate of the
Figure 4.10: The heat exchanger network for the data rectification example, Tjoa and Biegler [50]

missing values are also obtained.

The initial estimates are made based on the observed set of data and using the EM iterations. The missing values were estimated such that the maximum likelihood estimates for mean, covariance and weight of each component of the mixture density and hence the pdf estimate were found.

The limiting effect of missing data on the learning process was also studied. It is shown that the EM estimator is capable of providing a reliable estimate of mixture parameters in the presence of missing data and the resulting pdf displays similar behavior to that of the pdf estimated in the
absence of missing data. This similarity in behavior does not hold when the presence of missing data passes beyond a certain limit. Beyond this limit the EM estimator fails to provide an acceptable estimate of the pdf.

The use of the EM algorithm does ensure that for any given percentage of missing data, the resulting probability density mixture is the most likely one. Given the maximum likelihood properties of the EM algorithm, it is also shown that after reaching a process dependent threshold of missing data, the resulting probability density mixture is not reliable.
Chapter 5

Adaptive Mixtures

5.1 Introduction

A large number of applications require the ability to recognize patterns within data, where the character of the patterns may change with time. Example applications include remote sensing, autonomous control and automatic target recognition in a changing environment. These applications have a common requirement and that is the need to recognize new entities as they enter the environment. A pattern recognition system in this type of environment must be able to change its representation of the classes dynamically in order to conform to changes in the classes themselves, as well as recognize and develop a representation for a new class in the environment. The adaptive mixtures approach uses density estimation to develop decision functions for supervised and unsupervised learning through the use of a great deal of a priori information about the structure of the data [11].

Parametric approaches are often used to estimate the underlying model
for the data. While these approaches produce impressive results, nonparametric approaches, free of a priori assumptions can be considered more powerful due to their increased generality and therefore wider applicability [85]. Developing a system for performing unsupervised learning nonparametrically is a daunting task. In fact, there are many instances in which no system can be assured of proper performance. For example, two classes with identical distributions cannot be identified as such based on purely unsupervised learning. Nevertheless, a nonparametric density estimation approach to unsupervised learning can lead to a general and powerful pattern recognition tool [11].

The adaptive mixtures approach is considered a nonparametric method. While there is some blurring of distinction between parametric, semiparametric and nonparametric approaches, an estimation approach which intermittently changes the list of parameters to be estimated based on the incoming observations and has no a priori upper bound on this parameter list, may be called a nonparametric approach.

In addition to the nonparametric assumption, one must also consider the problem of recursive estimation. That is, it is assumed that due to high data rates or large time constraints, estimates must be developed in such a way that they do not require the storage or processing of all observations to date. This limits the ability to develop optimal estimates, but often this is the only feasible approach for a given application.

By virtue of addressing the type of applications that can be termed nonparametric and recursive, the problem of density estimation has become more
difficult and challenging. The nonparametric assumption implies that only simple assumptions can be made about the data. The need for nonparametric techniques stems from a range of applications in which the experimenter is unwilling to assume a parametric family for the true underlying probability density function [86].

The recursive assumption eliminates the possibility of using iterative techniques. In recursive estimation of a probability density function, the estimate based on the \( n+1 \)th observation \( \{x_1, x_2, ..., x_{n+1}\} \) is a function of the \( n+1 \)th observation, \( y_{n+1} \), and the estimate obtained based on the \( n \) previous observations \( \{y_1, y_2, ..., y_n\} \). The procedure eliminates the need to store all the incoming observations, permitting high data rates.

### 5.2 Kernel estimation

The method of kernel (or Parzen) estimates was introduced by Rosenblatt [87] and Parzen [88], studied by many authors and summarized in detail by Silverman [69], and Scott [89]. For the univariate case, the kernel estimate is given by:

\[
\hat{f}_n(y; x_1, ..., x_n) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - x_i}{h}\right).
\]  

(5.1)

where \( K(.) \) is the kernel function. The asymptotic properties of kernel estimates, in particular its strong convergence under very weak conditions as indicated by Devroye and Gyorfi [90], are very appealing. However, there are computational disadvantages to this method. To evaluate \( \hat{f}_n \), all \( n \) observations of \( K(.) \) are necessary. For many problems with very large data sets,
such a constraint is unacceptable as the computational load quickly becomes infeasible to handle.

### 5.3 Finite mixture models

Assume that the true but unknown density is of the form

$$
\hat{f}_o(x) = \sum_{i=1}^{N} \pi_i \phi(x; \mu_i, \Sigma_i).
$$

(5.2)

where $N < \infty$ is known, the non-negative mixing coefficients sum to one, and

$$
\phi_i(x) = \phi(x; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^n/2 |\Sigma_i|} \exp \left( -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right).
$$

(5.3)

is the normal probability density with mean $\mu_i$ and standard deviation $\Sigma_i$.

The goal is to estimate the parameter vector $\theta$, which consists of $3N - 1$ components:

$$
\theta = \{ \pi_1, \mu_1, \Sigma_1, ..., \pi_{N-1}, \mu_{N-1}, \Sigma_{N-1}, \mu_N, \Sigma_N \}.
$$

(5.4)

Finite mixture models have been discussed at length by Everitt and Hand [91], Titterington et al. [70], and McLachlin and Basford [92].

A standard technique for estimating the parameter vector $\theta$ based on observations $Y_n = \{ y_1, ..., y_n \}$ is to maximize the log-likelihood $L(Y_n)$ over the family of $N$ mixtures $F$. Let $\hat{f}(x) = \hat{f}(x; \hat{\theta})$ be the estimate for $f_o(x)$, when
\( \hat{\theta} \) is the estimated parameter vector. The iterative expectation-maximization (EM) algorithm given by Dempster et al. [8], and Redner and Walker [10], is a method for maximum likelihood estimation of these parameters.

To recursively implement this estimation technique, following Titterington [93], let \( S(y, \theta) \) denote the vector of scores. That is, for each component \( \theta_i \) of the parameter vector \( \theta \), let

\[
  s_i(y, \theta_i) = \frac{\partial}{\partial \theta_i} \log(\hat{f}(y; \hat{\theta})).
\]

and consider recursive update formula as follows:

\[
  \hat{\theta}_{k+1} = \hat{\theta}_k + K_n S(y_{k+1}; \hat{\theta}_k)
\]

with \( K_n \), defined as a sequence converging to zero. This equation can be interpreted as a gradient ascent on the log-likelihood surface.

With the proper choice of the sequence \( K_n \), this stochastic approximation procedure can be made consistent. An example of this approximation formula is the following set of recursive updates for normal components from Titterington [93], and Titterington et al., [70]:

\[
\begin{align*}
\rho_{k+1}^i &= \pi_k^i \frac{\phi^i_k(y_{k+1})}{\hat{f}_k(y_{k+1})}. \tag{5.7a} \\
\pi_{k+1}^i &= \pi_k^i + \beta_k^i (\rho_{k+1}^i - \pi_k^i). \tag{5.7b} \\
\mu_{k+1}^i &= \mu_k^i + \frac{\beta_k^i \rho_{k+1}^i}{\pi_k^i} (y_{k+1} - \mu_k^i). \tag{5.7c} \\
\Sigma_{k+1}^i &= \Sigma_k^i + \frac{\beta_k^i \rho_{k+1}^i}{\pi_k^i} \left[ (y_{k+1} - \mu_k^i) (y_{k+1} - \mu_k^i)^T - \Sigma_k^i \right]. \tag{5.7d}
\end{align*}
\]
\[ \beta_k^i = 1/k. \quad (5.7e) \]

Equations (5.7a) to (5.7e) are called update rules \( \mathcal{U}_k(y_{k+1}; \theta_k) \) [11]. The superscript \( i \) used in the above equations indicates the \( i^{th} \) term in the mixture estimate. The objective of the update rules is to distribute the effect of the new observations to all the terms in proportion to their respective likelihoods. The mean, variance and mixing coefficients are then updated by these proportions. In the case of a single term with \( \rho = 1 \) and \( \pi = 1 \), the update rules become recursive versions of sample mean and sample variance calculations. Even if the true probability density function is not known to be a mixture of Gaussians, one might still wish to use the foregoing formulation to find an approximation to the density by a mixture. The kernel estimator is an extreme example of such an approximation, with \( N \), the number of kernels, equal to \( n \) and no maximum likelihood updating involved. Thus one could choose \( N \) large enough, start the estimate with some initial \( \hat{\theta}_0 \) and then recursively update the estimate using \( \mathcal{U}(\cdot) \) [94].

To obtain consistency, very good assumptions must be made about the underlying density and for the initial state of the estimator. In particular, the underlying density must be a mixture of the same type as the estimator [94].

### 5.4 The approach of adaptive mixtures

Given a probability density function \( \hat{f}_o \) belonging to some class \( \mathcal{F} \), the adaptive mixtures procedure produces a sequence of estimates \( \{ \hat{f}_k \} \) that is consistent. The kernel estimator is consistent for a large class of functions but the
computational complexity of the method is often prohibitive. The finite mixture approach to density estimation, on the other hand, has very appealing properties but can be consistent only when the number of terms in the model has been correctly chosen. The adaptive mixtures technique demonstrates the appealing properties of both approaches: consistency for a large class of functions combined with the low computational complexity associated with finite mixture models [94].

If the number of components of a mixture is fixed, then the mixture can be parameterized with a fixed-length vector of parameters, \( \theta \). The estimate can be written as:

\[
f(y; \theta) = \sum_{i=1}^{N} \pi_i K(y; \Gamma_i). \tag{5.8}
\]

where \( N \) is the number of terms and \( K(y; \Gamma_i) \) is some fixed density parameterized by \( \Gamma \), then

\[
\theta = \{ \pi_1, \Gamma_1, \ldots, \pi_{N-1}, \Gamma_{N-1}, \Gamma_N \}. \tag{5.9}
\]

\( K(.) \) is taken to be a normal distribution, in which case \( \Gamma_i \) becomes \( \{ \mu_i, \Sigma_i \} \).

The basic stochastic approach is to update recursively the estimate \( \hat{\theta} \) of the true parameters \( \theta_0 \) based on the latest estimate \( \hat{\theta}_k \) and the newest observation \( y_{k+1} \):

\[
\hat{\theta}_{k+1} = \hat{\theta}_k + U_k(y_{k+1}; \hat{\theta}_k). \tag{5.10}
\]

where \( U_k \) is the update function.

The specific form of interest of this recursive update formula, as suggested
by Titterington [93], and Novelson and Hansminkii [95] is

\[ \hat{\theta}_{k+1} = \hat{\theta}_k + (nI(\hat{\theta}_k))^{-1} \frac{\partial}{\partial \theta} \log(f(y_{k+1}; \hat{\theta}_k)). \] (5.11)

where \( I(\hat{\theta}_k) \) is the Fisher Information matrix and the derivative represents the vector of partial derivative with respect to the components of \( \theta \) [86].

It is well understood [94] that maximum likelihood estimation cannot be performed in an infinite dimensional space if unconstrained maximization is attempted. The recursive maximum likelihood technique converges to the desired resultant estimator when properly constrained. The relationship of this estimator to the EM algorithm is noted in Titterington [93].

In order to extend finite mixtures to nonparametric estimation with a variable number of components, the recursive update formula must be extended to allow the addition of new components. The stochastic approximation procedure of equation 5.12:

\[ \hat{\theta}_{k+1} = \hat{\theta}_k + [1 - p_k(y_{k+1}; \hat{\theta}_k)] U_k(y_{k+1}; \hat{\theta}_k) + p_k(y_{k+1}; \hat{\theta}_k) C_k(y_{k+1}; \hat{\theta}_k). \] (5.12)

is used to update recursively the density function. \( p_k(.) \) represents a create decision and takes on values 0 and 1. This penalty function serves to constrain the addition of new terms.

\( U_k(.) \) updates the current parameters while \( C_k(.) \) adds a new component to the model similar to the kernel estimation approach.

