Robust Probabilistic Principal Component Analysis Based Modeling with Gaussian Mixture Noises

by

Anahita Sadeghian

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

Doctor of Philosophy

in

Process Control

Department of Chemical and Materials Engineering

University of Alberta

Abstract

Most of industrial plants are heavily instrumented with a large number of sensors and analyzers to provide the data needed for process control and monitoring purposes. However, online and fast-rate measurements are not always available due to restricted availability and/or reliability of measurement techniques and devices. Even in cases where appropriate measuring devices are available, some key process variables are still determined offline by laboratory sample analysis or by means of often unreliable online analyzers. Such methods of process data acquisition are time consuming, often expensive in the long run, and introduce delays and discontinuities into their application. On top of that, the development of advanced process monitoring and control techniques is key to achieving profitability, meeting safety requirements and operating environmental friendly processes. This development stage requires the operational data to be recorded for the analysis of the problem.

A popular approach to make reliable data available fast and at lower cost is using predictive models. Predictive models are basically mathematical models of the process which can be developed based on the history of the plant and using available data. In some cases, if possible, adding first principles equations could better the accuracy of the model. It is important to have relevant data that are clean to an acceptable extent, and cover a meaningful time span of the process under study. These circumstances might not be available perfectly. Data quality, namely availability, accuracy, relevance, density, and frequency, is a pivotal determinant for the outcome of a model. Some common disputable phenomena are uncertainty, high-dimensionality in terms of the count of recorded features compared to that of sample points, outlying observation, missing records, nonlinearity, and non-Gaussianity. In this thesis we have

targeted a combination of the most relevant phenomenon in a chemical process record such as *uncertainty*, *high-dimensionality*, *outliers* and their*non-Gaussianity*.

Probabilistic models are potent in terms of dealing with uncertainties, so are principal component analysis (PCA) methods in handling high-dimensionality. As a result, probabilistic principal component analysis (PPCA) based models are considered as the motif for this research. Conventionally, for probabilistic principal component analysis based models, noise with a Gaussian distribution is assumed for both input and output observations. This assumption makes the model to be vulnerable to large random errors, earlier referred to as outliers. In this thesis, unlike the conventional noise assumption, a mixture noise model with a contaminated Gaussian distribution piece is adopted for probabilistic modeling to diminish the adverse effect of outliers, which usually occur due to irregular process disturbances, instrumentation failures or transmission problems. This is done by downweighing the effect of the noise component which accounts for contamination, on the model output prediction. This adoption is implemented in different settings: a scaled mixture noise model, a location mixture noise model and a switching noise model to account for the dynamic behaviour of noise, for either process noise or the measurement noise. More details will be cracked further in the main chapters.

Finally, in comparison with the conventional PPCA based model and specific robust PPCA based models, the prediction performance of the developed robust model is evaluated in presence of data contamination. To further appraise the model validity and practicality, two case studies were carried out for each development. A simulated set of data with predefined characteristics to highlight the presence of outliers was used to demonstrate the robustness of the model. The advantages of this robust model are further illustrated via a set of real process data set from our industrial partners.

Preface

This thesis is an original work conducted by Anahita Sadeghian. The work presented in this thesis is based on the research carried out under the supervision of Dr. Biao Huang. It is in a paper-based format, and the contribution and details of each chapter/paper are discussed below.

Chapter 2 of this work is published as *Anahita Sadeghian, Biao Huang, 2016, "Robust probabilistic principal component analysis for process modeling subject to scaled mixture Gaussian noise", Computers & Chemical Engineering.*

Chapter 3 of this work is published as *Anahita Sadeghian, Ouyang Wu, Biao Huang, 2018, "Robust probabilistic principal component analysis based process modeling: Dealing with simultaneous contamination of both input and output data", Journal of Process Control.*

Chapter 4 of this work is in preparation as *Anahita Sadeghian, Nabil Magbool Jan, Ouyang Wu, Biao Huang, Robust Probabilistic Principal Component Analysis for Process Modeling Subject to Switching Scaled Mixture Gaussian Noise.*

Chapter 5, finally will conclude the results and provide some future directions for the interested readers.

A co-authored paper is also published as *Shabnam Sedghi, Anahita Sadeghian, Biao Huang, 2017, "Mixture semisupervised probabilistic principal component regression model with missing inputs", Computers & Chemical Engineering.*

To ROBUSTNESS.

In memory of the passengers on PS752

Acknowledgments

I cherished the entire process of mapping my doctoral journey and it would have not been this memorable without the care and support I received from multiple dimensions. Principally, it is my great pleasure to acknowledge, with sincere gratitude, Dr. Biao Huang's leadership, inspiring professionalism, continuous support and patience. My journey was packed with so many life lessons I learned from him that I will remember in my entire personal and professional life; and will hold dear to my heart.

I would like to thank the members of my defense committee, Drs. Jinfeng Liu, Prashant Mhaskar, Qing Zhao, and Zhan Shu for their helpful commentaries and inspirational discussions. I would also like to thank Dr. Vinay Prasad, who empowered the momentum for my final steps when I had forgotten my fighter hat among myriads of outlying life challenges I had to conquer. I extend my gratitude to my role model, Aris Espejo from Syncrude Canada Ltd., whom I learned a lot on professional communication and teamwork from, and also to my mentors from International Society of Automation, Edmonton branch, Andy Bahniuk from Shell Canada Ltd., Bob Bahniuk from Spartan Controls, and Aaron Boser. I express my special appreciation to my former supervisor at University of Tehran, Dr. Farhang Jalali-Farahani, for opening a research window on process modeling and uncertainty to me on the way to investigate my dream of an autonomous refinery. And here we stand in front of roads yet to be paved and journeys yet to be taken for materializing that exciting ambition.

I greatly appreciate the fruitful discussions I had with, and the feedback I received during my journey from Drs. Sirish L. Shah, Bhushan Gopaluni, Zhiqiang Ge, Zukui Li, Shima Khatibisepehr, Nima Sammaknejad, Elham Naghoosi, Mohammad Rashedi, Omid Namaki, Alireza Fatehi, Ouyang Wu, Nabil Magbool Jan, Yousef Alipouri, and Mengqi Feng.

During this adventure, I had the honour to be a part of the Computer Process Control (CPC) group and work with members from different cohorts through

years. I would like to thank, with no specific order, Shabnam Sedghi, Atefeh Daemi, Agustin Vicente, Seraphina Kwak, Rishik Ranjan, Dereje Tefera, Junyao Xie, Arun Senthil Sundaramoorthy, Drs. Swanand Khare and Jiusun Zeng, Ruben Gonzalez, Fadi Ibrahim, Aditya Tulsyan, Yuri Shardt, Rahul Raveendran, Khushaal Popli, Bardia Hassanzadeh, Ruomu Tan, Yanjun Ma, Lei Fan, Kirubakaran Velswamy, Hongtian Chen, Yue Cao, and so many others.

Financial support from Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta Innovative Technology Futures (AITF) is sincerely acknowledged. I would also like to acknowledge University of Alberta, particularly the department of Chemical and Materials Engineering for giving me the opportunity to pursue my doctoral degree.

Finally yet importantly, all and all is owed to my beloved family: my mother, *Mina*, My father, *Masoud* and my dear brother, *Behrouz*, whom wholeheartedly encouraged me to be authentic, stand up for myself and for my dreams. I am grateful for their presence, understanding, unconditional love and support in spite of being far for so long. To me, they signify *robustness* and are a symbol of resilience.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"The antifragile loves randomness and uncertainty, which also means— crucially—a love of errors, a certain class of errors. Antifragility has a singular property of allowing us to deal with the unknown, to do things without understanding them— and do them well."

Antifragile: Things That Gain from Disorder (N. N. Taleb, 2012)

1.1 Motivation and Inspirations

Typically, the success of industries depends on their production rate and product quality. Along with this, they are required to comply with safety and environmental considerations in their efforts to meet the set goals on production and quality [24]. Development of advanced process monitoring and control techniques is a key to achieving these objectives. Effective process monitoring and control strategies require the operational data to be recorded for the analysis of the problem.

In general, industrial plants are heavily instrumented with a large number of sensors and analyzers to provide the data needed for process control, fault diagnosis, and monitoring purposes. However, on-stream and wired measurements are not always accessible due to restricted availability and/or reliability of measurement techniques and devices. Even in cases where appropriate measuring devices are installed, some key process variables are still determined offline by laboratory sample analysis or by means of often arguable existing online analyzers. Such methods of process data acquisition are time consuming and introduce delays and discontinuities into their applications. Moreover, implementation of these methods might be expensive and might demand frequent and costly maintenance. One example of delayed data is the slowly-processed measurements obtained from gas chromatographs.

As a result of the particular issues introduced above, there has been sprouting interest in setting up predictive models that can provide frequent estimates for quality variables of interest. These predictive mathematical models are known as *soft sensors*, *inferential sensors* or *virtual sensors* that provide online (real-time) estimates of the key quality variables based on some other process records that are already available. By using these soft sensor models, we could enhance the entire system's reliability and accuracy to develop tighter control policies for the system under study [24].

Generally, three main modeling approaches exist for design of soft sensors as mentioned by [24]: mechanistic or physical modeling that is performed by first principles analysis, empirical or data-driven modeling, and grey-box modeling. In the first group of models, we use physical laws governing the plant, such as mass and energy balances. These models, if possible to develop, may be reliable for a long period of time and perform well. This is because they represent the essence of the system based on the

main governing equations. However, usually deriving these models is challenging and considerably ambitious due to the uncertainties deep-seated in the nature of process dynamics. Data-driven models are useful based on the fact that the plant records contain information about factors affecting the operation, which is buried in the *big data* [23]. The term big data is coined to depict records that have specific characteristics such as *Volume*, *Variety*, *Velocity*, *Value*, and *Veracity*-usually referred to as $5 - Vs$. Whilst the groundwork matures, different research communities add more Vs to the list. Recent addenda are *Variability* and *Visualization*. When relevant and high-quality data is collected, and if appropriate analysis and investigation is carried on, it is feasible to achieve reliable and accurate models through statistical analysis of the historical data. These two steps, namely, the collection of high quality and relevant data and accomplishment of an appropriate analysis, are quite substantial in soft sensor or any style of model development.

When collecting high-quality data, it is required that the effect of noise and disturbances in the recording procedure be minimized; in such manner, the data meet the requirements for the next steps, for instance, accurate model identification. Different filtering methods have been developed to address this challenge. Furthermore, proper analysis of the data enables the detection and management of the adverse effects associated with outliers, missing data, redundancy, low accuracy, and many other possible issues [24]. Collected data could then be processed in myriads of ways, and there are a number of choices that should be made. These choices comprise the selection of an appropriate model structure from a suitable model class (linear or nonlinear, static or dynamic, ...), and estimation of model parameters by applying a suitable identification technique. The last step, that has the same significance in the procedure is model validation. Further detailed discussions could be found in [67] for missing data; in [90, 19, 56, 70, 13] for outlier detection; in [59, 32] for pre-filtering; in [88, 90, 23, 58, 53, 1] for variable and model structure selection; in [9, 47] for model order selection; in [82, 81, 41] for model identification and in [12, 7, 8] for model validation.

Over the course of time, a vast area of research has been devoted to the above-mentioned steps to obtain an adequate model of the process under study. However, there are challenging issues associated with each step that make this area of research

open for further explorations. For instance, in relation to the quality of the data used in modeling, issues such as measurement noises, missing values, outliers, etc. might arise. Several ad hoc solutions to these potential issues are available. A more general, robust and well-established approach of dealing with these issues is yet to be developed. In the context of model identification, it is required to pay special attention to the modeling of hybrid systems. Hybrid systems experience discrete changes during their continuous operation. Examples of such systems are provided by [46] from polymers industries. In this field, modeling of hybrid systems is of interest since some production policies drive a single polymer manufacturing plant to undergo different processes with different operating conditions, to produce different grades of polymer. Therefore, to describe this switching act, that reflects the multiple-mode characteristic of the process, a multi-modal representation is required. The other application of modeling of multi-modal systems in the literature is to represent the behavior of nonlinear dynamic processes by approximating them with individual piecewise linear models.

Soft sensors have several advantages including: *no capital cost* is needed for development stage compared to that of hardware measuring devices, since they are *software* sensors. They can *work in parallel* with hardware sensors for process control and monitoring purposes. They also *provide online estimates* for the process variable measurements based on a previously recorded historical data. Moreover, by considering the historical data that may include some hidden information about the plant's operation, they can *add to the prior process knowledge* available from physical laws and first principles equations. In addition, it is *easy to implement* soft sensors on the readily available hardware, e.g. microcontrollers, and update them as the system behavior changes. Along with these advantages, they would *improve plant operation* by solving the previously mentioned issues faced in process control and monitoring. As a result, soft sensors assist industries by *increasing market success* and *profitability* along with reducing off-specification products over environmentally acceptable operating conditions [22, 54, 69, 49].

So far, the important role of measuring variables in a process has been introduced. One other common challenge in practice is high-dimensionality of the process data. Searching for the most relevant variables to the target variable or key performance

index is a solution which requires domain knowledge. An alternative to removing the less relevant variables is to make a solution out of a type of combinations of variables in order to reduce the dimension of the data set.

In this thesis, approaches to dealing with the issues mentioned in this section are reviewed and the development of a solution to these issues is discussed through upgrading and reformulation of the existing modeling approach in a robust fashion. The work also constitutes testing of the developed solution on existing industrial data.

1.2 Outlying or Atypical Observation

Outliers, outlying or atypical observations are those that seem unusual or extreme with respect to other observations, as well as the prior knowledge about the possible typical range of measured values [29]. A very simple application of detecting outliers is to monitor a specific feature in a production line, similar to monitoring credit card usage to prevent fraudulent use [33].

Nowadays, there is a common issue in different disciplines and their commercial applications: the concept of *Big Data*. Researchers indicated that almost all large datasets contain outliers [91]. Manual evaluation of the outliers is difficult or sometimes impossible. The outlying observations mostly represent a random error caused by hardware failure, operator's incorrect reading from devices, transmission issues or infrequent changes in the dynamic behaviour of the system.

Observation data is usually multivariate, and the methods used to detect outliers in this category of data sets require distance metrics such as Euclidean or Mahalanobis distances. The former only considers the location information, while the latter is more reliable as it considers the dependencies between the attributes by incorporating the covariance matrix in the calculation. The general idea of such methods, is to assign a scalar distance to each observation and to pinpoint the observations with a distance larger than a threshold, as outliers.

Another group of researchers provided a profound review on outlier detection algorithms [33]. In conventional multivariate modeling methods, such as least square or other methods with similar loss functions, a single outlying data point can have a huge impact on the outcoming model. This is why robust methods and outlier

detection are of considerable importance. Some basic methods, such as trimming and winsorizing [87], are introduced in the early literature that simply discard data. The drawback of such methods is that due to this simple omission the estimates might be biased. It is worthwhile to indicate that the method of dealing with outliers can depend on the application domain. For example, if the entry clerk causes a typographical error, the person responsible could be notified to correct the error and the outlier is then restored to a normal record. Alternatively, for outliers resulting from a hardware failure, the treatment might be their removal.

The same group of researchers categorized three fundamental approaches to outlier detection. The first approach is analogous to unsupervised clustering, as it only looks at the static data and pinpoints the most remote points as possible outliers. They portray the method to be similar to a batch processing system that requires all data to be available to start the detection. However, the second approach is analogous to supervised classification and models both normality and abnormality. The application of such approaches requires pre-labeled data which are tagged as normal and abnormal. The last category, known as novelty detection, models only the normality or in a very few number of cases, models abnormality [33]. This approach is believed to be analogous to a semi-supervised detection task; semi-supervised in the sense that in this approach, the normal class is learned but the algorithm learns to detect abnormality. This approach requires pre-classified data and it only learns one class, the normal ones. In dynamic cases, it tunes the model of normality to improve the fit as each new data point becomes available. This defines a so-called boundary of normality. They also provide an overview of different statistical methods from their early univariate application to a one-dimensional data set.

Grubbs' method as in [31], is an example of univariate outlier detection methods. The basic idea of this method is to first, calculate the difference between the mean value of the attributes and the query value divided by standard deviation of the attribute. The mean and standard deviation are calculated by using all attribute values including the query value. Then, this distance is compared to a 1 to 5% of significance level. Previously mentioned research team in [33], defined a statistical class of methods of outlier detection such as proximity-based techniques (like k-nearest neighbor, k-NN), parametric (e.g. principal component analysis (PCA)),

non-parametric (e.g. Dasgupta and Forrest method) and semi-parametric methods (e.g. Gaussian mixture models (GMM)). Proximity-based methods are inefficient in dealing with high-dimensional data sets due to the computational complexity of calculating all vector distances. On the other hand, parametric methods are known to be suitable for large data sets since they allow the model to be quickly evaluated when new instances are received. The main shortcoming of these methods is that their applicability is limited by the availability of a pre-specified distribution model of the data. Such *a priori* knowledge is not often available since many data sets simply do not follow a specific and known distribution model. Non-parametric methods are more flexible and self-governing. Finally, semi-parametric methods apply different local kernel models instead of a single distribution model. This would result in a combination of speed and complexity growth of parametric methods [33].

Many robust approaches have been developed as an alternative to previously addressed naive methods. Another research indicated that appreciable attention has been given to replacing the non-robust least squares estimates with robust alternatives [21]. A number of methods for robust regression have been proposed in the literature. Examples include: M-estimates [35], the Stahel-Donoho estimate [84, 18], least median of squares [72] and S-estimates [15, 60, 21]. A review of all these methods and some others can be found in [66]. This review mostly attempts to improve the traditional multivariate regression methods, such as principal component regression (PCR) and partial least squares (PLS).

In addition to *Big Data*, there is another issue for the statistical methods to be aware of and that is named as *Curse of Dimensionality* by [5]. This publication states that the number of data and also the number of computations required for a predictive model grows exponentially with the dimensionality of the feature vectors; and that this increases the data processing time. As the dimensionality increases the data points are spread over a larger volume and become less dense. This makes the recognition of the spreading convex hull of the data distribution more challenging. PCR and PLS reduce the number of dimensions in data but they still do not provide the probabilistic model underlying the data generation (generative probabilistic model) to determine how well the new data will fit within the model. Research has been done on this subject to develop robust models in the framework of latent variable modeling.

There is still opportunities to improve this robustness against outliers and missing data. The outliers can be modeled in a probabilistic representation instead of simply being discarded. To model the outliers, different distributions could be used; some of these distributions will be discussed in the next sections. The use of these concepts is especially needed in process industries to develop more robust models. Development of more reliable models would result in more effective process monitoring and control policies. Consequently, the plant operation becomes safer and more profitable. Parameters of such models, could be estimated through an expectation-maximization algorithm.

One of the directions of this thesis is to address the presence of outliers by assuming different patterns of distributions for them and incorporating them in the main model of the system, instead of considering their removal or substitution, to develop a robust model. Considering different noise models to represent the outliers will be further discussed in Chapters 2, 3, and 4.

1.3 Latent Variable Models

In addition to the potential outlier problem, high-dimensionality of data space is another challenging issue in modeling applications. Latent variables or hidden variables, are generally known as variables that are not observed directly but are inferred from the directly-measured ones instead. A model that investigates the dependence of a set of observed variables on a set of latent variables in a statistical framework is a latent variable model [20]. We will further discuss the functionality of latent variable models on high-dimensional data, and also will propose the application of our developed method on such data sets.

A general latent variable could be formulated as:

$$p(x) = \int p(x|t)h(t)dt \tag{1.1}$$

where $x = [x_1, ..., x_M]^T$ is the observed variable vector and $t = [t_1, ..., t_P]^T$ represents a vector of latent variables. The dimension of the latent space, P, is usually smaller than the number of variables in the observation set, M.

Essentially all latent variable models assume that observations have a joint probability distribution conditional on the latents, that is $p(x|t)$. The other point is that

latent variable models lean on a key assumption of conditional independence. That is, the observable variables are independent of each other, given the values of the latent variables. Conditional and marginal density functions, p and h, could be inferred from the known or assumed density of x, based on some assumptions, to describe the probabilistic dependency between the observed and latent variables. The interdependence between the observable variables is the result of their common dependence on the latent variables. Therefore, when the latent variables are fixed, the behavior of the observable variables is random in essence. So we have:

$$p(x) = \int h(t)(\prod_{i=1}^{M} p(x_i|t))dt \tag{1.2}$$

Different classes of latent variable models are generated by different assumptions about latent variables distribution. The mostly known class is *factor analysis* which was developed by psychologists at first. Authors in [21], review the recent research on multivariate statistical techniques that are related to latent variable models as independent component analysis (ICA), Kalman filter model, and hidden Markov models (HMMs).

1.3.1 PCA-based Regression Model

To start the data mining process, there are some preparatory steps to be taken to ensure a correct interpretation. The compulsory step is pre-processing the data by dealing with outliers after their detection and also by handling the missing part(s) of the data. The next optional step, is data reduction to only consider the informative parts of the *Big Data*. Beside reducing the number of records by downsampling, dimension reduction is one of the common data reduction approaches. Principal component analysis is a suitable method for reducing the dimension of a high-dimensional data set by identifying correlated features in the data and then projecting it onto a lower dimensional subspace [33]. This method, known as an ideal method to select a subset of features for use in modeling, could be a preprocessing step for the methods that otherwise suffer from the curse of dimensionality.

1.3.2 Deterministic PCA

Let us consider a set of input and output variables, $X \in \Re^{n \times m}$ and $Y \in \Re^{n \times r}$, where n is the number of data samples or the observation length, and m and r are the number of input and output variables, respectively. One can perform dimension reduction on the data set X. PCA is considered only for input data and in PCR regression is done between the input and output data after performing the dimension reduction. To be consistent with the notations used in this thesis, we call PCR as a deterministic PCA based regression model.

The PCA based regression model structure is given as follows:

$$X = TP^T + E \tag{1.3a}$$

$$Y = TC^T + F \tag{1.3b}$$

where $P \in \Re^{m \times q}$ is the loading matrix, $T \in \Re^{n \times q}$ is the principal components matrix, $C \in \Re^{r \times q}$ is the regression matrix and scalar q is the number of selected principal components (i.e. the latent space dimension). E and F are the residual matrices with appropriate dimensions [27]. The dimension of the latent space is usually less than that of the original observation space.

PCA could be seen as a restricted version of factor analysis (FA) with respect to assumptions. The underlying assumptions of FA are randomly distributed latent variables, as well as random noise (or observation) with unit variance, and zero-centered latent variables. When we restrict FA to have isotropic error models each with variance σ^2 and force its latent variables to be deterministic, we build a PCA model. The goal is to find the loading matrix and also to determine the noise variance.

PCA is a convenient method, when graphical representations are not available or suitable in the representation and interpretation of data. It is a useful way to check the quality of the data with respect to existence of any clusters. In the context of process engineering, it is useful when univariate charts cannot sufficiently show the effect of process variables on the system and also it is useful for constructing a quality index in process monitoring.

The crux, to solve the unknowns, is to look at the data to see if PCA is needed and applicable at first hand. PCA would be applicable if the data has any insignificant

components. This evaluation is done by checking the rank of the covariance matrix of the observation data. If the covariance matrix is not full-ranked we cannot claim that the variables of the data matrix are independent, therefore a linear relation exists between them. In this case, matrix T is defined as $T = [t_1, ..., t_q] \in \Re^{n \times q}$,

- where $t_1 \in \Re^{n \times 1}$ is the first principal component, that is a linear combination of the process variables such that the maximum variance happens for the covariance matrix of the data and $\|p_1\| = 1$.
- and $t_2 \in \Re^{n \times 1}$ is the second principal component, that is a linear combination of the process variables such that the next biggest variance happens for the covariance matrix of the data and $\|p_2\| = 1$ and $t_1 \perp t_2$.
- $t_q \in \Re^{n \times 1}$ is the q^{th} (last) principal component, that is a linear combination of the process variables such that the last biggest value of variance happens for the covariance matrix (uncertainty is presented here) of the data and $\|p_q\| = 1$ and all $t_{i=1:q}$ are perpendicular to each other.

Principal components are uncorrelated and orthogonal, each with unit length to build an orthonormal coordinate system for the reduced dimension space. By finding $p_{i=1:q}$ that satisfy the above mentioned conditions, the loading matrix is determined. It is proven that the $p_{i=1:q}$, that are vectors of the loading matrix P, are eigenvectors of the covariance matrix of X. Meanwhile, the eigenvalues of the covariance matrix are equivalent to the variance of t_i. This indicates that the extension of the data over the corresponding direction, and this is the reason that maximizing the variance is desired [40]. One obvious weakness of this model, is the uncertainty situated in the number of principal components chosen for the reduced space. There are many methods mentioned in the relevant literature, such as Scree plot, to estimate this number via some thresholds. By trial and error, one can get an appropriate number of latent variables to capture the data set at its best fit. In the following chapters, we will investigate the use of Bayesian methods and maximum a posteriori estimation to reach the number of reduced dimensions while determining the loadings.

1.3.3 PPCA-based Regression Model

Principal component analysis (PCA) is widely used in dimension reduction in high dimensional data sets to ease the analysis. Researchers have combined the PCA method with a maximum likelihood solution for a generative latent variable model known as probabilistic PCA or PPCA [86]. PPCA is a more complex version of latent variable model, but has some distinct advantages. By using PPCA, there is the possibility of developing mixture models that could show better performance on nonlinear processes. PPCA is a probabilistic model that is more general and applicable. PPCA can be used in problems with maximum likelihood formulation, and since it is a probabilistic model, it also allows for the deployment of Bayesian methods when needed. In [27], solving PPCA was extended to a mixture PPCA (MPPCA) where multiple PPCA based models were considered; each PPCA based model was determined individually and then combined with appropriate weights to form the global model. In this thesis, however, we will consider a single unique model that is optimized simultaneously in the presence of both regular and outlying noises. A discussion on MPPCA based process model and the issue of missing observations has been discussed in [79].