Addition of new components due to \( C_k(.) \) adds new parameters to \( \theta \) and that in turn, changes the character of the likelihood surface. The fact that
5.4 The approach of adaptive mixtures

\( p_k(.) \) may depend on \( y_{k+1} \) implies that this change can be data driven. The creation rule \( C_k(.) \) is chosen so that the proportion and variance of new component decreases with the number of components [94].

If \( p_k(y_{k+1}; \hat{\theta}_k) = 1 \) for \( k < N \) and \( p_k(y_{k+1}; \hat{\theta}_k) = 0 \) for \( k \geq N \), the algorithm will fit \( N \) terms to the data. Alternatively, \( K \) components could be chosen using a priori knowledge and then set \( p_k(y_{k+1}; \hat{\theta}_k) = 0 \) for all \( k \).

On the other hand, if \( p_k(y_{k+1}; \hat{\theta}_k) = 1 \), the algorithm always creates a new component, centered at the new data point, and the estimate then becomes:

\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} k \left( \frac{y_i - x}{h_i} \right). \tag{5.13}
\]

The process of fitting a mixture density can be made consistent as the number of components goes to infinity. Although this estimator is recursive, it has no practical advantage over a kernel estimator unless the estimator is desired at only a finite number of predetermined points. Therefore, the principle of adaptive mixtures is to reduce the number of components, so that the computational requirements necessary to compute the estimate at any time is lessened and the estimate is more compactly represented. Fewer points in the kernel estimator could be used and the extra points discarded. But this is unacceptable because all the points have to be used to improve the estimator.

Observations for which \( p_k(y_{k+1}; \hat{\theta}_k) = 1 \) correspond to a jump in the likelihood surface which can be useful for guiding the estimator away from local maxima which may lead to erroneous solutions.
5.4 The approach of adaptive mixtures

5.4.1 Update rule

The update function \( U(.) \) guides a traversal of the estimate of the likelihood surface provided by the observations \( \{y_i\}_{i=1}^n \) and based on the likelihood equations [11]. This recursive maximum likelihood technique converges to the desired estimator when properly constrained.

5.4.2 Create rule

Assuming that the algorithm has decided to add a term \( (p(.) = 1) \), a create threshold \( C(.) \), can be derived from the fact that the kernel estimator based on \( n + 1 \) observations is closely related to the kernel estimator based on \( n \) observations. The differences are a new kernel or term, centered at the newest observation \( y_{n+1} \), updated proportionality constants for each term, from \( n^{-1} \) to \( (n+1)^{-1} \) and possibly different variances. This analogy is captured by the create rules defined as:

\[
\begin{align*}
\mu_{n+1}^{N+1} &= y_{n+1}. \quad (5.14a) \\
\Sigma_{n+1}^{N+1} &= \Sigma_o^{N+1}. \quad (5.14b) \\
\pi_i^{n+1} &= \pi_i^n (1 - \beta_{n+1}) \quad i = 1, ..., N \quad (5.14c) \\
\pi_{n+1}^{N+1} &= \beta_{n+1}. \quad (5.14d) \\
N &= N + 1 \quad (5.14e)
\end{align*}
\]

Thus the new term is centered at the newest observation and given a small mixing coefficient and an initial standard deviation \( \Sigma_o \), which may be user
5.4 The approach of adaptive mixtures

defined or derived from the terms in the neighborhood of the observation. All of the other mixing coefficients must be updated so that they sum to unity.

\( C_n(.) \) adds new parameters to \( \hat{\theta} \), changing the character and dimensionality of the likelihood surface [11]. The fact that \( p_n(.) \) depends on \( y_{n+1} \) implies that this change is data driven. The create rule \( C_n(.) \) is chosen so that the proportion and variance of the new term decreases with the number of terms.

5.4.3 Decision rule

Observations for which \( p_n(y_{n+1}; \hat{\theta}_n) = 1 \) in equation (5.12) correspond to jumps in the likelihood surface. This can be useful for guiding the estimator toward a good solution and away from local maxima corresponding to poor solutions. The consistency of adaptive mixtures hinges on the fact that these jumps in fact propel the recursive EM algorithm into a high quality estimate.

The decision to add a term, \( p(.) \), a kind of clustering criterion, can be made in a number of ways. The simplest is to check the Mahalanobis distance from the observation to each of the terms. If the minimum of these distances exceeds a threshold, called the create threshold, \( T_c \), then the point is in some sense too far away from the existing terms and a new term should be created [11]. The square of Mahalanobis distance is defined as the distance between a point \( y \) and a term with mean \( \mu^i \) and covariance matrix \( \Sigma^i \):

\[
M^i(y) = (y - \mu^i)^T\Sigma^{-1}_i(y - \mu^i).
\]  
(5.15)
Therefore if the create threshold is $T_c$, then a new term is created at the point $y_{n+1}$ if and only if

$$M(y_{n+1}) = \min_i (M^i(y_{n+1})) > T_c.$$  \hfill (5.16)

$T_c = 1$ implies creation of a new term for any observation that is at least one standard deviation away from the mean of each term. Similarly, $T_c = 4$ implies creating a new term for any observation that is at least two standard deviations away from the mean of each term. The former yields a faster rate of increase in the number of terms in the model than the latter.

Considering $\Lambda(y) = \exp(-\frac{1}{2}M(y))$, $p(.)$ may be defined as $p(y_{n+1}) = 1$ if and only if $\Lambda(y_{n+1}) < T_c$ and $p(y_{n+1}) = 0$ otherwise. In this case a threshold of $T_c \approx \exp(-0.5) = 0.605$ translates into a threshold of one standard deviation and $T_c \approx \exp(-2) = 0.1353$ translates into a threshold of two standard deviations.

\section{5.5 The adaptive mixtures algorithm}

\subsection*{5.5.1 Assumptions}

The assumptions made for training a network of adaptive mixtures are as follows:

- The kernels are assumed to be Gaussians.

- The number of existing classes is known and is assumed to be $C$. 

5.5 The adaptive mixtures algorithm

- The number of unknown classes, $U$, is assumed large enough so that it can accommodate the creation of new classes as they may appear.

- The model for each class $j$, is defined to be of the form:

$$ f^i(y) = \sum_{j=1}^{N_i} \pi^i_j \phi(y; \mu^i_j, \Sigma^i_j). \tag{5.17} $$

where $N_i$ is the number of terms for the $j^{th}$ class and $\phi$ is a Gaussian distribution. The overall estimate is then would be of the form:

$$ \tilde{f}(y) = \sum_{i=1}^{C+U} \sum_{j=1}^{N_i} \pi^i_j \phi(y; \mu^i_j, \Sigma^i_j). \tag{5.18} $$

and

$$ \sum_{i=1}^{C+U} \sum_{j=1}^{N_i} \pi^i_j = 1. \tag{5.19} $$

- The input to the system is in the format of pairs of observation and class and can be represented as $\{y_n, C_n\}$. If the observed point $y_n$ belongs to a known class, the value of $C_n$ for that observation is known.

5.5.2 Initialization

To start the training process, a number of parameters need to be initialized:

- The total number of supervised classes, $C$.

- The total number of unsupervised classes, $U$.

- The creation threshold, $T_C$. 
5.5 The adaptive mixtures algorithm

- The inclusion threshold, $T_I$.

- The discrimination threshold, $T_D$.

- The number of components for each class is set to zero.

5.5.3 The main loop

The whole training process consists of three main parts, namely Classification, Learning and Normalization. The process goes through all of the data points and the components are updated iteratively as observation points are processed. It can be summarized as [11]:

while the pairs of $\{y_n, C_n\}$ are available:

Step 1: Classify observation $y_n$

Step 2: If the class is known:

Supervised Learn.

Otherwise:

Unsupervised Learn.

Step 3: Normalize.

5.5.4 Classification

- Determine the probability of the observation belonging to each class as in equation 5.17.

- If $\max f^i(y_k) < T_D$ then the class is unknown. That is, there is no class above the discrimination threshold.
5.5 The adaptive mixtures algorithm

- For the observation $y_k$, calculate

$$\frac{f^i(y_k)}{f(y_k)} = \frac{f^i(y_k)}{\sum_{i=1}^{C+U} f^i(y_k)} \quad (5.20)$$

and return the vector of probabilities.

5.5.5 Supervised learn

Depending on the Mahalanobis distance of the observation and the value of the create threshold, either supervised update or supervised create steps may be required. The supervised learn step can be summarized as [11]:

If $\min \ M^j_i(y_k) \leq T_C$:

Supervised Update.

Otherwise:

Supervised Create.

Supervised update

For each component $i$ of class $c_k$ ($i = 1, ..., N_{ck}$), update mean, covariance and weight according to equations (5.7a) to (5.7f) proportionately.

Supervised create

Allocate a new component to the class $c_k$ with its properties as defined in equations (5.14aa) to (5.14e). The covariance matrix of the newly created component is the weighted average of all components of the class $c_k$. If the class $c_k$ does not have any components, then a predetermined value of $\Sigma^o$ is chosen as the initial covariance matrix.
5.5.6 Unsupervised learn

For the observation \( y_k \), if the smallest Mahalanobis distance of all the clusters of all the classes is smaller than or equal to the create threshold, an unsupervised update will take place. An unsupervised update is the same as a supervised update but is done over all of the classes.

If, on the other hand, the minimum was greater than the create threshold, the supervised create takes place. Based on the Mahalanobis distance of the observation point from existing components and the inclusion threshold, this step may result in a new component for the existing classes or lead to the creation of a new class. This step may be summarized as [11]:

If \( \min_i \min_j M_{ij}(y_k) \leq T_C \):

Unsupervised Update.

Otherwise:

Unsupervised Create.

Unsupervised update

Update mean, covariance and weight of all the components of all classes proportionately. This step is the same as supervised update except that it takes place over all classes.

Unsupervised create

In this step the smallest Mahalanobis distance of all components of all classes is compared to the inclusion threshold and if greater, a new class is created.
and its first component is initialized with the following properties:

\begin{align}
\mu_{N_j+1}^i &= y_k. \quad (5.21a) \\
\Sigma_{N_j+1}^i &= \Sigma_0^j = \text{weighted average over all classes} \quad (5.21b) \\
N_j &= 1. \quad (5.21c) \\
p_{N_j+1}^i &= 1. \quad (5.21d)
\end{align}

where \( p_{N_j+1}^j \) is the unnormalized weight of the \( i^{th} \) component of the \( j^{th} \) class [11].

If the smallest Mahalanobis distance of all components of all classes is smaller than or equal to the inclusion threshold, then for each class whose \( \min M_j^i(y_k) \leq T_C \), allocate a new component. This is the same as supervised create except that the covariance matrix of the newly created component is the average of the covariance matrices of other components of that class.

### 5.5.7 Normalization

After processing each observation point, the weights of all of the components of each class should be normalized so that the summation of weights of components of each class is equal to unity.

Two different methods to normalize weights are considered. One is to make the overall estimate a density function, and as a result, the sum of all weights over all classes would be unity. Another way is to make each class estimate a density, such that the sum over each class is unity, assuming equal weights for different classes.
The flowchart for the entire training process of a network of adaptive mixtures is shown in figure 5.1.

Figure 5.1: Training procedure for estimating probability density estimation using Adaptive Mixtures
5.6 Summary

The nonparametric technique for estimation of probability density functions provides an useful tool for density estimation where the form of the underlying density function is not known. There are many methods for nonparametric density estimation. Some, like kernel estimators, impose a large computational load and require large amounts of storage. Some, like histograms, are non-smooth and many are not recursive in nature.

The method of adaptive mixtures is a recursive nonparametric method for probability density estimation. This technique combines the desired properties of both kernel estimates and finite mixtures in that it converges for a large class of probability density functions while maintaining the computational simplicity associated with finite mixtures [94].