In a PPCA based model, given the input data $X = [x_1, x_2, ..., x_n]^T \in \Re^{n\times m}$ and the output data $Y = [y_1, y_2, ..., y_n]^T \in \Re^{n\times r}$, the model is derived via the following generative equations [27]:

$$x = Pt + e \tag{1.4a}$$

$$y = Ct + f \tag{1.4b}$$

where $P \in \Re^{m\times q}$ and $C \in \Re^{r\times q}$ are the weighting matrices, $t \in \Re^{q\times 1}$ is the latent variable vector, and $e \in \Re^{m\times 1}$ and $f \in \Re^{r\times 1}$ are measurement noises for the input and output data, respectively. For this to be a probabilistic model, there are some assumptions for the random variables involved in the model. In this general model, it is assumed that the latent variables and the measurement noise both follow a Gaussian distribution as $t \sim \mathcal{N}(0, I)$, $e \sim \mathcal{N}(0, \sigma_x^2 I)$, and $f \sim \mathcal{N}(0, \sigma_y^2 I)$, where I is the identity matrix, and σ_x^2 and σ_y^2 are noise variances of the input and output variables, respectively. To estimate the optimal unknown parameters of the model,

loading matrices and the variances, maximization of the likelihood of the data is performed. EM algorithm is usually used for this estimation because of the presence of hidden variables [27].

1.4 Thesis Outline

This thesis is organized as a collection of three primary chapters and a chapter on conclusions, aside from this introductory chapter. The upcoming chapters are organized as follows:

Chapter 2: Robust Probabilistic Principal Component Analysis for Process Modeling Subject to Scaled Mixture Gaussian Noise

In this chapter, unlike the conventional noise assumption, a mixture noise model with a contaminated Gaussian distribution is adopted for probabilistic modeling to diminish the adverse effect of outliers, which usually occur due to irregular process disturbances, instrumentation failures or transmission problems. This is done by downweighing the effect of the noise component which accounts for contamination on the output prediction. Outliers are common in process industries; therefore, handling this issue is of practical importance. When compared with the conventional PPCA based regression model, the prediction performance of the developed robust probabilistic regression model is improved in presence of data contamination. To evaluate the model performance, two case studies were carried out in this chapter. A simulated set of data with specified characteristics that highlight the presence of outliers was used to demonstrate the robustness of the developed model. The advantages of this robust model are further illustrated via a set of real industrial process data.

Chapter 3: Robust Probabilistic Principal Component Analysis Based Process Modeling: Dealing with Simultaneous Contamination of Both Input and Output Data

In this chapter, possible *location* outliers are considered for both input and output data in contrast to the traditional robust algorithms that have focused on output outliers only, such as the scale outliers that are discussed in Chapter 2. Probabilistic principal component analysis based regression is used for the predictive model in this chapter and Expectation Maximization algorithm is applied to solve a complex robust estimation problem. Finally, the performance of the developed robust predictive

model is evaluated by simulated and industrial case studies. This chapter is a generalization to the traditional robust probabilistic principal component analysis based regression modeling work which considered a different type of outliers that occur in the output only.

Chapter 4: Robust Probabilistic Principal Component Analysis for Process Modeling Subject to Switching Scaled Mixture Gaussian Noise

This chapter considers another common behaviour of measurement noise patterns. Robust PPCA based model is developed under the assumption of switching noise model. This more broad glimpse of the behaviour of a measurement noise is a closer reconstruction of the reality of complex chemical processes. The dynamic behaviour of the noise is designed to be switching between two states; and the two states are sourced from different Gaussian distributions, one representing regular noise and the other representing outliers. Similar to Chapter 2, a scaled Gaussian mixture model is studied. Here as in the previous chapters, the derived model is evaluated under a simulated case study and then is used in a real industrial application. Results confirm the robustness of this approach.

Chapter 5: Conclusions and Future Directions

Final chapter is dedicated to summarizing the contributions and providing an overall view of the results at a glance.

It is noted that the thesis is based on the paper-format and follows the rules set by Faculty of Graduate Studies and Research at University of Alberta. Therefore, to maintain the paper-format and ensure completeness, each chapter is self-contained. Some parts of the chapters might have overlaps, especially in the fundamentals section. The overlap was not removed in order to provide a smooth flow of the thesis to the readers and ease the understanding for the material.

1.5 Contributions

The main objective of this work is unfolding some process data quality matters and developing robust models through an improved formulation of a PPCA based process model. Uncertainty, high-dimensionality, and outlying data points are the central

focus of this thesis. I was responsible for conceptualization, literature review, data curation, formal analysis, and investigation, methodology, simulation, visualization, validation, and writing of the original manuscript.

As Chapter 2 elaborates, the first step was to develop a robust PPCA based model through the consideration of a scaled Gaussian mixture noise for the process measurements.

Next contribution, as elaborated in Chapter 3, is the consideration of a more broad noise formulation. A combination Gaussian mixture noise model was used for both input and output data to formulate a robust PPCA based model. I was responsible for conceptualization, data curation, formal analysis, and investigation, methodology, visualization, validation, and the writing of the original manuscript. My coauthor for the corresponding published paper, Ouyang Wu, contributed to the simulation, review, and editing of the manuscript.

For the next contribution, a more realistic case for the noise model was considered that contains its dynamic behaviour. As Chapter 4 elaborates, the robust PPCA based process model developed in Chapter 2 was reformulated with a switching noise model to mimic the dynamic nature of process noise. I was responsible for conceptualization, data curation, formal analysis, and investigation, methodology, visualization, validation, and the writing of the original manuscript. My coauthors for the corresponding published paper, Ouyang Wu and Nabil Magbool Jan, contributed to conceptualization, reviewing the formulae derivations and the simulation. Nabil Magbool Jan also participated in creating an initial draft of the manuscript.

To make this research journey happen, conceptualization, resources, software, project supervision, submission review, and funding acquisitions were provided by Dr. Biao Huang.

Chapter 2

Robust Probabilistic Principal Component Analysis for Process Modeling Subject to Scaled Mixture Gaussian Noise*

"I have two ways of learning from history: from the past, by reading the elders; and from the future, thanks to my Monte Carlo toy."

Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets (N. N. Taleb, 2001)

*A version of this chapter is published as Anahita Sadeghian, Biao Huang, 2016, "Robust probabilistic principal component analysis for process modeling subject to scaled mixture Gaussian noise", Computers & Chemical Engineering

2.1 Introduction

The profitability of many industries depends on their production throughput and product quality. Moreover, they are required to comply with safety and environmental considerations in their efforts to meet the goals they have set on production and quality [24]. Development of advanced process monitoring and control techniques is a key to achieving these objectives. This development stage requires the operational data to be recorded for the analysis of the problem.

In general, industrial plants are heavily instrumented with a large number of sensors and analyzers to provide the data needed for process control and monitoring purposes. However, online and fast-rate measurements are not always available due to restricted availability and/or reliability of measurement techniques and devices. Even in cases where appropriate measuring devices are available, some key process variables are still determined offline by laboratory sample analysis or by means of often unreliable online analyzers. Such methods of process data acquisition are time consuming and introduce delays and discontinuities into their application.

As a result of the issues stated above, there has been a growing interest in setting up predictive models that can provide frequent estimates for quality variables of interest. These predictive mathematical models are known as *soft sensors*, *inferential sensors* or *virtual sensors* that provide online (real-time) estimates of the key process variables based on some other process records that are already available. By using these models, we can enhance the entire system's reliability and accuracy to develop tighter control policies for the system under study [24].

Generally, three main modeling approaches exist for the design of soft sensors as mentioned by [24]: mechanistic or physical modeling that is performed by first principles analysis, data-driven modeling, and grey-box modeling. In the first group of models, we use physical laws governing the plant, such as mass and energy balances. These models, if possible to develop, may be reliable for a longer period of time. This is because they represent the essence of the system based on the main governing equations. However, usually deriving these models is difficult due to the uncertainties deep-seated in the nature of the process dynamics. Data-driven models are useful based on the fact that the plant records contain information about factors affecting

the operation, which is buried in the data. However, the relevancy and quality of data affects the model performance, along with other factors such as the data analysis method used. For high fidelity modeling, it is required to reduce the effect of noise and disturbances in the recording of the data. This might not be possible at all times. Although different filtering methods have been developed to address this issue, proper analysis of the data enables one to detect and deal with the effect of outliers, missing data, redundancy, low accuracy, and many other possible issues [24]. Several ad hoc solutions to these potential issues are available in the literature [42, 33, 10].

Outliers, or outlying observations are one of the main causes of low quality data [44, 58]. These points are usually the data points that do not seat in the range of most of data points and are randomly off the statistics of the majority. The outliers could be result of instrument failure, field operator and/or laboratory technician errors, and other factors [45, 13, 96, 52].

Addressing all data issues together in one step is a highly complicated task. This chapter addresses the problem of outliers and develops a robust predictive model which is invulnerable to the presence of outliers. There are different types of outlying observations as discussed in [45], and the focus of this article would be on scaled outliers. This contribution is done in a probabilistic modeling framework under a different noise formulation.

In the next sections, the basics of probabilistic principal component analysis (PPCA) and PPCA based regression models are reviewed. Then, the robust model is formulated based on one of the most common outlier models. The developed model is next solved with the use of an Expectation Maximization (EM) algorithm. Finally, to evaluate the developed robust model two case studies are discussed.

2.2 Fundamentals

To start the process of working with data, there are some preparatory steps to be taken to ensure a correct interpretation. The compulsory step is preprocessing of the data by dealing with outliers after their detection and also by handling the missing data. Outliers, are those observations that seem unusual or extreme with respect to other observations. They are also extreme with respect to the prior knowledge

about the possible typical range of measured values [29]. A very simple application of detecting outliers is to monitor a specific feature in a production line, or monitoring a credit card usage to prevent fraudulent use [33]. The next recommended step, is data reduction to only consider the informative parts of the *Big Data*. Beside reducing the number of records by downsampling, dimension reduction is one of the common data reduction approaches. Principal component analysis is a suitable method for reducing the dimension of a high dimensional data set by identifying correlated features in the data and then projecting them onto a lower dimensional subspace [33]. This method, which has been extensively discussed in Section 1.3, is known as an ideal method to select a subset of features for use in modeling, could be a pre-processing step for the methods that otherwise suffer from the curse of dimensionality.

The preliminaries of the probabilistic modeling approach have been previously discussed in 1.3. The fundamentals of outliers have been previously reviewed in 1.2.

In this chapter, under the framework of probabilistic principal component analysis (PPCA)-based regression, we aim to develop a robust PPCA-based regression model to capture the output outliers by applying appropriate distributions for them and cautiously incorporating them into the main model instead of considering the removal or substitution of a whole sample point. Principles of how to consider different noise models in comparison with that of a regular PPCA-based model will be discussed in Section 2.4.

2.3 Problem Statement

The authors in [45] classify outliers to two general classes, scale and location outliers. These classes are result of a shift in variability and/or location of the measurement noise. They consider scale outliers to be generated from process measurements which are violating the physical limitations of a process operation unit, while stating that the other class is caused by process measurements that do not conform to the technological extent of the measuring device. The second class under the above-mentioned circumstances usually generates a symmetric location outlier. In addition, the measurements which are obtained from a jammed instrumentation device could be considered as an asymmetric location outlier. In this article, one of the most common

outlier categories, scaled outlier, is considered for the output noise model of a PPCA based regression model to advance its predictive role in presence of contaminated data.

2.4 Robust Model Development

In this section a data-driven generative regression model based on probabilistic principal component analysis is formulated which is robust to scaled mixture type of outliers. As seen in Section 2.2, probabilistic principal component regression model is formulated by [99, 86] as:

$$\begin{cases} x_i = Pt_i + \mu_x + e_i \\ y_i = Ct_i + \mu_y + f_i \end{cases} \tag{2.1}$$

In this generative model, $t_i \in \Re^{q\times 1}$ is latent variable vector, $x_i \in \Re^{m\times 1}$ and $y_i \in \Re^{r\times 1}$ are observed as input and output data vectors, respectively. μ stands for mean vectors. In a matrix form, the model could be written as:

$$\begin{cases} X = TP^T + M_x + E \\ Y = TC^T + M_y + F \end{cases} \tag{2.2}$$

where, $X = [x_1, \cdots, x_n]^T \in \Re^{n\times m}$, $Y = [y_1, \cdots, y_n]^T \in \Re^{n\times r}$ and $T = [t_1, \cdots, t_n]^T \in \Re^{n\times q}$. In Section 1.3.3, the assumptions for the input/output noise models were reviewed. In order for the model to be robust to outlying observations, we consider a more general output noise model.

The noise is assumed to follow a mixture distribution which consists of two Gaussian components. One component of this noise model is a Gaussian distribution with a specific mean and variance, whereas the other one is a Gaussian with same mean as the first component but a different variance (often taking an extreme value). Variance of the latter one, which we call as contaminating part, is inflated with respect to that of the first component by a factor of ρ^{-1}, where $\rho \in (0, 1]$. In other words, this mixture output noise model is assumed to incorporate a scaled contaminating portion to acknowledge the outliers which originate from measurements violating the physical limitation of a process unit, such as a very large flow rate.

$$\begin{aligned} e_i &\sim \mathcal{N}(0, \sigma_x^2 I) \\ f_i &\sim (1-\delta_y)\,\mathcal{N}(0, \sigma_y^2 I) + \delta_y\,\mathcal{N}(0, \rho_y^{-1}\sigma_y^2 I) \end{aligned} \tag{2.3}$$

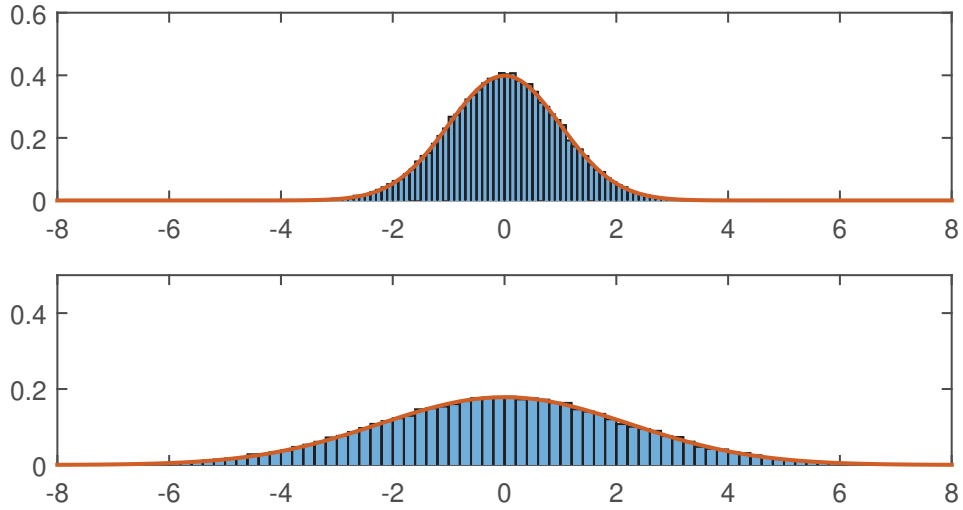


Figure 2.1: A regular standard Gaussian distribution (top) and its scaled counterpart with $\rho = 0.2$ (bottom)

A regular Gaussian distribution is shown in Figure 2.1, along with a superimposed scaled Gaussian distribution which is responsible to account for the outliers. According to its fat-tailed shape and wider spread, this scaled distribution is capable of considering the noises which are located farther from the mean. Thus, the noise models would become as that in (2.3); such a noise model will help downweighing the effect of outliers in parameter estimation.

A vector of variables $Q_Y = [q_{y_1}, \cdots, q_{y_n}]^T \in \Re^{n\times 1}$, $(q_{y_i} \in \Re^{1\times 1})$ is introduced as a vector of binary indicators which implies identity of each sample point. When this indicator $q_{y_i} = 1$, the output noise f_i is distributed as $\mathcal{N}(0, \sigma_y^2)$; and when $q_{y_i} = \rho_y$, f_i is distributed as $\mathcal{N}(0, \rho_y^{-1}\sigma_y^2)$. [45] represents a Bernoulli distribution for q_{y_i} as:

$$P(q_{y_i}|\delta_y, \rho_y) = \delta_y^{1-\frac{q_{y_i}-\rho_y}{1-q_{y_i}\rho_y}}(1-\delta_y)^{\frac{q_{y_i}-\rho_y}{1-q_{y_i}\rho_y}}, \tag{2.4}$$

where δ_y is the probability that the output observation y_i is generated by the contaminated Gaussian noise component; ρ_y is the variance inflation factor for that component which could vary between 0 and 1; the smaller ρ_y is, the bigger the magnitude of noise (that leads to an outlying observation) would be.

RPPCA regression model parameters consist of P, C, and the hyperparameters of noise terms e and f which are the noise variances σ_x^2, σ_y^2 and the output noise variance

inflation factor ρ_y, as well as one other parameter δ_y which adjusts the proportion of the two Gaussian components in the mixture noise. Assumptions which were discussed for PPCA regression model hold for this model as well. Establishment of this problem will be delineated in Section 2.4.2.

2.4.1 Expectation Maximization Algorithm

In a broad variety of problems with incomplete data, EM algorithm is applied to iteratively solve a maximum likelihood problem. This is an alternative solution when the maximum likelihood problem is hard to tract due to the absence of some data or variables. The incompleteness of data could arise from different evident factors, namely missing observation data, truncated distributions, censored or grouped observations, or from statistical models such as random effects, mixtures, convolution and latent variable structures [62].

As the name suggests, the algorithm consists of two main steps in its iterations. First, an expectation step which is called E-step for short; second, a maximization step which is shortened as M-step. These two steps work towards formulating a complete-data problem out of the given incomplete-data problem. Thus, a succession of optimizations is done on an augmented or completed data instead of performing one complex optimization task [16]. This is accomplished by considering hidden variables. Repetition of these two steps is stopped when a convergence criterion is satisfied; this criterion could be relative difference of the parameters or the complete-data likelihood function of two successive iterations. The latter is known to be faster; nevertheless, [94] in particular devoted effort to discuss whether the convergence of likelihood can automatically involve the convergence of estimated parameters.

Numerical stability is among the properties that make this algorithm to be favorable, such that in each iteration the likelihood function is increased until the convergence occurs. Easy implementation, low computational cost per iteration and the ability to provide estimates for missing data are other beneficial properties [62]. However, this algorithm might require the user to start from different initial sets of parameters and to use Monte Carlo simulations to achieve convergence in the case of multiple local optimum points (multimodal complete-data likelihood function). More details on EM algorithm will be discussed in Section 3.2.2.

2.4.2 Parameter Estimation

For the RPPCA regression model to be developed, the available observations are X and Y. Unknown information is:

$$\{P, C, \sigma_x^2, \sigma_y^2, \mu_x, \mu_y, \delta_y, \rho_y, T, Q_Y\},$$

from which $\{t_i\}_{i=1}^n$ and $\{q_{y_i}\}_{i=1}^n$ are treated as latent variables of the model and $\theta = \{P, C, \sigma_x^2, \sigma_y^2, \mu_x, \mu_y, \delta_y, \rho_y\}$ is a set of parameters of the model to be estimated.

The EM algorithm is employed in this formulation to estimate the parameters of the model, since there are some unknowns to be resolved which could be treated as hidden variable in this approach. We need to construct the $Q - function$ based on (2.5) first, and then maximize that function to obtain the estimation of parameters during an iterative procedure.

$$\mathbb{Q} = \underset{T,Q_Y|X,Y,\theta^{old}}{\mathbb{E}} (logP(\underbrace{\overbrace{X,Y}^{Observed}, \overbrace{T,Q_Y}^{Hidden}}_{CompleteData}|\theta)), \tag{2.5}$$

where $P(X, Y, T, Q_Y|\theta)$ is the complete data likelihood and is obtained as described in (2.6). The noise of the input and output is assumed to be independent and identically distributed. The same assumption holds for the latent variable and the sample indicators.

$$\begin{aligned} &P(X, Y, T, Q_Y|\theta) = P(X, Y|T, Q_Y, \theta)P(T, Q_Y|\theta) \\ &= P(X|T, Q_Y, \theta)P(Y|T, Q_Y, \theta)P(T|\theta)P(Q_Y|\theta) \\ &= \textstyle\prod_{i=1}^n P(x_i|t_i, q_{y_i}, \theta)P(y_i|t_i, q_{y_i}, \theta)P(t_i|\theta)P(q_{y_i}|\theta) \end{aligned} \tag{2.6}$$

Therefore,

$$\begin{aligned} \mathbb{Q} &= \mathbb{E}_{T,Q_Y|X,Y,\theta^{old}} \Big(logP(X, Y, T, Q_Y|\theta) \Big) \\ &= \mathbb{E}_{T,Q_Y|X,Y,\theta^{old}} \Big(\textstyle\sum_{i=1}^n logP(x_i|t_i, q_{y_i}, \theta) \\ &\qquad + \textstyle\sum_{i=1}^n logP(y_i|t_i, q_{y_i}, \theta) \\ &\qquad + \textstyle\sum_{i=1}^n logP(t_i|\theta) \\ &\qquad + \textstyle\sum_{i=1}^n logP(q_{y_i}|\theta) \Big) \\ &= \mathbb{E}_{T,Q_Y|X,Y,\theta^{old}} \Big(\textcircled{\text{I}} + \textcircled{\text{II}} + \textcircled{\text{III}} + \textcircled{\text{IV}} \Big) \end{aligned} \tag{2.7}$$

Now, the conditional expectation is taken based on the definition of expectation as in (2.8). Note that the variable T is continuous unlike Q_Y which is discrete. Thus, integration over the indicator variables would transfer to summation.

$$\begin{aligned}\mathbb{Q} &= \mathop{\mathbb{E}}_{T,Q_Y|X,Y,\theta^{old}}(function(x_i, y_i, t_i, q_{y_i}, \theta)) \\ &= \int_{t_i} \sum_{q_{y_i}} (function(x_i, y_i, t_i, q_{y_i}, \theta)) P(t_i, q_{y_i}|x_i, y_i, \theta^{old})\, dt_i \end{aligned} \tag{2.8}$$

Thus, in the RPPCA regression model with Gaussian-mixture output noise model the $Q - function$ would be as:

$$\begin{aligned}\mathbb{Q} &= \mathop{\mathbb{E}}_{T,Q_Y|X,Y,\theta^{old}}(logP(X, Y, T, Q_y|\theta)) \\ &= \int_{t_i} \sum_{q_{y_i}} \sum_{i=1}^{n} logP(x_i|t_i, q_{y_i}, \theta) \times P(t_i, q_{y_i}|x_i, y_i, \theta^{old})\, dt_i \\ &\quad + \int_{t_i} \sum_{q_{y_i}} \sum_{i=1}^{n} P(y_i|t_i, q_{y_i}, \theta) \times P(t_i, q_{y_i}|x_i, y_i, \theta^{old})\, dt_i \\ &\quad + \int_{t_i} \sum_{q_{y_i}} \sum_{i=1}^{n} P(t_i|\theta) \times P(t_i, q_{y_i}|x_i, y_i, \theta^{old})\, dt_i \\ &\quad + \int_{t_i} \sum_{q_{y_i}} \sum_{i=1}^{n} P(q_{y_i}|\theta) \times P(t_i, q_{y_i}|x_i, y_i, \theta^{old})\, dt_i \\ &\triangleq \mathbb{Q}_1 + \mathbb{Q}_2 + \mathbb{Q}_3 + \mathbb{Q}_4 \end{aligned} \tag{2.9}$$

As in the complete log-likelihood expansion (2.6), the minor terms are based on known indicator values. Indicators are binary discrete variables and the minor likelihood terms are being separately defined for each value of indicator. This is done by using *chain rule of probability* for joint posterior probability of hidden variables as in (2.10).

$$P(t_i, q_{y_i}|x_i, y_i, \theta^{old}) = P(t_i|x_i, y_i, q_{y_i}, \theta^{old}) P(q_{y_i}|x_i, y_i, \theta^{old}) \tag{2.10}$$

So, (2.9) would be expanded to (2.15), as detailed below, where $\mathbb{Q}_1^{first}$ indicates the first term of $Q - function$, $\mathbb{Q}$, when the indicator $q_{y_i} = 1$, and $\mathbb{Q}_1^{second}$ indicates the second term of $Q - function$ when the indicator $q_{y_i} = \rho_y$ (i.e. second case. Simply

follow $\mathbb{Q}^{case}_{term}$). Note that $\mathbb{Q}^{first}_{i=1,\dots,4} + \mathbb{Q}^{second}_{i=1,\dots,4}$ will equal to $\mathbb{Q}_{i=1,\dots,4}$ as defined in (2.9). Each of terms in (2.15) can be further expanded as shown in (2.16) to (2.23). The probability distribution terms in the complete data likelihood function could be found from (2.1) by recalling the model assumptions.

$$P(x_i|t_i, q_{y_i}, \theta) \sim \begin{cases} \mathcal{N}(Pt_i + \mu_x, \sigma_x^2 I) & , q_{y_i} = 1 \\ \mathcal{N}(Pt_i + \mu_x, \sigma_x^2 I) & , q_{y_i} = \rho_y \end{cases} \tag{2.11}$$

Same procedure occurs for the other three terms:

$$P(y_i|t_i, q_{y_i}, \theta) \sim \begin{cases} \mathcal{N}(Ct_i + \mu_y, \sigma_y^2 I) & , q_{y_i} = 1 \\ \mathcal{N}(Ct_i + \mu_y, \rho_y^{-1}\sigma_y^2 I) & , q_{y_i} = \rho_y \end{cases} \tag{2.12}$$

$$P(t_i|\theta) \sim \mathcal{N}(0, I) \tag{2.13}$$

$$P(q_{y_i}|\theta) \sim \mathcal{B}(1, 1 - \delta_y) \tag{2.14}$$

Based on these distributions and their logarithm, terms of (2.15) would be as in (2.16) to (2.23).