Adaptive mixtures algorithm require less data storage than kernel estimators and is not limited to the distributional assumption that finite mixture densities are. It automatically determines the number of terms present in the density mixture and is capable of estimating the smoothing parameters for separate terms [11].

The method of adaptive mixtures recursively fits a mixture of Gaussian distributions to the data. This mixture is not constrained to a fixed size and can be expanded through adding new terms where the incoming data requires it. This enables the resulting density mixture to track the entities in a changing environment and adapt to new operating conditions.
Chapter 6

Probabilistic Data Rectification

6.1 Introduction

Data rectification is the process of correcting measured process data. A maximum likelihood data rectification technique that poses the rectification problem in a probabilistic framework and maximizes the probability of the estimated plant states given the plant measurements is developed. Errors that appear in the plant measured variables are categorized as random errors and gross errors. The maximum likelihood data rectification finds the most likely values of the plant states such that the existing gross errors are eliminated and random errors are minimized.

Instead of using different classes of operation for a sensor that operates in normal mode and gross error mode, a bimodal probability density function is used to model a sensor error. The bimodal probability density function consists of two elements, one representing the normal mode of sensor operation and the other characterizing the sensor when it is operating in gross error
mode. Each mode is modeled by a Gaussian distribution and is weighted according to its probability of occurrence.

The method of adaptive mixtures is used to estimate the probability density function of the plant states, \( p(x) \). The method of adaptive mixtures uses plant historical data to establish a mixture of weighted Gaussians. The plant history provides the physical relationship as well as the statistical correlation among measured process variables. The method of adaptive mixtures utilizes this knowledge to build the mixture of Gaussians. The method of adaptive mixtures also uses the new plant observations to update its components. The method of adaptive mixtures was explained in Chapter 5. This method enables the process to recognize the new entities in the measured process data when the process operating conditions change. The use of new measured plant data to update the probability density function representing the probabilistic model has previously always been ignored and the model has been previously built solely based on historical steady state plant data.

The updating technique is also extended to the probability density functions representing sensor errors. As the new process measurements arrive, the probability density functions of the sensors errors are updated. In order to preclude outliers from the updating process, a robust recursive technique for updating the probability density parameters was developed.

The resulting maximum likelihood data rectification problem can be reduced to a surrogate quadratic objective function [96] if a set of indicators is known. The method of Expectation-Maximization is then used to iteratively find the indicators and the rectified values of the measured process variables.
6.2 Sensor model

The application of the maximum likelihood data rectification technique to a simulated heat exchanger network and its performance in rectifying the measured process data is described in examples and it is shown that in the absence of a process model and in a changing environment, this technique is capable of providing reliable estimates of the process actual states.

6.2 Sensor model

The measurement error in the maximum likelihood rectification problem is represented by \( p(y|x) \). For a given measurement \( y_i \) that comes from sensor \( i \), the error associated with that measurement can be modeled as:

\[
    y_i = x_i + \epsilon_i.
\]  

(6.1)

where \( x_i \) is the true value for the measured variable and \( \epsilon_i \) represents the associated error.

The error in a sensor is normally assumed as an additive random value, but multiplicative errors, where the size of error is dependent on the size of the signal are also possible [97]. Equation 6.1 can be used to model a multiplicative error by taking \( \epsilon_i \) as a function of \( x_i \). When \( \epsilon_i \) is independent of \( x_i \), it can be stated that:

\[
    p(\epsilon_i) = p(y_i|x_i) = \sum_j p(y_i|x_i,m_j)p(m_j).
\]  

(6.2)

where \( j \) counts the number of sensor modes of operation, and \( p(m_j) \) is the
probability of a sensor operating in the $j^{th}$ mode of operation. When a sensor is operating in its normal mode of operation, the error term, $\epsilon_i$, represents the sensor noise. Noise is usually modeled as a zero-mean Gaussian distribution. The magnitude of sensor noise is provided by the manufacturer and contains the effects of conformity, hysteresis, dead band and repeatability errors [97].

The central limit theorem states that the cumulative effects of errors will approach a Gaussian distribution as the number of sources of error increases. Since the error in a sensor is a result of a number of different errors, the Gaussian distribution would appropriately represent the sensor model. If the sensor error on the other hand, had a skewed distribution the underlying assumptions in deriving the sensor error model are not true and so would be the algorithm based on that. Additional failed modes such as failure to a fixed value or failure to a random value, can also be considered. Failure to a fixed value can be modeled by a delta function and failure to a random value can be modeled by a uniform distribution, but the exact distribution of other failure models is difficult to characterize [1]. If over a period of time, smaller gross errors are more likely than large gross errors, the distribution of gross errors may be approximated by a Gaussian with a standard deviation greater than that of the normal noise [50]. It should be noted that this Gaussian is not the probability density function representing the outlier model. It is only used to reflect the effect of presence of the outliers so that a separate outlier detection scheme is not required. This assumption leads to modeling
6.3 Estimating the pdf of the plant states

$p(y_i|x_i)$ as a bivariate Gaussian as follows:

$$p(y_i|x_i) = p(\epsilon_i) = (1 - p_i) \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{(x_i - y_i)^2}{2\sigma_i^2}\right)$$

$$+ p_i \frac{1}{\sqrt{2\pi b_i \sigma_i}} \exp\left(-\frac{(x_i - y_i)^2}{2b_i^2 \sigma_i^2}\right).$$ (6.3)

where $p_i$ is the probability of a gross error in sensor $i$, $\sigma_i$ is the standard deviation of normal noise in sensor $i$ and $b_i$ is the ratio of the standard deviation of the gross error in sensor $i$ to the standard deviation of the normal noise in sensor $i$. The second term of $p(y_i|x_i)$, i.e. the gross error probability density can be modeled using other distributions but regardless of the choice of the distribution, assuming that sensor biases are independent of each other, the joint density of the bias is the product of the individual biases:

$$p(y|x) = p(\epsilon) = \prod_i p(\epsilon_i) = \prod_i p(y_i|x_i).$$ (6.4)

6.3 Estimating the pdf of the plant states

The joint probability density function (pdf) of the process states, $p(x)$, captures the physical relationships of the process such as mass and energy balances. It also describes the statistical characteristics among process variables and in plant operation, e.g. correlation among process variables and distribution properties of these.

To enhance the estimation of $p(x)$, the method of adaptive mixtures is used. This method broadens the estimation process resulting from the stan-
6.4 Robust recursive update of sensor error pdf parameters

standard offline training of \( p(x) \) and improves it with on-line updating of \( p(x) \).

The method of adaptive mixtures is a recursive nonparametric method for probability density estimation. It fits a mixture of Gaussian distributions to the sensor data. This mixture is not constrained to a fixed number of components and new components may be added when the incoming process measurements point to addition of components.

The method of adaptive mixtures, its comparison with other methods of pdf estimation, the structure and steps in constructing the mixture and its properties are explained in Chapter 5.

6.4 Robust recursive update of sensor error pdf parameters

Each sensor in the process is modeled by a Gaussian distribution. To reflect any changes in the process operating conditions, this network of Gaussian distributions should be updated. Iterative updating techniques require all the measured process data to be stored, are slow and impose large storage requirements. On the other hand, recursive updating eliminates the need for storing all the measured process data and generates faster updates.

If one chooses not to update the basic structure of the network of Gaussians for sensor errors, obtained from offline training, one loses the ability of tracking and reflecting changes in the process condition. The updating process in practice translates to updating the parameters of the Gaussian distributions, i.e. mean and variance. But what if the incoming measured
process data is erroneous? Updating the parameters when the data contains outliers leads to an unreliable parameter set and in turn to a faulty description of the process condition. To overcome these problems, a robust recursive updating scheme is desirable.

### 6.4.1 Non-robust update

Assuming that the mean and variance of a measured process variable, using \( k \) observations, are \( \mu_k \) and \( \sigma_k^2 \) and given \( k + 1 \)th observation, \( y_{k+1} \), the non-robust recursive formula for updating mean and variance are as follows:

\[
\mu_{k+1} = \mu_k + \frac{1}{k+1} (y_{k+1} - \mu_k). \quad (6.5)
\]

\[
\sigma_{k+1}^2 = \frac{k - 1}{k} \sigma_k^2 + \frac{1}{k+1} (y_{k+1} - \mu_k)^2. \quad (6.6)
\]

### 6.4.2 Robust mean update

The most likely values of mean can be estimated robustly as follows [64] [65]:

\[
\frac{1}{n} \sum_{i=1}^{n} h_i(d_i)(y_i - \mu) = 0. \quad (6.7)
\]

where \( d_i \) is the Mahalanobis distance defined as:

\[
d_i^2 = \frac{(x_i - \mu)^2}{\sigma_i^2}. \quad (6.8)
\]
and $h_1$ is the first Huber-type coefficient [9] defined as:

$$h_1(d_i) = \begin{cases} 
1, & d_i \leq k. \quad (a) \\
\frac{k}{d_i}, & d_i > k. \quad (b)
\end{cases}$$

(6.9)

where $k^2$ is the 90% point of $\chi^2_p$ distribution with $p$ degrees of freedom. Rearranging equation 6.7 and solving for the "up to the $k^{th}$ observation", $\mu_k$ is found as follows:

$$\mu_k = \frac{\sum_{i=1}^{k} h_1(d_i) y_i}{\sum_{i=1}^{k} h_1(d_i)}.$$  

(6.10)

By the same token, the mean for the "up to the $k + 1^{th}$ observation", $\mu_{k+1}$, is found as:

$$\mu_{k+1} = \frac{\sum_{i=1}^{k+1} h_1(d_i) y_i}{\sum_{i=1}^{k+1} h_1(d_i)}.$$  

(6.11)

By rearranging and manipulating equation 6.11, $\mu_{k+1}$ can be calculated recursively from $\mu_k$ and $y_{k+1}$ as follows:

$$\mu_{k+1} = \mu_k + \frac{h_1(d_{k+1})}{\sum_{i=1}^{k+1} h_1(d_i)} (y_{k+1} - \mu_k).$$

(6.12)

### 6.4.3 Robust variance update

The most likely estimate for variance is calculated as follows [64] [65]:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} h_2(d_i)(y_i - \mu)^2.$$  

(6.13)
where $h_2$ is the second Huber-type coefficient and is defined as:

$$h_2(d_i) = (h_1(d_i))^2 / \beta.$$  \hfill (6.14)

where $\beta$ is defined by equation 6.15:

$$\beta = G(k^2/2, 1.5) + 2k^2(1 - \Phi(k)).$$  \hfill (6.15)

and $\Phi(k)$ is the standard normal distribution function and $G(x, f)$ is the Gamma distribution function with $f$ degrees of freedom [66].

The variance for the "up to the $k^{th}$ and $k+1^{th}$ observation" is respectively calculated as follows:

$$\sigma_k^2 = \frac{1}{k-1} \sum_{i=1}^{k} h_2(d_i)(y_i - \mu_k)^2.$$  \hfill (6.16)

$$\sigma_{k+1}^2 = \frac{1}{k} \sum_{i=1}^{k+1} h_2(d_i)(y_i - \mu_{k+1})^2.$$  \hfill (6.17)

By rearranging equation 6.17 and incorporating equation 6.16, the recursive formula for the variance update is obtained as follows:

$$\sigma_{k+1}^2 = \frac{k-1}{k} \sigma_k^2 + \frac{1}{k}\beta \left( \sum_{i=1}^{k+1} h_1(d_i) \right)^2 \left[ \sum_{i=1}^{k} h_1^2(d_i) + ... \right] \left( \sum_{i=1}^{k} h_1(d_i) \right)^2 (y_{k+1} - \mu_k)^2.$$  \hfill (6.18)
6.4.4 Numerical examples

Without outliers

A measurement set of 100 points is drawn from a Gaussian distribution of \( N(800, 5) \). A noise of the distribution \( N(0, 4) \) is added to the measurement. The measurement set of 100 points is shown in figure 6.1 and it is assumed that no outlier is present in the data set. Figures 6.2 and 6.3 show the results for both the non-robust and robust recursive update for mean and variance, respectively. It can be seen that both of the robust recursive updates closely follow the non-robust updates. As such, in the absence of outliers, the estimated parameters by the robust recursive method are as reliable as those estimated by the non-robust recursive update and can be used for the probability distribution function that represents the sensor error.