$$
\begin{aligned}
\mathbb{Q} &= \sum_{i=1}^{n} \int_{t_i} logP(x_i|t_i, q_{y_i} = 1, \theta) \times P(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) P(q_{y_i} = 1|x_i, y_i, \theta^{old})\ dt_i \\
&+ \sum_{i=1}^{n} \int_{t_i} logP(y_i|t_i, q_{y_i} = 1, \theta) \times P(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) P(q_{y_i} = 1|x_i, y_i, \theta^{old})\ dt_i \\
&+ \sum_{i=1}^{n} \int_{t_i} logP(t_i|\theta) \times P(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) P(q_{y_i} = 1|x_i, y_i, \theta^{old})\ dt_i \\
&+ \sum_{i=1}^{n} \int_{t_i} logP(q_{y_i} = 1|\theta) \times P(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) P(q_{y_i} = 1|x_i, y_i, \theta^{old})\ dt_i \\
&+ \sum_{i=1}^{n} \int_{t_i} logP(x_i|t_i, q_{y_i} = \rho_y, \theta) \times P(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) P(q_{y_i} = \rho_y|x_i, y_i, \theta^{old})\ dt_i \\
&+ \sum_{i=1}^{n} \int_{t_i} logP(y_i|t_i, q_{y_i} = \rho_y, \theta) \times P(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) P(q_{y_i} = \rho_y|x_i, y_i, \theta^{old})\ dt_i \\
&+ \sum_{i=1}^{n} \int_{t_i} logP(t_i|\theta) \times P(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) P(q_{y_i} = \rho_y|x_i, y_i, \theta^{old})\ dt_i \\
&+ \sum_{i=1}^{n} \int_{t_i} logP(q_{y_i} = \rho_y|\theta) \times P(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) P(q_{y_i} = \rho_y|x_i, y_i, \theta^{old})\ dt_i \\
&\triangleq \mathbb{Q}_1^{first} + \mathbb{Q}_2^{first} + \mathbb{Q}_3^{first} + \mathbb{Q}_4^{first} + \mathbb{Q}_1^{second} + \mathbb{Q}_2^{second} + \mathbb{Q}_3^{second} + \mathbb{Q}_4^{second} \qquad (2.15)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{Q}_1^{first} &= \mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(\sum_{i=1}^{n} P_1 \times logP(x_i|t_i, q_{y_i} = 1, \theta) \right) \\
&= \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(P_1 \times logP(x_i|t_i, q_{y_i} = 1, \theta) \right) \right] \\
&= \sum_{i=1}^{n} P_1 \times \left[\frac{-1}{2} log((2\pi)^m |\sigma_x^2 I|) - \frac{1}{2}\sigma_x^{-2} \Big((x_i - \mu_x)^T (x_i - \mu_x) \right. \\
&\quad - \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old})^T P^T (x_i - \mu_x) - (x_i - \mu_x)^T P \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) \\
&\quad + tr(P^T P(\underset{1}{E}(t_i t_i^T|x_i, y_i, q_{y_i} = 1, \theta^{old}) \\
&\quad - \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old})^T)) \\
&\quad \left. + \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old})^T P^T P \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) \Big) \right], \qquad (2.16)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{Q}_1^{second} &= \mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(\sum_{i=1}^{n} P_\rho \times logP(x_i|t_i, q_{y_i} = \rho_y, \theta) \right) \\
&= \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(P_\rho \times logP(x_i|t_i, q_{y_i} = \rho_y, \theta) \right) \right] \\
&= \sum_{i=1}^{n} P_\rho \times \left[\frac{-1}{2} log((2\pi)^m |\sigma_x^2 I|) - \frac{1}{2}\sigma_x^{-2} \Big((x_i - \mu_x)^T (x_i - \mu_x) \right. \\
&\quad - \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})^T P^T (x_i - \mu_x) - (x_i - \mu_x)^T P \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \\
&\quad + tr(P^T P(\underset{\rho}{E}(t_i t_i^T|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \\
&\quad - \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})^T)) \\
&\quad \left. + \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})^T P^T P \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \Big) \right], \qquad (2.17)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{Q}_2^{first} &= \mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(\sum_{i=1}^{n} P_1 \times logP(y_i|t_i, q_{y_i} = 1, \theta) \right) \\
&= \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(P_1 \times logP(y_i|t_i, q_{y_i} = 1, \theta) \right) \right] \\
&= \sum_{i=1}^{n} P_1 \times \left[\frac{-1}{2} log((2\pi)^r |\sigma_y^2 I|) - \frac{1}{2}\sigma_y^{-2} \Big((y_i - \mu_y)^T (y_i - \mu_y) \right. \\
&\quad - \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old})^T C^T (y_i - \mu_y) - (y_i - \mu_y)^T C \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) \\
&\quad + tr(C^T C(\underset{1}{E}(t_i t_i^T|x_i, y_i, q_{y_i} = 1, \theta^{old}) \\
&\quad - \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old})^T)) \\
&\quad \left. + \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old})^T C^T C \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) \Big) \right], \qquad (2.18)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{Q}_2^{second} &= \mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(\sum_{i=1}^{n} P_\rho \times logP(y_i|t_i, q_{y_i} = \rho_y, \theta) \right) \\
&= \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(P_\rho \times logP(y_i|t_i, q_{y_i} = \rho_y, \theta) \right) \right] \\
&= \sum_{i=1}^{n} P_\rho \times \left[\frac{-1}{2} log((2\pi)^r |\rho_y^{-1}\sigma_y^2 I|) - \frac{1}{2}\rho_y\sigma_y^{-2} \Big((y_i - \mu_y)^T (y_i - \mu_y) \right. \\
&\quad - \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})^T C^T (y_i - \mu_y) - (y_i - \mu_y)^T C \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \\
&\quad + tr(C^T C(\underset{\rho}{E}(t_i t_i^T|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \\
&\quad - \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})^T)) \\
&\quad \left. + \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})^T C^T C \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \Big) \right], \qquad (2.19)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{Q}_3^{first} &= \mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(\sum_{i=1}^{n} P_1 \times logP(t_i|q_{y_i}=1,\theta) \right) \\
&= \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(P_1 \times logP(t_i) \right) \right] \\
&= \sum_{i=1}^{n} P_1 \times \left[\frac{-1}{2} log((2\pi)^k|I|) - \frac{1}{2} \Bigg(tr(\underset{1}{E}(t_i t_i^T|x_i,y_i,q_{y_i}=1,\theta^{old}) \right. \\
&\quad - \underset{1}{E}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old}) \, \underset{1}{E}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old})^T) \\
&\quad \left. + \underset{1}{E}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old})^T \, \underset{1}{E}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old}) \Bigg) \right], \qquad (2.20)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{Q}_3^{second} &= \mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(\sum_{i=1}^{n} P_\rho \times logP(t_i|q_{y_i}=\rho_y,\theta) \right) \\
&= \sum_{i=1}^{n} \left[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \left(P_\rho \times logP(t_i) \right) \right] \\
&= \sum_{i=1}^{n} P_\rho \times \left[\frac{-1}{2} log((2\pi)^k|I|) - \frac{1}{2} \Bigg(tr(\underset{\rho}{E}(t_i t_i^T|x_i,y_i,q_{y_i}=\rho_y,\theta^{old}) \right. \\
&\quad - \underset{\rho}{E}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old}) \, \underset{\rho}{E}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old})^T) \\
&\quad \left. + \underset{\rho}{E}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old})^T \, \underset{\rho}{E}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old}) \Bigg) \right], \qquad (2.21)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{Q}_4^{first} &= \mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \Bigg(\sum_{i=1}^{n} P_1 \times logP(q_{y_i} = 1|\theta) \Bigg) \\
&= \sum_{i=1}^{n} \Bigg[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \Big(P_1 \times logP(q_{y_i} = 1|\theta) \Big) \Bigg] \\
&= \sum_{i=1}^{n} P_1 \times \Bigg[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \Big(log(1 - \delta_y) \Big) \Bigg] \\
&= \sum_{i=1}^{n} P_1 \times \Big[log(1 - \delta_y) \Big], \qquad (2.22)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{Q}_4^{second} &= \mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \Bigg(\sum_{i=1}^{n} P_\rho \times logP(q_{y_i} = \rho_y|\theta) \Bigg) \\
&= \sum_{i=1}^{n} \Bigg[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \Big(P_\rho \times logP(q_{y_i} = \rho_y|\theta) \Big) \Bigg] \\
&= \sum_{i=1}^{n} P_\rho \times \Bigg[\mathop{\mathbb{E}}_{T|X,Y,Q_Y,\theta^{old}} \Big(log\delta_y \Big) \Bigg] \\
&= \sum_{i=1}^{n} P_\rho \times \Big[log\delta_y \Big], \qquad (2.23)
\end{aligned}
$$

where $P_1 = P(q_{y_i} = 1|x_i, y_i, \theta^{old})$ and $P_\rho = P(q_{y_i} = \rho_y|x_i, y_i, \theta^{old})$.

To obtain the parameter update equations, (2.9) should be maximized with respect to the set of parameters. Simply,

$$
\begin{aligned}
\theta &= \operatorname*{argmax}_{\theta} \mathop{\mathbb{E}}_{T,Q_Y|X,Y,\theta^{old}} (logP(X, Y, T, Q_Y|\theta)) \\
&= \operatorname*{argmax}_{\theta} \mathbb{Q} \qquad (2.24)
\end{aligned}
$$

which is equal to solving a set of equations (2.25). For each parameter, partial differentiation of a specific term of $Q - function$, which contains the parameter of interest, is used. Results for the parameter update equations, after solving (2.25), are

given as in (2.26) to (2.37).

$$\begin{array}{ll} \text{P:} & \frac{\partial \mathbb{Q}_1}{\partial P} = 0 \\ \text{C:} & \frac{\partial \mathbb{Q}_2}{\partial C} = 0 \\ \sigma_x^2\text{:} & \frac{\partial \mathbb{Q}_1}{\partial \sigma_x^2} = 0 \\ \sigma_y^2\text{:} & \frac{\partial \mathbb{Q}_2}{\partial \sigma_y^2} = 0 \\ \mu_x\text{:} & \frac{\partial \mathbb{Q}_1}{\partial \mu_x} = 0 \\ \mu_y\text{:} & \frac{\partial \mathbb{Q}_2}{\partial \mu_y} = 0 \\ \delta_y\text{:} & \frac{\partial \mathbb{Q}_4}{\partial \delta_y} = 0 \\ \rho_y\text{:} & \frac{\partial \mathbb{Q}_2}{\partial \rho_y} = 0 \end{array} \tag{2.25}$$

$$\begin{aligned} P^{new} &= \Bigg[\sum_{i=1}^{n}\Bigg(2(x_i-\mu_x)\big(P_1 \underset{1}{E}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old})^T \\ &\quad +P_\rho \underset{\rho}{E}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old})^T\big)\Bigg)\Bigg] \\ &\quad \times\Bigg[\sum_{i=1}^{n}\Bigg(P_1\, \underset{1}{Coeff}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old}) \\ &\quad +P_\rho\, \underset{\rho}{Coeff}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old})\Bigg)\Bigg]^{-1}, \end{aligned} \tag{2.26}$$

where $P_1 = P(q_{y_i}=1|x_i,y_i,\theta^{old})$ and $P_\rho = P(q_{y_i}=\rho_y|x_i,y_i,\theta^{old})$ and

$$\begin{aligned} \underset{*}{Coeff}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old}) &= \underset{*}{S}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old}) + \underset{*}{S}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old})^T \\ &\quad +2\,\underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old}) \times \underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old})^T, \end{aligned} \tag{2.27}$$

and

$$\begin{aligned} \underset{*}{S}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old}) &= \underset{*}{E}(t_i t_i^T|x_i,y_i,q_{y_i}=*,\theta^{old}) \\ &\quad -\underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old}) \times \underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old})^T, \end{aligned} \tag{2.28}$$

in which the asterisk subfix indicates either cases of $q_{y_i} = 1$ or $q_{y_i} = \rho_y$.