With outliers

The same measurement set as in the previous section was used but in order to introduce the outliers to the data set, some measurements were randomly replaced by erroneous values. The data set is shown in figure 6.4. Figure 6.5 shows the mean update for both the non-robust and robust recursive updating methods. As can be seen, each outlier introduces a large deviation in the value of mean when it is updated by the non-robust method. On the other hand, when the robust method is used, the effect of outliers is minimized. The variance of the measurement set when updated by the non-robust recursive method, as shown in figure 6.6, dramatically changes while the robust recursive method of updating keeps the variance in its normal
Figure 6.1: Process measurements, no outlier present.

range.

6.5 The maximum likelihood data rectification approach

This approach poses the data rectification problem in a probabilistic framework and exploits the laws of probability and maximum likelihood to find the most likely values of the measured process variables. The maximum likelihood data rectification approach maximizes the probability distribution
of the process states, \( x \), given the process measurements, \( y \), to rectify the measured process variables. The problem is posed as follows:

\[
\max_x p(x|y).
\]  

(6.19)

Using Bayes’ theorem, \( p(x|y) \) can be written in terms of the probability distribution of the measurements, \( p(y|x) \), and the probability distribution of the states, \( p(x) \):

\[
\max_x \frac{p(y|x)p(x)}{p(y)} \equiv \max_x p(y|x)p(x).
\]  

(6.20)
Figure 6.3: Recursive variance update, no outlier present

\( p(y) \) is independent of the process states and as such, does not affect the maximization problem. \( p(y|x) \) represents the probability density of the measurements given the process states. Assuming that only random noise is present, \( p(y|x) \) is the distribution of the measurement error. As explained in Section 6.2, the presence of gross errors in the measured process data cannot be ruled out. Given the set of mutually exclusive modes of sensor operation \( m_j \), a measurement model for sensor \( i \) is expressed by a weighted sum over
its modes [1] (normal, failure to a fixed value, failure to a random value):

\[ p(y_i|x_i) = \sum_j (y_i|x_i, m_j)p(m_j). \]  
(6.21)

Assuming that sensor errors are independent, the product of these probabilities over all sensors results in the term \( p(y|x) \):

\[ p(y|x) = \prod_i p(y_i|x_i). \]  
(6.22)

The other term of equation 6.20, i.e. \( p(x) \), is estimated as explained in
Figure 6.5: Recursive mean update, outliers present.

Section 6.3. In all previous data reconciliation techniques, \( p(x) \) was considered as a binary variable that takes a value of 1 if the process constraints are met and 0 otherwise. Under this assumption the data rectification problem reduces to a constrained optimization as follows:

\[
\max_{x} p(y|x). \tag{6.23}
\]

\[
s.t. \quad f(x) = 0.
\]

where \( f(x) \) represents the set of process constraints. If in addition to the binary assumption for \( p(x) \), sensors are assumed to follow normal distrib-
Figure 6.6: Recursive variance update, outliers present.

ution, then the optimization problem of equation 6.23 further reduces to a constrained least squares problem as follows:

\[
\min_{\hat{x}} (y - x)^{-1}Q^{-1}(y - x).
\]

\[\text{s.t. } f(x) = 0.\]  \hspace{1cm} (6.24)

For a set of linear constraints expressed as \( f(x) = Ax = 0 \), the solution of the constrained least squares objective function is found as [17]:

\[
\hat{x} = (I - QA^T(AQA^T)^{-1}A)y.
\]

\hspace{1cm} (6.25)
6.5 The maximum likelihood data rectification approach

The maximum likelihood data rectification approach relaxes the following assumptions:

1. Sensors contain random noise only, i.e. the sensor error is modeled by a zero-mean Gaussian distribution.

2. The prior probability density function of process states, $p(x)$, is assumed to be a binary variable having a value of 1 when the process constraints are met and 0 otherwise.

To relax the first assumption, the measurement error is modeled by the multimodal distribution function of equation 6.21. Assuming that the sensor operates either in its normal operation mode or in gross error mode, the multi-modal distribution of equation 6.21 is then reduced to the bimodal distribution of 6.3. This however, results in a distribution function that is not normal even when the sensor is in its normal operating mode.

To relax the second assumption, the prior probability distribution, $p(x)$, is modeled as a weighted sum of the probability density function corresponding to different conditions of the plant operation. It is important that all possible scenarios for the plant operations are considered. For a plant operating at different steady states, each of these steady states is represented by its corresponding pdf. Assuming that the plant operates at a number of mutually exclusive operating modes, $p(x)$ is calculated as:

$$
p(x) = \sum_{j} p(x|o_j)p(o_j). \quad (6.26)
$$
where \( p(x|o_j) \) is the probability density function of the plant states when the plant operating mode is \( o_j \) and \( p(o_j) \) is the weighting factor and represent the probability of plant operating at \( o_j \) [1].

Should the maximum likelihood data rectification be used for an operating condition that is not reflected in \( p(x) \), the resulting rectified values would be incorrect. In such circumstances, the rectified values contain large adjustments that are required to increase the contribution of \( p(x) \) to the objective function and that in turn, leads to small values of \( p(y|x) \). Consequently, the resulting objective function is small and the rectified values are not dependable.

The maximum likelihood data rectification problem is then stated by:

\[
\max_x p(x)p(y|x) = \max_x p(x)p(\epsilon). \tag{6.27}
\]

\[
s.t. \quad f(x) = 0.
\]

where \( p(y|x) \) and \( p(x) \) are expressed by equations 6.22 and 6.26 respectively.

An interesting observation in the formulation of equation 6.27 is the opposing balance between its terms. There is a trade off between the likely rectified states, \( p(x) \), and the likely adjustments to the measurements, \( p(\epsilon) \). The higher the likelihood of the state, the lesser the likelihood of its adjustment. The optimization seeks a solution that results in a probable rectified state, and also is a probable adjustment to the measurement set.

The trade off between the likely adjustments to the measurements and the likely measurements is conceptually depicted in figure 6.7. When a mea-
6.6 Solving the maximum likelihood rectification (MLR) problem

When a measurement arrives, the highest obtainable value for \( p(\epsilon) = p(y|x) \) for that measurement is when \( \epsilon = 0 \). This is shown in figure 6.7 with the plus sign and the concentric circles around it indicating equal probability contours of \( p(\epsilon) \). The elliptical contours represent the equal probability contours of \( p(x) \) in the \( x_1 - x_2 \) space. Given the prior probability distribution of states, if the measurements fall short of higher values of \( p(x) \) and hence the contribution of \( p(x) \) in the objective function is small, the optimization process tries to adjust the measurement so that the contribution of \( p(x) \) increases. This, in turn, leads to a decrease in contribution of \( p(\epsilon) \) in the objective function. The objective function is the product of these two contours. As these two terms are changing, the maximum likelihood objective function goes through a maximum point. This point represents the most likely rectified state given the measurement and a priori knowledge of \( p(x) \).

6.6 Solving the maximum likelihood rectification (MLR) problem

Given the observed plant measurements, the most likely values for the plant states are found by maximizing the conditional probability of the plant states expressed by the objective function of equation 6.27.

\[ p(y|x) \]
represents the probability of the adjustments to the measured values. Assuming that sensor errors are independent of measured process values, \( p(y|x) \) is equal to \( p(\epsilon) \) and is modeled by equation 6.4.

The functional form of \( p(x) \) generally depends on the method used for esti-
Figure 6.7: Trade off between the likely adjustment of measurements and the likely process rectified states.

mating the probability density function for plant states. The non-parametric probability density estimation technique of adaptive mixtures is used to produce an estimate of \( p(x) \). This method depicts the state probability density function as a mixture of weighted Gaussian functions. It is also capable of detecting new entities in the data and conforming to the new state of the system. The choice of method for state probability density estimation and its merits is thoroughly explained in Section 5.4. Consequently, \( p(x) \) is modeled
6.6 Solving the maximum likelihood rectification (MLR) problem

as:

\[ p(x) = \sum_j \omega_j \phi_j. \quad (6.28) \]

where \( j \) counts the number of components and \( \phi_j \) is a Gaussian:

\[ \phi_j(x; \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{n/2}|\Sigma_j|} \exp \left( -\frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) \right). \quad (6.29) \]

If the MLR is subject to constraints \( f(x) = 0 \), then to implement the constraints, the \( x \) data may be projected to a lower dimension by being premultiplied by \( p \), where \( p \) is the projection matrix whose columns span the null space of the constraints surface represented by \( f(x) \). Then the MLR problem of equation 6.27 can be rewritten as:

\[
\max_x \left\{ \prod_{i=1}^s (1 - p_i) \phi(x_i; y_i, \sigma_i^2) + p_i \phi(x_i; y_i, b_i^2 \sigma_i^2) \right\} \left\{ \sum_{h=1}^H \omega_h \phi_h(px; \mu_h, \Sigma_h) \right\}. \quad (6.30)
\]

subject to the constraints.

6.6.1 An alternative approach to solving the MLR problem

As the number of sensors and the dimension of the system expands, solving the optimization problem of equation 6.30 becomes computationally intensive. Johnston and Kramer [96] proposed an alternative method that exploits the simplifications that arise when the objective function is a product of exponential terms. A new set of variables was proposed to represent the objective function in equation 6.30 as a product of exponentials with the same global
maximum as equation 6.30 [98]. Taking the natural logarithm of the new objective function results in a quadratic objective function in \( x \). The solution to the linearly constrained or unconstrained quadratic objective function can be found analytically.

The probability distribution of plant states, \( p(x) \), is defined as a mixture of Gaussians and is given by equation 6.28. \( p(x) \) can be considered as the marginal distribution of \( x \) derived from the joint probability distribution function of \( x \) and \( \gamma \), \( p(x, \gamma) \). There is one \( \gamma \) associated with each vector of measurements, \( x \). Elements of \( \gamma \) can only be one or zero indicating if the regarded vector of measurements belong to a component of the mixture density or not. As a result \( p(x, \gamma) \) can be defined as:

\[
p(x, \gamma) = \prod_{h=1}^{H} (\omega_h p(x|h))^{y_h}.
\]

where \( y_h \) is the \( h^{th} \) element of the vector of \( \gamma \). Then \( p(x) \), defined as the marginal probability derived from \( p(x, \gamma) \), is related to it as follows:

\[
p(x) = \sum_{\gamma} p(x, \gamma) = \sum_{\gamma} \prod_{h=1}^{H} (\omega_h p(x|h))^{y_h} = \sum_{h=1}^{H} \omega_h p(x|h) = \sum_{h=1}^{H} \omega_h \phi_h(px; \mu_h, \Sigma_h).
\]
6.6.2 Probability distribution function for error

Similar to the approach followed in the removal of the product of the summation terms for \( p(x) \), the error probability distribution for error, \( p(\epsilon) \) can be re-arranged so that \( p(\epsilon) \) includes only a product of Gaussians.