Similarly,

$$\begin{aligned} C^{new} &= \Bigg[\sum_{i=1}^{n}\Bigg(2(y_i-\mu_y)\big(P_1 \underset{1}{E}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old})^T \\ &\quad +\rho_y P_\rho \underset{\rho}{E}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old})^T\big)\Bigg)\Bigg] \\ &\quad \times\Bigg[\sum_{i=1}^{n}\Bigg(P_1\, \underset{1}{Coeff}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old}) \\ &\quad +\rho_y P_\rho\, \underset{\rho}{Coeff}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old})\Bigg)\Bigg]^{-1}. \end{aligned} \tag{2.29}$$

$$\sigma_x^{2\,new} = \frac{\sum_{i=1}^{n}\left(P_1 \underset{1}{A}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old}) + P_\rho \underset{\rho}{A}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old})\right)}{n\,m}, \tag{2.30}$$

where,

$$\begin{aligned} \underset{*}{A}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old}) &= (x_i-\mu_x)^T(x_i-\mu_x) - \underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old})^T P^T(x_i-\mu_x) \\ &\quad -(x_i-\mu_x)^T P\, \underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old}) \\ &\quad +tr\Bigg(P^T P(\underset{*}{E}(t_i t_i^T|x_i,y_i,q_{y_i}=*,\theta^{old}) \\ &\quad -\underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old})\, \underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old})^T)\Bigg) \\ &\quad +\underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old})^T P^T P\, \underset{*}{E}(t_i|x_i,y_i,q_{y_i}=*,\theta^{old})\,. \end{aligned} \tag{2.31}$$

$$\sigma_y^{2\,new} = \frac{\sum_{i=1}^{n}\left(P_1 \underset{1}{B}(t_i|x_i,y_i,q_{y_i}=1,\theta^{old}) + \rho_y P_\rho \underset{\rho}{B}(t_i|x_i,y_i,q_{y_i}=\rho_y,\theta^{old})\right)}{n\,r}, \tag{2.32}$$

where,

$$\begin{aligned}
\underset{*}{B}(t_i|x_i, y_i, q_{y_i} = *, \theta^{old}) &= (y_i - \mu_y)^T(y_i - \mu_y) - \underset{*}{E}(t_i|x_i, y_i, q_{y_i} = *, \theta^{old})^T C^T (y_i - \mu_y) \\
&\quad -(y_i - \mu_y)^T C \, \underset{*}{E}(t_i|x_i, y_i, q_{y_i} = *, \theta^{old}) \\
&\quad +tr\Bigg(C^T C(\underset{*}{E}(t_i t_i^T|x_i, y_i, q_{y_i} = *, \theta^{old}) \\
&\quad - \underset{*}{E}(t_i|x_i, y_i, q_{y_i} = *, \theta^{old}) \, \underset{*}{E}(t_i|x_i, y_i, q_{y_i} = *, \theta^{old})^T) \Bigg) \\
&\quad + \underset{*}{E}(t_i|x_i, y_i, q_{y_i} = *, \theta^{old})^T C^T C \, \underset{*}{E}(t_i|x_i, y_i, q_{y_i} = *, \theta^{old}) \, .
\end{aligned} \tag{2.33}$$

$$\begin{aligned}
\mu_x^{new} &= \Bigg[\sum_{i=1}^{n} \Bigg(x_i - P\big(P_1 \, \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) + P_\rho \, \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})\big) \Bigg) \Bigg] \\
&\quad \times \Bigg[\sum_{i=1}^{n} \Big(P_1 + P_\rho \Big) \Bigg]^{-1} \\
&= \frac{\sum_{i=1}^{n} \Bigg(x_i - P\big(P_1 \, \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) + P_\rho \, \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})\big) \Bigg)}{n}.
\end{aligned} \tag{2.34}$$

Similarly,

$$\begin{aligned}
\mu_y^{new} &= \Bigg[\sum_{i=1}^{n} \Bigg(y_i(P_1 + \rho_y P_\rho) - C\big(P_1 \, \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) \\
&\quad +\rho_y P_\rho \, \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old})\big) \Bigg) \Bigg] \\
&\quad \times \Bigg[\sum_{i=1}^{n} \Big(P_1 + \rho_y P_\rho \Big) \Bigg]^{-1},
\end{aligned} \tag{2.35}$$

$$\delta_y^{new} = \frac{\sum_{i=1}^{n} P_\rho}{\sum_{i=1}^{n} P_1 + P_\rho} = \frac{\sum_{i=1}^{n} P_\rho}{n}, \tag{2.36}$$

and

$$\rho_y^{new} = r\Bigg[\sum_{i=1}^{n} P_\rho \Bigg] \times \Bigg[\sum_{i=1}^{n} \sigma_y^{-2} P_\rho \, \underset{\rho}{B}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \Bigg]^{-1}. \tag{2.37}$$

2.4.3 A-Posteriori Distributions of Hidden Variables

As shown in Section 2.4.2, Q_Y and T in (2.1) were treated as hidden variables to carry out EM algorithm in the stage of parameter estimation. For using the developed model to predict the outputs corresponding to the new inputs, as will be shown in Section 2.4.4, and to finalize the solution of EM algorithm, we need to calculate the posterior probabilities of the hidden variables. Moreover, for output prediction, the expectation of the variable matrix T is needed. In essence, the goal of the expectation step in EM is to bring the updated parameters of the previous maximization step to recalculate the a-posteriori probability distribution of latent variables and then their expected values. The distributions are obtained via *Bayes' rule* and *chain rule of probability* as in (2.38),

$$\begin{aligned} P(t_i|x_i, y_i, q_{y_i}, \theta^{old}) &= \frac{P(x_i, y_i|t_i, q_{y_i}, \theta^{old})}{P(x_i, y_i|q_{y_i}, \theta^{old})} \\ &= \frac{P(x_i|t_i, q_{y_i}, \theta^{old})P(y_i|t_i, q_{y_i}, \theta^{old})}{P(x_i, y_i|q_{y_i}, \theta^{old})} \end{aligned} \tag{2.38}$$

in which, all terms of the numerator have a Gaussian distribution and the denominator acts as a normalizing constant [27]. Expected mean and variance-related terms of posterior distribution (2.38) are given in (2.39) to (2.42) for different values of the discrete hidden variable Q_Y.

$$\begin{aligned} \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) = (&\sigma_x^{-2}P^TP + \sigma_y^{-2}C^TC + I)^{-1} \\ &\times (\sigma_x^{-2}P^T(x_i - \mu_x) \\ &+ \sigma_y^{-2}C^T(y_i - \mu_y)) \end{aligned} \tag{2.39}$$

$$\begin{aligned} \underset{1}{E}(t_it_i^T|x_i, y_i, q_{y_i} = 1, \theta^{old}) = (&\sigma_x^{-2}P^TP + \sigma_y^{-2}C^TC + I)^{-1} \\ &+ \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old}) \\ &\times \underset{1}{E}(t_i|x_i, y_i, q_{y_i} = 1, \theta^{old})^T \end{aligned} \tag{2.40}$$

$$\begin{aligned} \underset{\rho}{E}(t_i|x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) = (&\sigma_x^{-2}P^TP + \rho_y\sigma_y^{-2}C^TC + I)^{-1} \\ &\times (\sigma_x^{-2}P^T(x_i - \mu_x) \\ &+ \rho_y\sigma_y^{-2}C^T(y_i - \mu_y)) \end{aligned} \tag{2.41}$$

$$\underset{\rho}{E}(t_i t_i^T | x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) = (\sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} + \underset{\rho}{E}(t_i | x_i, y_i, q_{y_i} = \rho_y, \theta^{old}) \times \underset{\rho}{E}(t_i | x_i, y_i, q_{y_i} = \rho_y, \theta^{old})^T \quad (2.42)$$

A similar approach is taken for the other hidden variable as follows in (2.43) onwards.

$$P(q_{y_i} | x_i, y_i, \theta^{old}) = \frac{P(x_i, y_i | q_{y_i}, \theta^{old}) P^*(q_{y_i} | \theta^{old})}{P(x_i, y_i | \theta^{old})} \quad (2.43)$$

in which P^* is the prior probability for this hidden variable Q_Y and the first term on the numerator is the joint conditional probability of the input and output which is easily calculated using multivariate Gaussian distribution properties. The prior for Q_Y is a Bernoulli distribution since this variable is a binary variable. The distribution is shown in (2.4). The above mentioned posterior distribution for every value of q_{y_i} acts as the proportion of the corresponding output noise component.

2.4.4 Online Prediction

According to the developed RPPCA based regression model, we can construct an online inference method to predict the desired output variable which can be the quality variable in a process, based on some input variables. This method is also known as soft sensing [24]. To do so, we need the posterior probabilities of the hidden variables as formulated in Section 2.4.3, given the new input variables, namely $P(t_i | x_i, y_i, q_{y_i}, \theta^{old})$ and $P(q_{y_i} | x_i, y_i, \theta^{old})$ which could be calculated from (2.38) and (2.43), respectively. Next, the hidden variables should be estimated. In our developed model, the discrete hidden variable gave the equations two distinct terms as seen in terms (2.39) and (2.41). This happened when the joint probability of the hidden variables was factorized using *chain rule of probability*, as in (2.44). So for prediction we only need to estimate the continuous hidden variable t, given the different scenarios for the other hidden variable, q_y. This estimation is done via *law of total expectation* as in (2.45).

$$P(t_i, q_{y_i} | x_i, y_i, \theta^{old}) = P(t_i | x_i, y_i, q_{y_i}, \theta^{old}) \times P(q_{y_i} | x_i, y_i, \theta^{old}) \quad (2.44)$$

$$
\begin{aligned}
\hat{t}_i &= E(t_i|x_i, \theta^{old}) \\
&= \sum_{q_{y_i}} E(t_i|x_i, q_{y_i}, \theta^{old}) P(q_{y_i}|x_i, \theta^{old}) \\
&= \underset{1}{E}(t_i|x_i, q_{y_i} = 1, \theta^{old}) P(q_{y_i} = 1|x_i, \theta^{old}) \\
&\quad + \underset{\rho}{E}(t_i|x_i, q_{y_i} = \rho_y, \theta^{old}) P(q_{y_i} = \rho_y|x_i, \theta^{old})
\end{aligned} \tag{2.45}
$$

The desired quality variable prediction could be calculated as in (2.46), with the prediction error ϵ, as in (2.47) which is the difference between predicted and real values of the output variable. To evaluate the developed model prediction performance, there are a variety of measures available. Correlation between predicted and real test values is usually checked to make sure the trend of the data is captured. R-squared and mean squared error (MSE) are the other often used measures. Root mean square error (RMSE) is also a well-known measure since it has the same unit as the output variable and might be helpful to give a quantitative sense in comparisons.

$$
\hat{y}_i = C\hat{t}_i + \hat{\mu}_y \tag{2.46}
$$

$$
\epsilon = \hat{y} - y \tag{2.47}
$$

RMSE is defined as

$$
RMSE \triangleq \sqrt{\frac{\sum_{i=1}^{n'} \|\hat{y}_i - y_i\|^2}{n'}} \tag{2.48}
$$

where, n' is the total number of test samples and $\hat{y}$ and y are predicted and real output values, respectively.

2.5 Case Studies for Model Performance Assessment

In this section, two examples are discussed to evaluate the robustness of our developed method. In the first case study, a data set is generated by a generative PPCA based regression model; in the second case study, a data set from an industrial steam-assisted gravity drainage process plant is employed for evaluation.

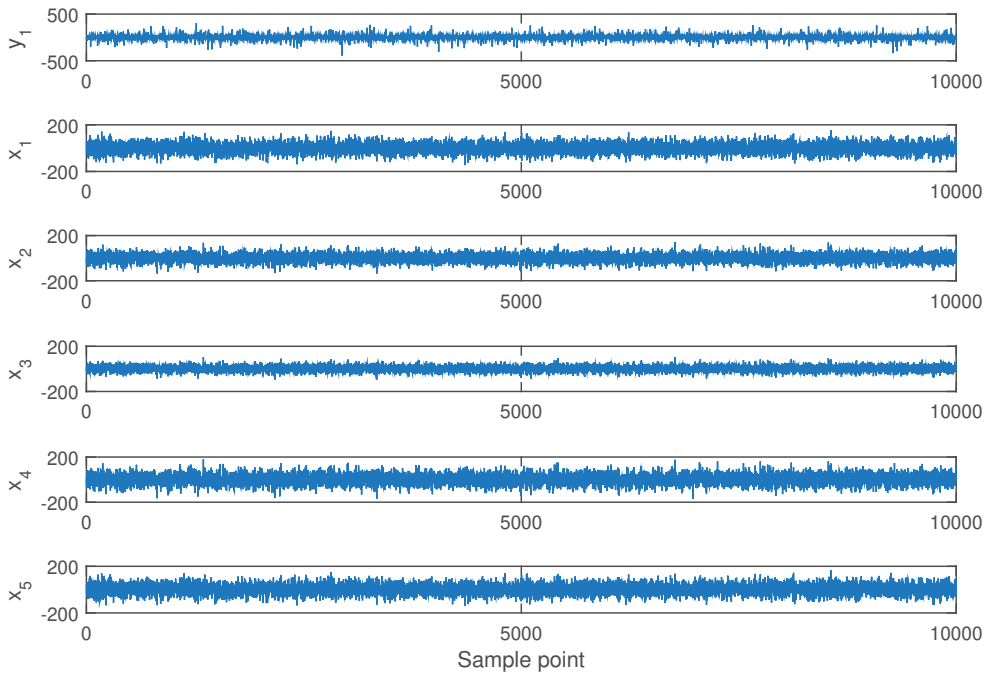


Figure 2.2: Generated data for numerical example

2.5.1 Case I: Numerically Generated Data Example

For this example, a data set of 10000 sample points from 5 inputs and 1 output variable is generated as follows by (2.1)

$$\begin{aligned} x_i &= Pt_i + \mu_x + e_i \\ y_i &= Ct_i + \mu_y + f_i \end{aligned} \tag{2.49}$$

where the loading matrix is $P_{(m=5)\times(k=2)} = [40, 10; 20, 30; 15, 20; 20, 40; 40, 15]$, regression coefficient matrix is $C_{(r=1)\times(k=2)} = [10, 20]$ and input/output mean vectors are $\mu_x = [1, 2, 3, 4, 5]$ and $\mu_y = [0]$, respectively. Input/output noises are distributed as $e_i \sim \mathcal{N}(0, 30I)$, $f_i \sim \mathcal{N}(0, 30I)$, respectively. Each latent variable vector has the standard Gaussian distribution as $t_{i_{k\times 1}} \sim \mathcal{N}(0, I)$, making the latent variable matrix, $T_{n\times k} = [t_1, \cdots, t_n]^T$, to become a set of $i.i.d$ vectors. Data from the generative model (2.49) is shown in Figure 2.2. To model this data by the probabilistic principal component analysis regression model, knowing the dimension of the latent space is necessary. Generally, a rough estimate of the latent space dimension is obtained by performing principal component analysis on data. Here in this example, the real