A new variable, \( \nu_i \), is introduced such that for each sensor \( i \), \( \nu_i = 1 \) if sensor contains a gross error and \( \nu_i = 0 \) otherwise. \( p(\epsilon) \), can be regarded as the marginal distribution derived from \( p(\epsilon, \nu) \):

\[
p(\epsilon, \nu) = \prod_{i=1}^{s} \left( (1 - p_i) \phi(x_i; y_i, \sigma_i^2) \right)^{1 - \nu_i} \left( p_i \phi(x_i; y_i, b_i^2 \sigma_i^2) \right)^{\nu_i} \tag{6.33}
\]

since for all \( \nu_i \) (\( i = 1, ..., s \)):

\[
(1 - p_i) \phi(x_i; y_i, \sigma_i^2) + p_i \phi(x_i; y_i, b_i^2 \sigma_i^2) \equiv \left( (1 - p_i) \phi(x_i; y_i, \sigma_i^2) \right)^{1 - \nu_i} \left( p_i \phi(x_i; y_i, b_i^2 \sigma_i^2) \right)^{\nu_i}
\]

\[
p(\epsilon) = \prod_{\nu} \prod_{i=1}^{s} \left( (1 - p_i) \phi(x_i; y_i, \sigma_i^2) \right)^{1 - \nu_i} \left( p_i \phi(x_i; y_i, b_i^2 \sigma_i^2) \right)^{\nu_i} \tag{6.35}
\]

\[
= \prod_{i=1}^{s} \left( (1 - p_i) \phi(x_i; y_i, \sigma_i^2) + p_i \phi(x_i; y_i, b_i^2 \sigma_i^2) \right).
\]

By introducing equations 6.31 and 6.33 into equation 6.30, the resulting alternative objective function would be of the form:

\[
\max_{x, \gamma, \nu} \left\{ \prod_{i=1}^{s} \left( (1 - p_i) \phi(x_i; y_i, \sigma_i^2) \right)^{1 - \nu_i} \left( p_i \phi(x_i; y_i, b_i^2 \sigma_i^2) \right)^{\nu_i} \right\} \cdot \left\{ \prod_{h=1}^{H} \omega_h \phi_h(p_x; \mu_h, \Sigma_h)^{\gamma_h} \right\}, \tag{6.36}
\]
After taking the natural logarithm, the objective function of equation 6.36 is re-written as:

\[
\max_{z,\gamma,\nu} \left\{ - \left( \sum_{h=1}^{H} \gamma_h (px - \mu_h) \Sigma_h^{-1} (px - \mu_h) + (P^T px - y) \Lambda^{-1} (P^T px - y) + C \right) \right\}. \tag{6.37}
\]

In equation 6.37 all the constants which are independent of \( x \) are collected in term \( C \), and \( \Lambda^{-1} \) is defined as:

\[
\Lambda^{-1} = \text{diag} \left( \frac{1 - \nu_i}{\sigma_i^2} + \frac{\nu_i}{b_i^2 \sigma_i^2} \right). \tag{6.38}
\]

To make the format of observed data consistent for both terms of equation 6.37, the \( x \) data is presented as \( px \), but has been pre-multiplied by \( P^T \). So whether a null space exists or not, \( P^T P = I \) always ensures the correct format of equation.

By re-arranging equation 6.37 and removing the constant terms that have no effect on the maximization, the new form of MLR objective function is found to be as follows:

\[
\min_{z,\gamma,\nu} (px - \omega)^T S^{-1} (px - \omega). \tag{6.39}
\]

where:

\[
S = \left\{ \sum_{h=1}^{H} \gamma_h \Sigma_h^{-1} + P\Lambda^{-1}P^T \right\}^{-1}. \tag{6.40}
\]
and:

$$
w = S \left\{ \sum_{h=1}^{H} \gamma_h \Sigma_h^{-1} \mu_h + P\Lambda^{-1} y \right\}. \quad (6.41)$$

Providing the values of $\gamma$ and $\nu$ are known, the objective function of equation 6.39 is a quadratic function in $x$ that provides an analytical solution to the MLR problem.

Johnston and Kramer [96] proposed using the method of Expectation-Maximization (EM) to find the most likely values for parameters $\gamma$ and $\nu$ as well as the rectified states, $x$. The proposed EM algorithm consists of two steps. In the E-step, a set of unknowns, called missing data are estimated using the model. Missing data vary from a portion of the observed data that is not reported due to some failures, to the whole data when one tries to find the most likely values of the whole data, to some parameters that are introduced to facilitate the treatment of the data. The model itself and its parameters are updated in the M-step using the most likely values of the missing data obtained in the E-step. The EM algorithm proceeds alternatively between these two steps until the convergence is achieved. It is guaranteed that after each iteration the set of the model parameters obtained are at least as likely as the set of model parameters obtained at the previous iteration [8]. Given the observed data and the starting point, the monotonic increase in the likelihood function guarantees that the obtained result is the most likely one.
6.6 Solving the maximum likelihood rectification (MLR) problem

6.6.3 The expectation step

In this step, the expected values for the new variables are estimated using the values for $x$ data obtained at the previous iteration. For the first iteration, the $x$ data is approximated by the measured values, $y$. The Expectation relations are as follows:

$$\gamma^k = E(\gamma(x^k, y)) = \sum_{\gamma} \gamma p(\gamma|x^k, y). \quad (6.42)$$

$$\nu^k = E(\nu(x^k, y)) = \sum_{\nu} \nu p(\nu|x^k, y). \quad (6.43)$$

The summation in equation 6.42 is for all $H$ possible combinations of $\gamma$. The summation in equation 6.43 is for all $2^s$ combination of $\nu$. Equations 6.42 and 6.43 can be re-written as:

$$\gamma^k = \sum_{\gamma} \gamma p(\gamma|x^k, y) = \sum_{\gamma} \gamma \frac{p(\gamma, x^k|y)}{p(x^k, y)} \quad (6.44)$$

$$= \sum_{\gamma} \gamma \frac{p(\gamma, x^k)p(y|\gamma, x^k)}{p(x^k)p(y|x^k)}.$$ 

$$\nu^k = \sum_{\nu} \nu p(\nu|x^k, y) = \sum_{\nu} \nu \frac{p(\nu, y|x^k)}{p(y|x^k)}. \quad (6.45)$$

The information about the component of the prior distribution, $p(x)$, (which an observation comes from) can be extracted from the estimate of the current states of the process which is another expression for the independence of the measurement error from the process states. As a result equation 6.44
may be reduced to:

$$
\gamma^k = \sum_{\gamma} \gamma \frac{p(\gamma, x^k)}{p(x^k)}.
$$

(6.46)

The $j^{th}$ element of $\gamma$ is found by substituting equations 6.31 and 6.32 into equation 6.46:

$$
\gamma_j^k = \frac{\omega_j \phi(px^k; \mu_j, \Sigma_j)}{\sum_{j=1}^{H} \omega_j \phi(px^k; \mu_j, \Sigma_j)}.
$$

(6.47)

where $\sum_{j=1}^{H} \gamma_j^k = 1$.

Similarly, by substituting equation 6.3 into equation 6.45, the $j^{th}$ elements of $v$ can be written as:

$$
v_j^k = \frac{p_j \phi(x_j^k; y_j, b_j^2 \sigma_j^2)}{(1 - p_j) \phi(x_j^k; y_j, \sigma_j^2) + p_j \phi(x_j^k; y_j, b_j^2 \sigma_j^2)}
$$

(6.48)

The last two equations constitute the Expectation part of the EM algorithm.

### 6.6.4 The maximization step

Given the values of the auxiliary variables $\gamma^k$ and $\nu^k$ estimated in the Expectation step, the Maximization step produces the most likely values for the $x$ data. The solution to the quadratic objective function of equation 6.39 with no constraints is found as:

$$
p_x = \omega
$$

(6.49)
and by replacing $\omega$ from equation 6.41, $x$ is found as:

$$x^{k+1} = p^T \left\{ S^k \left( \sum_{h=1}^{H} \gamma^k_h \Sigma^{-1}_h \mu_h + PA^{-1,k} y \right) \right\}.$$  \hfill (6.50)

If $x$ is in its full dimension (that is $p$ is equal to the identity matrix) and a set of linear constraints $Ax = c$ exists, then the solution to the quadratic objective function of equation 6.39 is found as:

$$x^{k+1} = \omega^k - S^k A^T (A S^k A^T)^{-1} (A \omega^k - c).$$  \hfill (6.51)

In equations 6.50 and 6.51, $S$, $\omega$ and $A$ are evaluated by equations 6.40, 6.41 and 6.38, respectively. If there is no prior distribution for $x$ available, the related terms (that is the summations over $H$ components of the mixture distribution) are ignored.

### 6.6.5 Initialization of the EM algorithm

The obvious choice for initializing the $x$ data to evaluate $\gamma$ is the observed measurements themselves:

$$\gamma^0_j = \frac{\omega_j \phi(p y_j; \mu_j, \Sigma_j)}{\sum_{j=1}^{H} \omega_j \phi(p y_j; \mu_j, \Sigma_j)}$$  \hfill (6.52)

Initializing the $\nu$ values using the observed measurements, $x = y$, however, results in a set of $\nu_j$'s which are close to zero because the assumption of $x = y$ suggests that there are no gross errors in the observed measurements.
Consequently, using $x = y$ as the starting point biases the algorithm toward no gross errors in the plant data. That in turn, affects the compromise between the two terms of the MLR objective function (as discussed earlier and shown in equation 6.27). Although by assuming $x = y$ it is presumed that there are no gross errors in the observed data, the information stored in $p(x)$ can be exploited to avoid the bias and preserve the ability to detect the existing gross errors in the data. If the observed data contains only noise, then it is likely to lie in the probable region of $p(x)$ and if it contains gross errors it is likely to lie in the low probability or close to zero probability region of $p(x)$. It is proposed that the best initial estimate for $\nu$ is its expected value given the initial values of $\gamma$ and $p(x)$ [96]:

$$\nu^0 = E(\nu | y, \gamma^0, \mu, \Sigma)$$  \hspace{1cm} (6.53)

and therefore:

$$\nu_i = \frac{p_j \phi(y_j; u_j, v_j^2 + b_j^2 \sigma_j^2)}{(1 - p_j) \phi(y_j; u_j, v_j^2 + \sigma_j^2) + p_j \phi(y_j; u_j, v_j^2 + b_j^2 \sigma_j^2)}$$  \hspace{1cm} (6.54)

where $v_i^2$ is the $i^{th}$ element on the diagonal of the matrix $V$ which is defined as:

$$V = \sum_{h=1}^{H} \gamma_h (p^T \Sigma_h^{-1} p).$$  \hspace{1cm} (6.55)

and $u_i$ is the $i^{th}$ element of the vector $u$:

$$u = \left[ \sum_{h=1}^{H} \gamma_h (p^T \Sigma_h^{-1} p) \right]^{-1} \left[ \sum_{h=1}^{H} \gamma_h p^T \Sigma_h^{-1} \mu_h \right].$$  \hspace{1cm} (6.56)
6.6 Solving the maximum likelihood rectification (MLR) problem

6.6.6 On obtaining the global maximum

Since the mean for $p(y|x)$ is the observed value itself and the measured value presumably contains errors, the highest value for $p(y|x) = p(\varepsilon)$ is at the measured points. As the measurements get rectified, the probability of $p(y|x) = p(\varepsilon)$ decreases while the $p(x)$, the probability of the estimate of the process states getting closer to the true process states, increases.

The MLR objective function is a product of these two probabilities. Most of the probability density functions including a mixture of Gaussians, asymptotically approach zero as they move away from the bulk of the probability mass. As a result, in the tails of the probability distribution the probability values change very slowly. So when the measurement contains an error, it lies in the tails of $p(x)$. As the measured values get rectified, the $p(x)$ changes very slowly while $p(\varepsilon)$ goes through a maximum. Similarly, if the measured value does not contain an error (or the error is very small), it lies in the tails of $p(\varepsilon)$ and while $p(\varepsilon)$ is slowly changing for that measurement, $p(x)$ changes much more quickly and goes through its maximum. Consequently, the product of $p(x)$ and $p(\varepsilon)$ often shows its maximum in the region between these two extremes and a local maxima may appear [1].

The presence of local maxima can in turn lead to a local maximum solution for the MLR objective function which would result in an erroneous solution. To avoid this possibility, one approach is to start the optimization by randomly selecting different starting points to arrive at different solutions and then pick the solution with the largest MLR objective function value as the solution. The shortcoming of this approach is that it can be very
6.6 Solving the maximum likelihood rectification (MLR) problem

computationally demanding. A two step approach has been proposed \[96\] to remedy this problem. In the first step the likelihood of the local solution is evaluated to determine whether it is a likely global maximum. If in this first step the global solution was not reached, in a second step a new starting point is selected from the prior distribution of the rectified plant states \(p(x)\).

To decide on whether a global solution has been reached, a histogram of MLR objective function values found in estimating \(p(x)\) from a set of calibration data is formed. Since the MLR objective function of equation 6.30 is the likelihood of the rectified states given the measurements and prior distribution, its value would be an indication of how likely a particular solution is.

A lower objective function limit is set to include a certain percentage of the calibration cases (for example 97.5%). If the final objective function value is below this limit, then it is concluded that the solution reached is a local maxima and a new starting point is required. The first starting point used is the measurement, \(y\), itself. The initial values for \(\gamma\) and \(\nu\) are obtained according to section 6.6.5. If the likelihood of the obtained solution is below the threshold, the problem is re-solved using a new initial guess and the EM optimization algorithm is started at the E-step (equations 6.47 and 6.48), instead of using the initialization equations for \(\gamma\) and \(\nu\).

Since all valid solutions to the MLR problem have to lie in \(p(x)\), instead of randomly choosing another starting point, a searching point is randomly drawn from \(p(x)\). The problem with entirely random selection of the starting points is that it may result in some points that are unlikely to yield a valid so-
olution to the MLR problem. To make this process more efficient, a number of points are drawn from \( p(x) \) and the likelihood of each point being a solution to the MLR problem given the measurement, \( y \), is calculated using equation 6.30. To find the likelihood of a starting point both \( p(x) \) and \( p(\varepsilon) \) are estimated. While \( p(\varepsilon) \) depends on the particular measurement being treated, a starting point has the same value of \( p(x) \) for all the measurements until \( p(x) \) itself is updated. In the process of estimating \( p(x) \), its value for each of the starting points is calculated off line. As a result, in calculating the likelihood of the starting points, only \( p(\varepsilon) \) needs to be calculated and that accelerates the whole rectification process. The points are then sorted according to their likelihood value in descending order and are used as starting points.

The process starts with the measurement, \( y \), as the initial estimate for the true plant states and initial estimates for \( \gamma \) and \( v \). Using these initial values, the MLR problem in its quadratic form is solved. The likelihood for the resulting solution is then checked and if its likelihood fell below the limit, it is concluded that the solution to the MLR problem is a local maxima. The next starting point is then selected from the ordered starting points and the problem is re-solved.

If no likely solution is found, then it is likely that the true state of the plant that has generated the measurements lies in the extreme tails of \( p(x) \) and as a result, the true state itself is very unlikely.
6.6.7 The overall data rectification scheme

The overall process of data rectification is illustrated in figure 6.8. Given the prior probability density function of the measured process data, $p(x)$, and the probability density functions representing sensor error, $p(\epsilon) = \sum_{i=1}^{S} p(\epsilon_i)$ where $S$ is the number of sensors, the incoming measured process data is rectified. Using the EM algorithm in the rectification scheme, the most likely value of the vector of the measured process data is found. The rectified data point, as explained in Sections 6.3 and 6.4, is then used to recursively update $p(x)$ and $p(\epsilon)$.

![Diagram](image)

**Figure 6.8:** Recursive parameter update for sensor errors, and process states, probability density functions in data rectification.
6.7 Examples

Two chemical engineering examples of maximum likelihood data rectification are presented in this section. The first example is a heat exchanger network adapted from Tjoa and Biegler [50]. The second example is a simplified ammonia synthesis loop taken from Mah [23]. In both cases, the performance criterion for assessing the efficiency of the data rectification scheme is the mean-squared-error and is calculated as follows:

$$MSE = \frac{1}{SN} \sum_{s=1}^{S} \sum_{n=1}^{N} \frac{(\hat{x}_{ns} - \mu_s)^2}{\sigma_s^2}$$  \hspace{1cm} (6.57)

where $S$ is the dimension of the measured data (number of sensors that measured data are obtained from), $N$ is the number of measurement vectors, $\mu_s$ and $\sigma_s^2$ are the estimated mean and variance for the $s^{th}$ sensor and $\hat{x}$ is the rectified values. For an ideally rectified set of data where the true value for every plant state is found, the MSE is equal to zero. When the data contains random noise only, the value of MSE approaches 1 as the number of measurements approaches infinity. For a data set that contains gross errors, the value of MSE is greater than one. A decrease in the MSE value is an indication of removing outliers from the data set. Data rectification must bring the MSE value below one and drive it toward zero to be effective.

In both examples, all of the sensor errors were modeled by a bivariate Gaussian distribution as indicated by equation 6.3. The values of this bivariate Gaussian distribution parameters, $b$ and $p$, were assumed to be 20 and 0.05, respectively.
6.7 Examples

6.7.1 The heat exchanger network example

The heat exchanger network of Tjoa and Biegler [50], with modifications as detailed in the following, is selected as a chemical process example to illustrate the rectification of the measured process data using the maximum likelihood technique and EM iteration.

The heat exchanger network as depicted in figure 6.9 consists of four heat exchangers. Fluid A is heated by fluid B in two of the heat exchangers (HX1 and HX2) and by fluid C and fluid D in two other heat exchangers (HX3 and HX4). The flow of stream D is manipulated to maintain the temperature of the stream A8 that leaves the heat exchanger network. The input flows and temperatures were drawn from Gaussian distributions as shown in table 6.1. It can also be seen from table 6.1 that the noise added to the training data set had a standard deviation of 0.75 for temperatures and standard deviations of 4.0, 2.0, 3.0 and 3.0 for flows $F_{A1}$, $F_{B1}$, $F_{C1}$ and $F_{D1}$ respectively. Stream A2 is split into streams A3 and A6 such that $F_{A3} = 0.42F_{A2}$. The process was simulated in HYSYS™ using hypothetical components and the values of the measured variables were recorded at the given input conditions.

A data set of 100 points was generated and used to train the mixture of probability density functions, $p(x)$, that represent the prior probability density function of the plant states using the method of adaptive mixtures as explained in Chapter 5. This data set was also used to find the parameters of the Gaussian distributions that represent sensor operation.

A separate data set of 1000 points was generated for rectification and performance evaluations. For this data set, it was assumed that sensors
contain only normal noise. To investigate the effect of the presence of gross errors on the data rectification scheme, gross errors were introduced into the data set. Three more data sets were generated, one with a bias in $F_{C1}$ drawn from $U[-50]$, the second with a sensor failure in $F_{B1}$ drawn from $N(350, 5)$ and the last one with two sensor failures in $F_{B1}$ and $T_{B1}$ drawn from $N(350, 5)$ and $N(665, 5)$ respectively.

![Diagram of a heat exchanger network](image)

**Figure 6.9:** The heat exchanger network for the first data rectification example, Tjoa and Biegler [50].

The data rectification for the network of heat exchangers is done without incorporation of any process constraints. Using principal component analysis, the size of the measured process data set was reduced from 13 to 2 cap-
Table 6.1: Probability distribution properties for process variables of the heat exchanger network.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Distribution</th>
<th>Noise Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{A1}$</td>
<td>N(465,5)</td>
<td>0.75</td>
</tr>
<tr>
<td>$F_{A1}$</td>
<td>N(800,5)</td>
<td>4</td>
</tr>
<tr>
<td>$T_{B1}$</td>
<td>N(620,5)</td>
<td>0.75</td>
</tr>
<tr>
<td>$F_{B1}$</td>
<td>N(200,5)</td>
<td>2</td>
</tr>
<tr>
<td>$T_{C1}$</td>
<td>N(670,5)</td>
<td>0.75</td>
</tr>
<tr>
<td>$F_{C1}$</td>
<td>N(240,5)</td>
<td>3</td>
</tr>
<tr>
<td>$T_{D1}$</td>
<td>N(688,5)</td>
<td>0.75</td>
</tr>
<tr>
<td>$F_{D1}$</td>
<td>N(514,5)</td>
<td>3</td>
</tr>
</tbody>
</table>

turing 70% of the existing variation in the measured variables. For process data rectification in the presence of random noise only, the MSE value was reduced from 1.0181 calculated for the measured process values to 0.8268 calculated for the rectified data. Figure 6.10 depicts the comparison between the rectified values and measured process data for variable $T_{A8}$. Figure 6.11 shows that the MLR objective function, which is defined as a log-likelihood function, increases monotonically during a typical EM iteration in the data rectification process, until it reaches a maximum plateau.

To examine the ability of the maximum likelihood data rectification in the presence of biases, a bias of $U[-50]$ was introduced to $F_{B3}$ and the resulting set of measured data was rectified. In this case, the MSE value of 1.0310 for the measured data was reduced to 0.8297. A comparison between rectified and measured values for $F_{C1}$ is displayed in figure 6.12.

Another test of MLR, failure to a fixed value, was simulated by drawing the measured values for $F_{B1}$ from $N(350,5)$. Figure 6.13 compares the rectified values for $F_{C1}$ with its erroneous measured values and demonstrates the ability of the rectification scheme to provide the rectified values when
Figure 6.10: Rectified and measured value of $T_{A8}$, no gross error, the heat exchanger network.

One sensor has failed. The MSE value is also reduced from 1.0988 to 0.7736. Even when two sensor failures for $T_{B1}$ and $F_{B1}$ were incorporated into the measured data, the MLR approach was capable of providing a reliable estimate of the process states and lowering the MSE value from 1.0870 to 0.7790. Figure 6.14 illustrates that the rectified values for $T_{B1}$ and $F_{B1}$ are reliable estimates of the true values for these variables. A summary of the MSE values for the heat exchanger network example is provided in table 6.2.

To investigate the effect of the bivariate Gaussian distributions used to model the sensors (equation 6.3) on the data rectification scheme, the MSE was calculated for different values of 'b' and 'p'. 'p' is the probability of a
Figure 6.11: The MLR objective function value in the EM iteration.

Table 6.2: MSE values for measured and rectified process variables for the heat exchanger network.

<table>
<thead>
<tr>
<th>MSE Value</th>
<th>Measured Data</th>
<th>Rectified Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Noise only</td>
<td>1.0181</td>
<td>0.8268</td>
</tr>
<tr>
<td>Bias in FC1</td>
<td>1.0310</td>
<td>0.8297</td>
</tr>
<tr>
<td>Failure in FB1</td>
<td>1.0988</td>
<td>0.7736</td>
</tr>
<tr>
<td>Failures in FB1 and TB1</td>
<td>1.0870</td>
<td>0.7790</td>
</tr>
</tbody>
</table>

gross error happening while 'b' is the width of the gross error distribution. The data sets generated for the heat exchanger network, for cases of normal noise only and one sensor failure, were also used in these calculations. The results for the measured data set containing normal noise only is presented in table 6.3 where table 6.4 shows the MSE values when the measured process data contained one sensor failure. The results show that the choice of 'p' and
Figure 6.12: Rectified and measured value of $F_{C1}$, with a bias of -50 in its measured values, the heat exchanger network.

Table 6.3: Effect of ‘b’ and ‘p’ on process data rectification results, no gross error, for the heat exchanger network.

<table>
<thead>
<tr>
<th>MSE</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>10</td>
</tr>
<tr>
<td>0.05</td>
<td>0.7856</td>
</tr>
<tr>
<td>0.10</td>
<td>0.7810</td>
</tr>
<tr>
<td>0.15</td>
<td>0.7806</td>
</tr>
</tbody>
</table>

'b' in the presence or absence of gross errors does not affect the efficacy of the maximum likelihood data rectification.
Figure 6.13: Rectified and measured value of $F_{B1}$, with a sensor failure, drawn from N(350,5), the heat exchanger network.

Table 6.4: Effect of ‘b’ and ‘p’ on process data rectification results, one failed sensor, for the heat exchanger network.