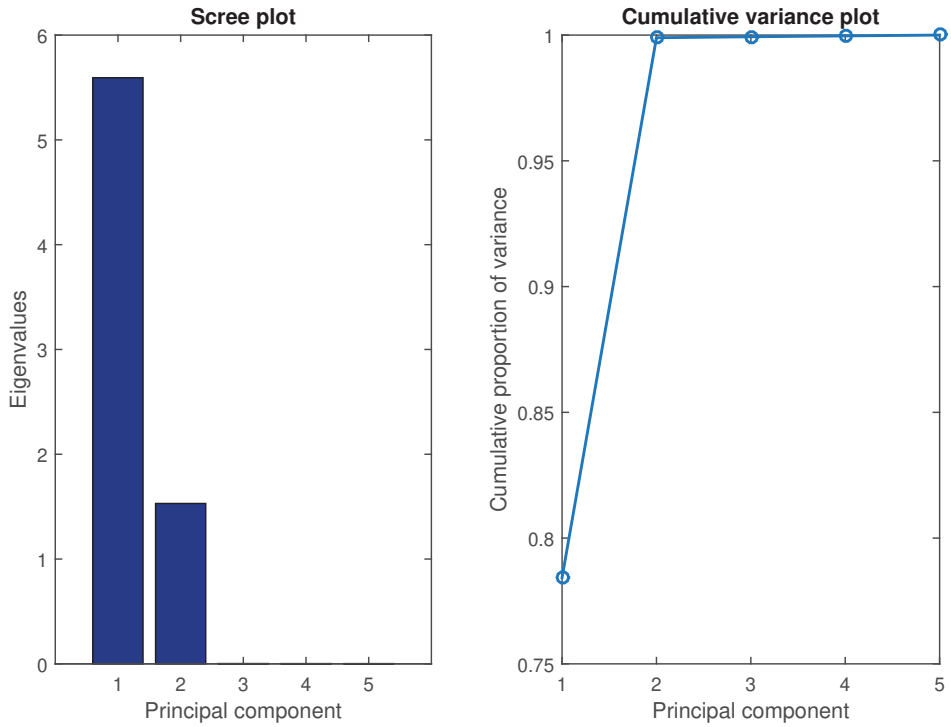


Figure 2.3: Scree plot (left) and cumulative variance explained plot (right) for generated data used in numerical example, obtained through PCA

latent space is two-dimensional as shown by the scree plot in Figure 2.3.

To see the robustness of the RPPCA based regression model, some random data points are replaced with outliers to contaminate the whole data set. This causes the performance of the PPCA regression model to diminish. Feeding contaminated data to the RPPCA based model results in better prediction performance compared to that of the PPCA based model, showing that the RPPCA based model tolerates the contamination. Figure 2.4 and Figure 2.5 illustrate the performance of the PPCA model when the generated data is used to estimate the parameters.

Data is contaminated by 10% outlying points which come from a model with a noise of scaled Gaussian distribution with same mean as the first component of the mixture output noise and a variance twice of that of the first component. This data makes the conventional PPCA based model perform poor. Figure 2.6 and Figure 2.7 show the performance of the RPPCA based model on the contaminated data set. Results are summarized in Table 2.1. Figure 2.8 shows the convergence of model parameters for the RPPCA based regression model.

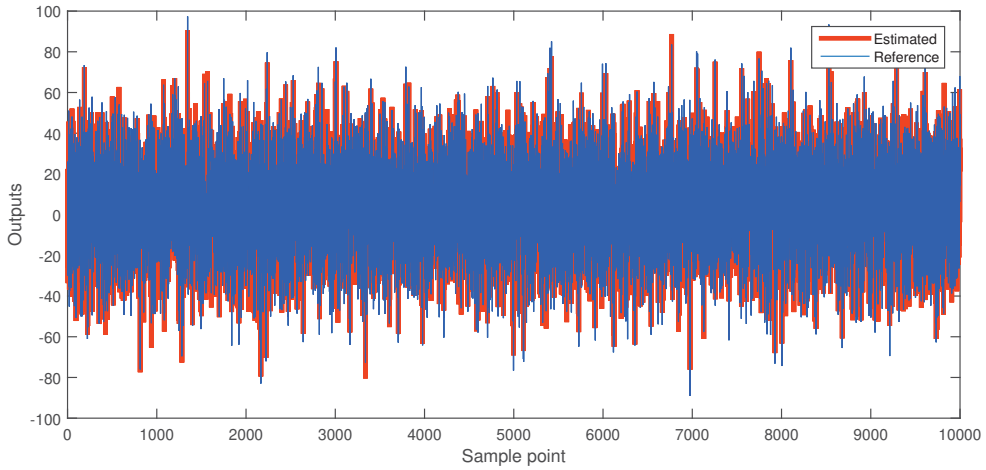


Figure 2.4: Prediction performance of PPCA regression model on generated data

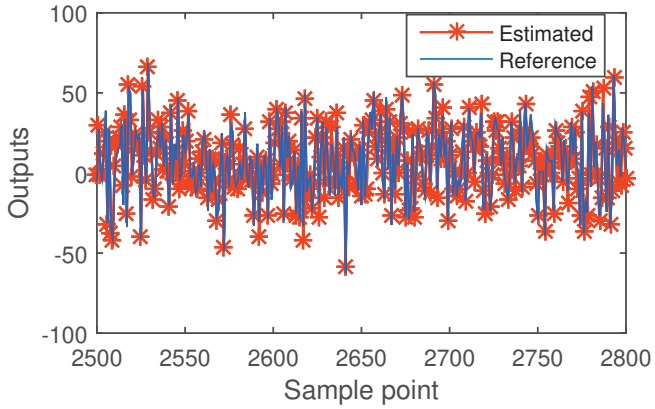


Figure 2.5: Prediction performance of PPCA regression model on generated data - Zoomed in for better comparison

Table 2.1: Prediction performance of regular and robust models in numerical example

	Regular data	Contaminated data	
	PPCA	PPCA	RPPCA
R^2	0.9346	0.2648	0.9348
$RMSE$	5.8255	33.6799	5.8155

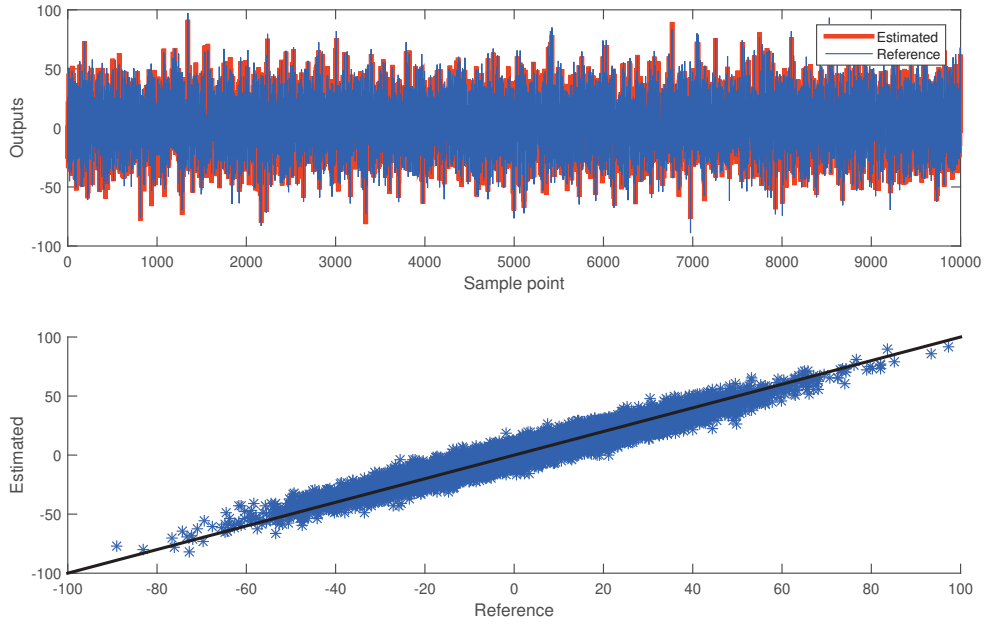


Figure 2.6: Prediction performance of RPPCA regression model on contaminated generated data

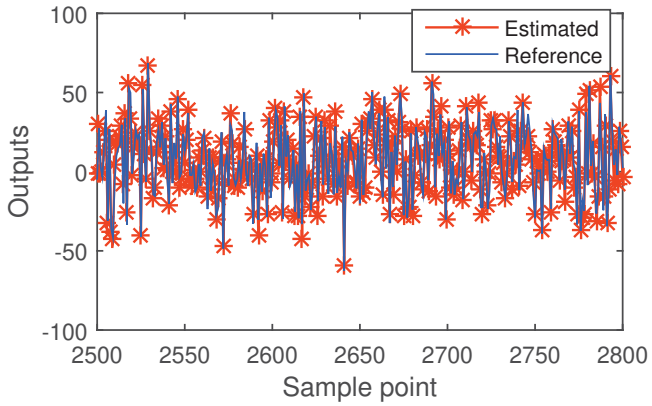


Figure 2.7: Prediction performance of RPPCA regression model on contaminated generated data - Zoomed in for better comparison

In this example we intended to show the effect of outlying observations of the scaled category. The results confirm the robustness of the developed RPPCA based regression model, while prediction performance of the regular PPCA based regression model was affected by outlying observations.

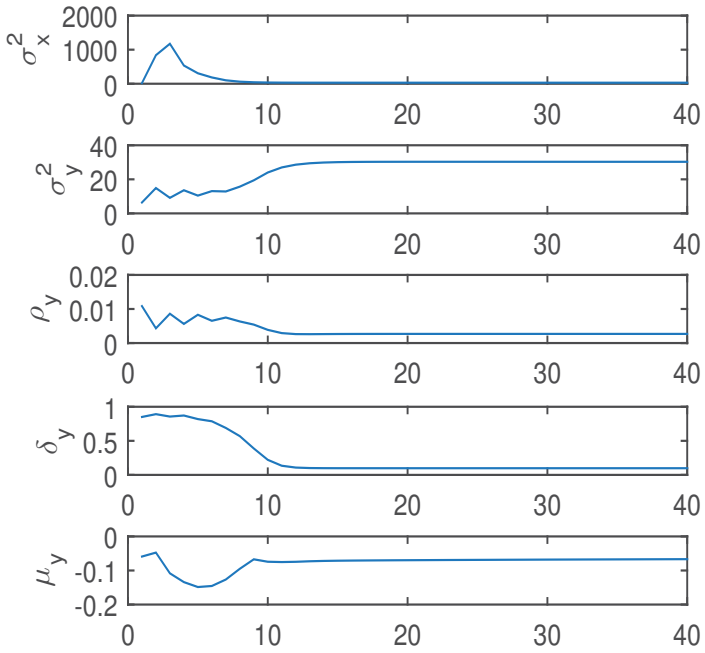


Figure 2.8: Parameter convergence for RPPCA model in first case study

2.5.2 Case II: Industrial Application

In this section, a set of data from an industrial plant is used to evaluate the robustness of the developed model. Data are collected from a steam-assisted gravity drainage (SAGD) operation.

SAGD process

SAGD is an in-situ method of oil recovery which is employed in heavy crude oil and bitumen production. This method is favorable to areas in which oil sands are deep seated in the ground and open pit mining is not feasible. This technology is becoming more common recently and it is forecast to reach higher production rates by the next ten years. The other beneficial aspect of this technology is that the water consumption during this kind of oil extraction is less than that of the conventional methods. Steam to oil ratio, which is an efficiency measure of steam-assisted gravity drainage operations, has been reduced in SAGD operation which is a result of high percentage of water recycling in this operation.