<table>
<thead>
<tr>
<th>MSE</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td>0.05</td>
<td>0.7704</td>
</tr>
<tr>
<td>0.10</td>
<td>0.7673</td>
</tr>
<tr>
<td>0.15</td>
<td>0.7669</td>
</tr>
</tbody>
</table>

6.7.2 The ammonia synthesis example

The simplified ammonia synthesis example is taken from Mah [23] and is shown in figure 6.15. The process stream F1 is mixed with the recycle stream F5 in a mixer. The stream exiting the mixer, F2, goes to the reactor. Stream F3 that leaves the reactor is a mixture of reactants and products. They are
Figure 6.14: Rectified and measured value of process variables $F_{B1}$ and $T_{B1}$, with sensor failures, drawn from $N(350,5)$ and $N(665,5)$ respectively, the heat exchanger network.

partially separated in the separator where the stream F4 leaves the system and the remaining material is recycled back to the mixer via stream F5.

The measured process variables in this example consisted of flows F1, F3 and F5. The flow of F1 was specified and the other flows were calculated. For this example, they were calculated as $F1 = 1000$, $F3 = 3333.2$ and $F5 = 2333.2$. The measured values were generated by adding normal Gaussian noise of variances 1, 2 and 1. A training data set of 100 points was generated to estimate the prior probability distribution function of the measured process states as explained in Chapter 5. Using the method of
principal component analysis, the dimension of data was reduced from 3 to 2 preserving 90\% of variation in the measured process variables. A separate data set of 1000 data points was generated for investigating the efficiency and performance of the rectification technique.

In the first case, a data set containing random noise only is rectified. The MSE value of 0.9990 for the measured process data was reduced to 0.0866 for the rectified data. Figures 6.16, 6.17 and 6.18 depict the rectified and measured process values for variables F1, F3 and F5, respectively, and compare them with their respective true values. These figures illustrate the ability of the rectification technique in providing reliable estimates and reducing the measured data variation. To examine the effect of the presence of gross errors, the F3 sensor was assumed to fail to a fixed value and its measurements drawn from N(2833,20). The maximum likelihood data rectification was able to reduce the MSE value from 1.0866 to 0.0505 and provide a reliable estimate of the true plant states in the presence of outliers. Figures 6.19, 6.20 and 6.21 demonstrate the competence of the rectification method in obtaining reliable estimates of variables F1, F3 and F5, respectively.

To compare the results of the maximum likelihood data rectification with a model-based data rectification technique, the measured process data was also reconciled using weighted least-squares (WLS) method. Since a number of process variables are not measured, the matrix projection method of Crowe et al. [18] was used to project the process constraints to a lower dimensional space where all the variables are measured. Using this approach for measured process data containing normal noise only, the MSE value was reduced from
0.9990 to 0.6533. Figure 6.16 depicts the result of the weighted least-squares data reconciliation for process variable F1 and compares it with the rectified values obtained from the MLR technique. The significant reduction in MSE value can also be observed as a substantial drop in the variation of rectified process values. The same effect can be seen from figures 6.17 and 6.18 for process variables F3 and F5, respectively.

In the presence of outliers, simulated as a failure to a fixed value in process variable F3, the weighted least-squares data rectification fails to perform any reconciliation on the measured data. The resulting estimates contained significant errors. Gross errors present in the process variable F3 propagated throughout the process, contaminating all of the estimated values of the true plant state. This effect is referred to as smearing [19]. Consequently, the MSE value increased from 1.0866 for the measured data to 637.9933 for WLS while it had a value of 0.0505 for the maximum likelihood rectified data. The same conclusions can be drawn from figures 6.19, 6.20 and 6.21, depicting the rectified values for measured process variables F1, F3 and F5 respectively. The estimates of the process variables obtained from weighted least-squares data reconciliation deviated significantly from the true plant states. The MLR data rectification, however, was capable of rejecting the outliers and providing a reliable estimate of the measured process values. The MSE values calculated for rectified and measured values of the ammonia synthesis example are summarized in table 6.5.
### 6.8 Summary

The MLR objective function consisted of two terms. One is the prior probability density of the process states, stated by a mixture of probability densities. The other is a product of probability density functions, each representing the likelihood of the adjustments to the measurements, i.e. the sensor errors. Given the complicated nature of the resulting objective function the use of nonlinear programming techniques could be time consuming. Alterna-
Figure 6.16: Rectified and measured values of F1, no gross error error present, the ammonia synthesis example.

tively, an iterative EM technique is used to solve the MLR problem. To reflect on changes in the process behaviour, the parameters of the MLR objective function are updated with the new measured process data within the framework of quasi steady state operation. A robust recursive method of updating the parameters of the bimodal probability density functions, that represent sensor errors, is introduced and its capability in isolating the outliers from the updating process is illustrated. The parameters of the probability density mixture that represented the prior pdf of the process states, are recursively updated by method of adaptive mixtures using the rectified measurements.

The examples provided demonstrate the ability of the maximum likeli-
Figure 6.17: Rectified and measured values of F3, no gross error error present, the ammonia synthesis example.

hood data rectification technique in rectifying measured process data and finding their most likely values both in the presence of random noise and gross errors.
Figure 6.18: Rectified and measured values of F5, no gross error present, the ammonia synthesis example.

Figure 6.19: Rectified and measured values of F1, sensor failure for F3, the ammonia synthesis example.
Figure 6.20: Rectified and measured values of F3, sensor failure for F3, the ammonia synthesis example.

Figure 6.21: Rectified and measured values of F5, sensor failure for F3, the ammonia synthesis example.
Chapter 7

Conclusions and

Recommendations for Future Research

7.1 Conclusions

In order to provide a description of the state of a chemical plant, process variables must be measured. The measured process data is the foundation upon which all the control and evaluation of process performance are based. The reliability of the measured process data is key to their use in process control, optimization or operation of a chemical plant. However, the measured process data always contain measurement errors. Errors are either small noise originating from sensor irreproducibility or large errors, i.e. gross errors, that stem from sensor failures, biases, malfunctionings or miscalibrations. Data rectification is the process of mitigating existing errors in the
measured process data in order to obtain an estimate of the true process state.

This research has established a new approach to data rectification that is capable of rectifying the measured process data where the process model is not known or partially known. This approach uses historical plant data to achieve data rectification and is capable of incorporating additional process information like process constraints if they become available. It poses the data rectification problem in a probabilistic framework and finds the most likely values of the measured process variables, given the process measurements. It uses Bayes' theorem to express the probability density function of the process states given the process measurements as a product of the probability density of the measurements given the states and the prior probability density of the states. The maximum likelihood objective function is defined as this product. There exists a trade off between the two components of the maximum likelihood rectification objective function, i.e. the likely process states represented by $p(x)$, and the likely adjustments made to the measurements represented by $p(\epsilon)$. The resulting maximum likelihood data rectification objective function is maximized to obtain the most likely values of process states.

Calculation of the measurement error covariance matrix and also estimation of the covariance matrices for the components of the probability density mixture that represents the process states, are essential requirements in probabilistic data rectification. The underlying research of this dissertation has led to a new robust method of covariance matrix calculation. This approach
directly treats the measured process variables and uses a maximum likelihood estimator to reject the outliers and tune the measured values for deviations from steady state. A reliable estimation of the covariance matrix also plays a fundamental role in utilizing principal component analysis for reducing the dimensionality of the problem.

The prior probability density of the plant states is constructed using historical plant data in an offline training scheme. In real-world processes, there are many imperfections in the observations and some of the measured process values are missing. An unsupervised method of learning probability density function parameters in the framework of mixture densities from incomplete data is developed in this research. Unsupervised learning can be considered as learning from observations as they are received. The Expectation-Maximization algorithm is used to iteratively find the maximum likelihood estimate of the missing values and the parameters of the probability density function. The reliability of the Expectation-Maximization algorithm and its convergence properties during the learning process from incomplete data are presented and the boundaries of reliably estimating the mixture probability density function parameters as the percentage of missing data increases are examined.

The method of adaptive mixtures is used as a recursive nonparametric method for probability density estimation, for both offline training and online updating to estimate and update the probability density function of the process states. The method of adaptive mixtures recursively fits a mixture of Gaussian distributions to the measured process data. This mixture is not
constrained to a fixed size and can be expanded by adding new components when it finds new entities in the measured data.

The maximum likelihood data rectification objective function consists of two terms, the probability density of the process states and the probability density of the sensor errors. The latter is modeled by bimodal Gaussian distributions. Normal sensor operation is represented by one mode while the other mode characterizes the gross error operation for sensor. All of erroneous modes of sensor operation are embodied in one single gross error mode of operation. The Expectation-Maximization method is then used to iteratively find the set of process variables that maximizes the maximum likelihood objective function. Both the probability density of the process states and those of the sensor errors are then updated using the rectified point.

The approaches developed were tested on numerical examples and also implemented on chemical process examples, thus quantifying robust estimates for covariance matrices, reliable estimate of mixture probability densities, rejecting the errors existing in the measurements and yielding reliable rectified values for the measured process variables.

7.2 Future studies

7.2.1 Dimensionality; what is the best approach?

Generally, the probability density of the process states is multi-dimensional and multi-modal and a representative model is not known. Nonparametric
or semiparametric density estimation techniques are capable of capturing the unknown density but they depend on the data.

As the dimension of the process data grows, the required number of calculations grows exponentially. This problem is known as the “curse of dimensionality” [99]. The curse of dimensionality refers to the exponential growth of hypervolume as a function of dimensionality. Processing data of high dimensions consume resources and the amount of required resources is proportional to the hypervolume of the data [100].

Unsupervised learning algorithms are typically prone to this problem. A partial treatment is to preprocess the data and scale it according to its importance [89]. It should be noted that even if the more important portions of data are processed, the higher the dimensionality of the data and the more data is needed to discover what is important in data.

One happenstance for reducing data dimensionality occurs when the maximum likelihood data rectification problem is solved subject to a set of linear constraints. The linear relationship that exists among the rectified values causes singularity in the covariance matrix of the Gaussian components of the mixture density and as such, $p(x)$ cannot be updated. To avoid the singularity, the measured process data is projected from its original dimension to a lower dimension of the constraints' hyperplane. The projection matrix is constructed such that its orthonormal columns span the null space of the plane described by $Ax = 0$ [101]:

$$Ax = (AP)(P^Tx) = 0.$$  \hspace{1cm} (7.1)
The probability density of process states, \( p(x) \), is then estimated for a new set of variables, \( z \), calculated as:

\[
z = P^T x.
\] (7.2)

The transformation from \( z \)-space back to \( x \)-space, however, is only valid when the equality constraints exist. When the process model is not available or the equality constraints do not exist, the above mentioned method is not applicable. Principal component analysis provides an alternative approach to reduce dimensionality. The fundamental idea of principal component analysis is to reduce the dimensionality of a data set which consists of a large number of interrelated variables, while retaining as much of the variation present in the data set as possible. This is achieved by transforming to a new set of variables, the principal components, which are uncorrelated and are ordered so that the first few retain most of the variation present in all of the variables [7]. The transformation matrix is constructed such that its columns are eigenvectors corresponding to the eigenvalues of the correlation matrix of the measured process variables. The number of columns of the transformation matrix is determined by the number of eigenvalues whose summation achieves a pre-determined required level of variation. But there is always the question of how many principal components adequately account for the total variation. Maintaining all the variations requires keeping all the principal components, \( i.e. \) no dimensionality reduction, and selecting a few principal components means losing some information. There are different methods for choosing
7.2 Future studies

a subset of principal components, but generally, they are \textit{ad hoc} rules-of-thumb, whose justifications are mainly intuitively plausible and that they work in practice [7].

7.2.2 Data rectification for dynamic systems

The maximum likelihood data rectification approach presented in this thesis is intended for quasi steady state for which the process is essentially at steady state except for slow drifts and sharp transitions [23]. At the quasi steady state condition, process variables may be a function of time but nevertheless obey steady state process constraints.

In case of multiple steady state operating conditions, all steady states should be considered and reflected in the learning process of $p(x)$. That, in turn, means that the transition from one steady state to another is conducted as a switch from one set of functions representing $p(x)$ and $p(\epsilon)$ to another.

Although the principles of maximum likelihood data rectification can be extended to dynamic systems, there are additional complications that have to be dealt with. The probability density function of the process states has to be fitted to the time varying process variables. The method of adaptive mixtures provides an appropriate tool for this objective but it has to be modified such that the weight assigned to the new incoming measured process data is increased. It also has to be coupled with a dynamic gross error rejection technique so that the outliers are not included in updating $p(x)$.

For batch processes, measured variables and time can be related by treating time as another variable [15] and applying the maximum likelihood data
rectification to the problem. This technique for general dynamic systems, however, needs to be further developed.

### 7.2.3 Statistical process control

Maximum likelihood data rectification as a probabilistic technique provides an ideal framework for statistical process control. Statistical process control is a systematic method of tracking, predicting and reducing variations.

Statistical process control consists of sampling and analyzing process data to determine if the process is in or out of control from a statistical point of view. If the process is found to be out of control, it is examined to determine the source of the unsought variations. Once these sources are determined, the required measures are taken to eliminate or minimize their effects. Statistical process control techniques can be used to identify gross errors and model errors.

Conventional statistical process control methods treat each measured variable independently and require no process knowledge [102]. This simplicity requires that multivariate interactions and correlations among process variables are ignored. These restrictions can be avoided if the process knowledge is incorporated into statistical process control and maximum likelihood data rectification provides a proper basis for achieving this objective.

### 7.2.4 Dynamic gross error detection

Most recognized gross error detection and elimination techniques are designed for steady state processes only. The maximum likelihood data rectification
is capable of rejecting outliers in steady state conditions but in transition from one steady state to another, it must be coupled with a dynamic outlier rejection method. Statistical analysis of residuals and use of neural networks are new approaches for dynamic gross error detection [52].
Bibliography


Appendix A

Expectation-Maximization

Suppose that one has a "Measure Space" $\mathcal{Y}$ of complete data and a measurable map $y \rightarrow x(y)$ of $\mathcal{Y}$ to a "Measure Space" $\mathcal{X}$ of incomplete data. Let $f(y|\Phi)$ be a member of a parametric family of probability density function defined on $\mathcal{Y}$ for $\Phi \in \Omega$, and suppose $g(x|\Phi)$ is a probability density function on $\mathcal{X}$ induced by $f(y|\Phi)$.

For a given $x \in \mathcal{X}$, the purpose of the EM algorithm is to maximize the incomplete data log-likelihood $L(\Phi) = \log(g(x|\Phi))$ over $\Phi \in \Omega$ by exploiting the relationship between $f(y|\Phi)$ and $g(x|\Phi)$.

For $x \in \mathcal{X}$, the set $\mathcal{Y}(x)$ is defined as $\mathcal{Y}(x) = \{y \in \mathcal{Y} : x(y) = x\}$. The conditional density $k(y|x, \Phi)$ on $\mathcal{Y}(x)$ is given by:

$$f(y|\Phi) = k(y|x, \Phi) g(x|\Phi).$$  \hspace{0.5cm} (A.1)

For $\Phi$ and $\Phi' \in \Omega$:

$$L(\Phi) = Q(\Phi|\Phi') - H(\Phi|\Phi').$$  \hspace{0.5cm} (A.2)
where

\[
Q(\Phi'|\Phi') = E(\log f(y|\Phi') | x, \Phi'). \quad (A.3a)
\]

\[
H(\Phi|\Phi') = E(\log k(y|x, \Phi) | x, \Phi'). \quad (A.3b)
\]

The general EM algorithm of Dempster et al. [8] is followed. In this algorithm, given a current approximation \(\Phi^c\) that maximizes \(L(\Phi)\), the next approximation \(\Phi^n\) is obtained by:

1. E-Step: Determine \(Q(\Phi|\Phi^c)\).

2. M-Step: Choose \(\Phi^n\), a set of values \(\Phi \in \Omega\) which maximizes \(Q(\Phi|\Phi^c)\).

The EM algorithm is used most often in applications which permit, relatively speaking, the easy maximization of \(\log f(y|\Phi)\) over \(\Phi \in \Omega\). This, in turn facilitate the M-step evaluation. The importance of EM algorithm lies in the fact that at any iteration \(L(\Phi^n) \geq L(\Phi^c)\). In fact, the way in which \(\Phi^n\) is determined guarantees that \(Q(\Phi|\Phi') \geq Q(\Phi|\Phi')\) and \(H(\Phi|\Phi') \leq H(\Phi|\Phi')\) [8].

This fact implies that \(L\) is monotonically increasing during any iteration sequence of the EM algorithm. The EM algorithm for a mixture density estimation problem associated with this family is derived by first interpreting the problem as one involving incomplete data and then obtaining the algorithm from its general formulation given above.

In calculating \(Q(\Phi|\Phi')\) of the E-step, the data set is considered as a sample of incomplete data by assuming each observation, \(x_i\), to be the
known part of a more general and incomplete observation $y_i$. For $\Phi = \{(\alpha_1, ..., \alpha_M), (\phi_1, ..., \phi_M)\} \in \Omega$, the sample variables $X = \{x_1, ..., x_N\}$ and $Y = \{y_1, ..., y_N\}$ have associated probability density function

$$g(x|\Phi) = \prod_{i=1}^{N} p(x_i|\Phi). \quad (A.4)$$

and

$$f(y|\Phi) = \prod_{i=1}^{N} \alpha_{ji} p_{ji} (x_i|\phi_{ji}) \quad j_1, ..., j_N = 1, ..., M. \quad (A.5)$$

where $j_i$ in equation. A.5 is an integer number between 1 and $M$ indicating the component population of origin.

As a result, for $\Phi' = \{(\alpha'_1, ..., \alpha'_M), (\phi'_1, ..., \phi'_M)\} \in \Omega$, the conditional density $k(y|x, \Phi')$ is given by:

$$k(y|x, \Phi') = \prod_{i=1}^{N} \frac{\alpha'_{ji} p_{ji} (x_i|\phi'_{ji})}{p(x_i|\Phi')}, \quad j_1, ..., j_N = 1, ..., M. \quad (A.6)$$

$Q(\Phi|\Phi')$ is then calculated as:

$$Q(\Phi|\Phi') = \sum_{j_1=1}^{M} ... \sum_{j_N=1}^{M} \sum_{i=1}^{N} \log (\alpha_{ji} p_{ji} (x_i|\phi_{ji})) \prod_{i=1}^{N} \frac{\alpha'_{ji} p_{ji} (x_i|\phi'_{ji})}{p(x_i|\Phi')}.$$

(A.7)

Considering the fact that $\sum_{j_1=1}^{M} ... \sum_{j_N=1}^{M} \alpha_{ji} p_{ji} (x_i|\phi_{ji}) = \alpha_j p_j (x_i|\phi_j)$:

$$Q(\Phi|\Phi') = \sum_{j=1}^{M} \sum_{i=1}^{N} \log (\alpha_j p_j (x_i|\phi_j)) \frac{\alpha'_j p'_j (x_i|\phi'_j)}{p(x_i|\Phi')} \quad (A.8)$$

$$= \sum_{j=1}^{M} \left[ \sum_{i=1}^{N} \frac{\alpha'_j p'_j (x_i|\phi'_j)}{p(x_i|\Phi')} \right] \log \alpha_j + \sum_{j=1}^{M} \sum_{i=1}^{N} \log p_j (x_i|\phi_j) \frac{\alpha'_j p'_j (x_i|\phi'_j)}{p(x_i|\Phi')}.$$
Equation A.8 permits the maximization problem to be divided into two separate maximization problems. The first one includes $\alpha_1, ..., \alpha_M$ and the second one deals with $\phi_1, ..., \phi_M$. Since $\log \alpha_1, ..., \alpha_M$ appear linearly in each function $Q(\Phi|\Phi')$ the first maximization problem has a unique solution. This solution can be explicitly determined regardless of the functional forms of the component densities $p_j (x_i|\phi_j)$. If $\phi_1, ..., \phi_M$ are mutually independent variables, then the second maximization problem further divides into $M$ component problems, each of which involves only one of the parameters $\phi_j$. Both the component problems and the maximization problem for the proportions have the appealing property that they can be regarded as weighted maximum likelihood estimation problems involving sums of logarithms, weighted by posterior probabilities that sample observations belong to appropriate components given the current approximate maximum likelihood estimate of $\Phi$.

Assuming the current maximum likelihood estimator maximizer, $\Phi^c$ as $\Phi^c = \{(\alpha_1^c, ..., \alpha_M^c), (\phi_1^c, ..., \phi_M^c)\}$, then the next set of parameters, $\Phi^n$ can be approximated as:

$$\alpha^n_j = \frac{1}{N} \sum_{i=1}^{N} \frac{\alpha^c_j p_j (x_i|\phi^c_j)}{p(x_i|\Phi^c)} \quad j = 1, ..., M. \quad (A.9)$$

$$\phi^n_j = \max_{\phi_j} \sum_{i=1}^{N} \log p_j (x_i|\phi^c_j) \frac{\alpha^c_j p_j (x_i|\phi^c_j)}{p(x_i|\Phi^c)} \quad j = 1, ..., M. \quad (A.10)$$

Each weight $\alpha^c_j p_j (x_i|\phi^c_j) / p(x_i|\Phi^c)$ is the posterior probability that $x_i$ originated in the $j^{th}$ component population, given the current approximate maximum likelihood estimate $\Phi^c$. 
Let’s consider \( p_j(x|\phi_j) \) as a multivariate normal density

\[
p_j(x_i|\phi_j) = \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right).
\]

(A.11)

and hence, the set of parameters

\[
\phi_j = (\mu_j, \Sigma_j).
\]

(A.12)

where \( \mu_j \in \mathbb{R}^n \) and \( \Sigma_j \) is a positive-definite \( n \times n \) symmetrical matrix. The same concept can be applied here to obtain the maximum likelihood estimates for parameters of the multivariate density function:

\[
\mu_j^n = \frac{\sum_{i=1}^{N} x_i \frac{\alpha^j_p(x_i|\phi_j^c)}{p(x_i|\phi^c)}}{\sum_{i=1}^{N} \frac{\alpha^j_p(x_i|\phi_j^c)}{p(x_i|\phi^c)}}.
\]

(A.13)

\[
\Sigma_j^n = \frac{\sum_{i=1}^{N} (x_i - \mu_j^c)(x_i - \mu_j^c)^T \frac{\alpha^j_p(x_i|\phi_j^c)}{p(x_i|\phi^c)}}{\sum_{i=1}^{N} \frac{\alpha^j_p(x_i|\phi_j^c)}{p(x_i|\phi^c)}}.
\]

(A.14)

\( \Sigma_j^n \) would be a positive-definite symmetric matrix, with possibility one, if \( N \), the number of data points (or observations) remains larger than \( n \), the dimensionality of the system. As can be seen from equation A.14, each \( \Sigma_j^n \) is in the convex hull of \( \{(x_i - \mu_j^c)(x_i - \mu_j^c)^T\}_{i=1,...,N} \), a set of rank-one matrices which are not positive definite. Thus there is no guarantee that a sequence of matrices \( \{\Sigma_j^k\}_{k=0,1,2,...} \) produced by the EM iteration will remain bounded from below. It is observed in practice that sequences of iterates
produced by the EM algorithm for a mixture of multivariate normal densities do occasionally converge to singular solutions, i.e. points on the boundary of $\Omega_j$ with associated singular matrices [68]. It is observed by Hosmer [103] that if enough labeled observations are included in a sample of a mixture of normal densities, then with probability one, the log-likelihood function attains its maximum value at a point at which the covariance matrices are positive definite. This consideration, in a similar manner, alleviates the problem of an EM iteration sequence having singular solutions as limit points.