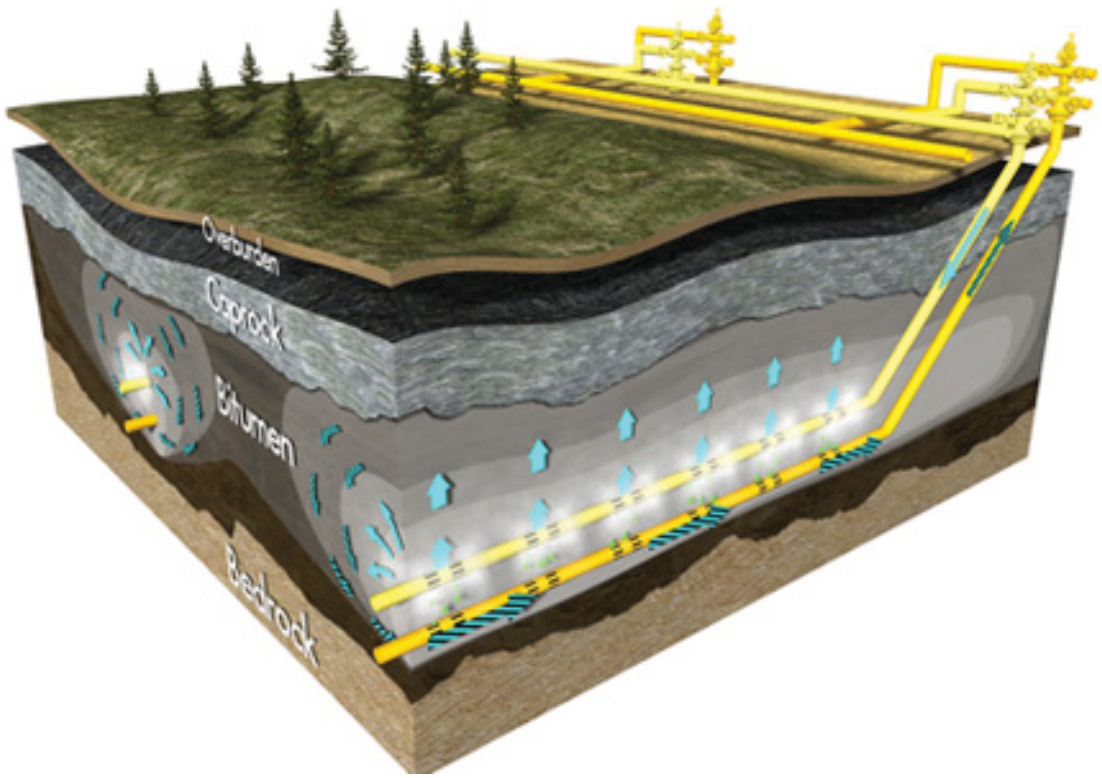


Figure 2.9: SAGD operation schematic (http://www.huskyenergy.com)

In this technique two horizontal wells, that are usually located one above each other with a four to six meter distance, are drilled under the ground surface. This well pair is connected to a central well pad. Steam is generated from steam generators located in the utilities section of the plant, and is then transferred into the ground via the top well as shown in Figure 2.9, also known as steam injection well. The heat released from injected steam brings heavy oil, deep seated under the ground, to a temperature point in which it would flow towards the bottom well, also known as production well. At the same time that steam heats the heavy oil, the specific situation of wells allows for gravity drainage; therefore, the appellation *SAGD* was adopted.

Robust modeling of SAGD process

The objective of this section is to develop a model to continuously predict produced fluid flow rates of the SAGD operation. The output stream of this operation is in the form of an emulsion which contains bitumen, water, and traces of sand particles and

is sent to downstream operations for further processing. The measurements for this stream flow rate are not frequently available and the available measurements might not be reliable due to the previously mentioned factors. Thus, to better operate the downstream operations it is critical to have frequent and more precise estimates of this flow rate and that is the goal of this case study.

The available historical data from this process are a set of flow rates, pressures, and temperatures for the injection and production well pair, which are constructed by a $10-min$ average data recorded from 1/1/2014 to 12/31/2014. Four of the mentioned variables which are highly correlated with the product oil flow rate are selected and smoothed for further analysis; among them three principal components are selected to result in a good performance for the conventional PPCA based regression model. The variable time series trends are shown in Figure 2.10. This data set consists of a total of 4 inputs and 1 output variables that have been recorded in 37690 sample points. For propitiatory reason, all y-axis values of the data plots have been removed.

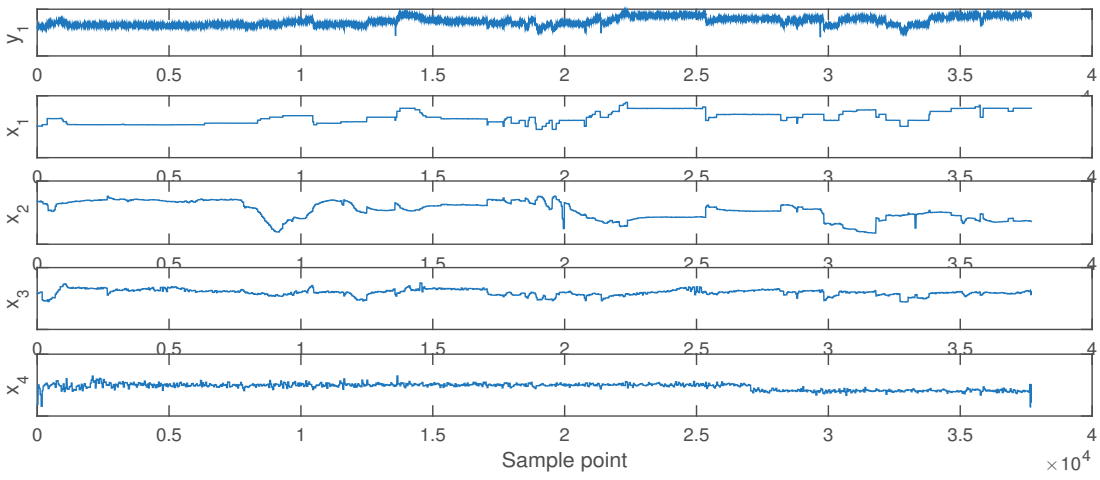


Figure 2.10: Data for industrial case study

In this case study, first, a PPCA based regression model is used to fit the data. The model shows a good prediction performance with 3 principal components, as scree plot shows in Figure 2.13. This performance is considered as a control case

and is regarded as basis of the comparison. Figure 2.11 shows the reference predicted emulsion flow rate versus the estimated flow rate from this control model. To evaluate

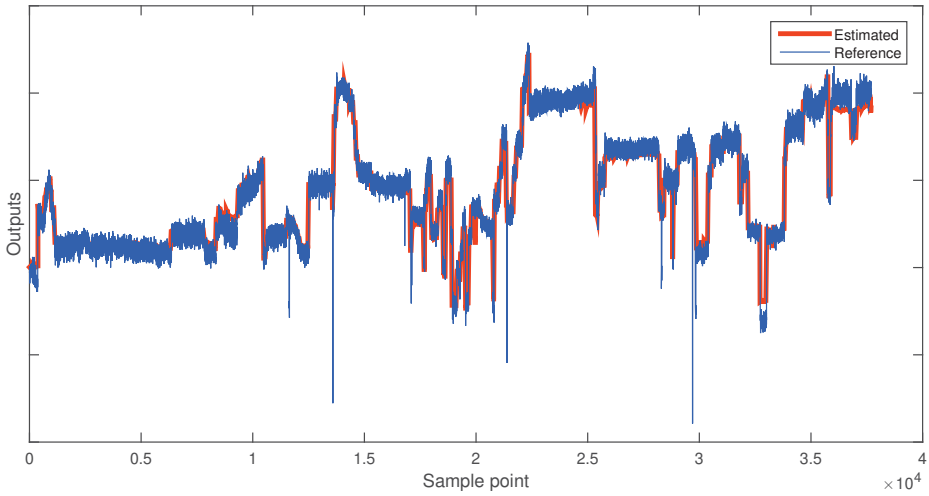


Figure 2.11: Prediction performance of PPCA regression model on industrial data

the robustness of RPPCA based regression model, outlying observations with inflated noise variance are used to randomly replace a portion of the observations. As shown in Figure 2.12, 2500 data points are replaced by outlying observations. This amount is around 7% of the whole sample size. The outliers are sampled from a Gaussian distribution whose variance is 5 times as large as the normal data variance. Note that in the case of the regular PPCA based regression model, the output noise was assumed to be a single Gaussian.

The percentage of outlying observations with mixture Gaussian noise with a scaled Gaussian component, and the amount of inflation considered for the variance of the contaminating noise component, are the two important factors which can affect the prediction performance of the regular PPCA based regression model. Though it is worthwhile to mention that contaminating the data (i.e., manipulating these two factors) would involve restrictions, specifically when real industrial data is being used. Increasing the variance inflation for the contaminating noise component or increasing the number of sample points which are replaced with abnormalities, could both be done to some extent after which the ability of the model to predict the trend will diminish. This behavior is due to the fact that information contained in the data is altered by too much (in terms of number of points) or by very abnormal (in terms of

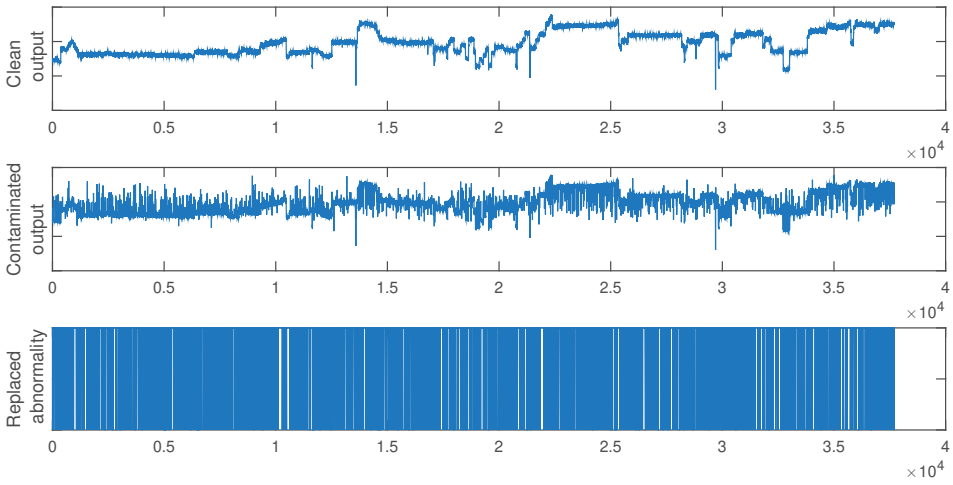


Figure 2.12: Data before and after introducing abnormality

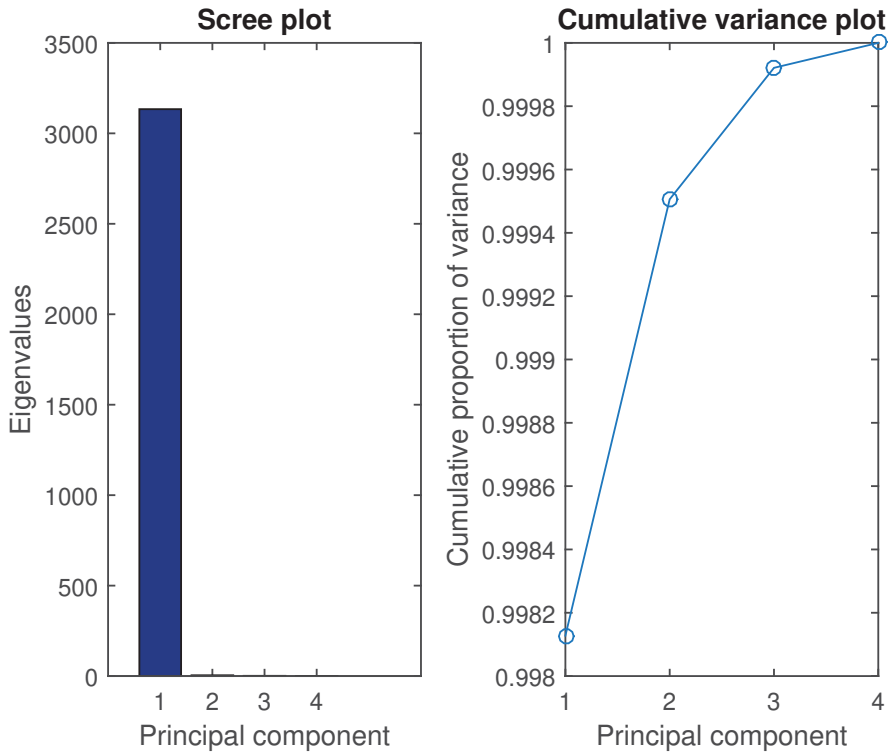


Figure 2.13: Scree plot (left) and cumulative variance explained plot (right) for industrial data used in second case study, obtained through PCA

inflation) noise. The more outlying observations replace historical data and/or the more inflated variance of the contaminated noise component is, the more deviation would be seen in predictions of the model. However, when the amount of abnormal

data increases noticeably, the model may lose credibility and reliability. To have a fair comparison, in both case studies, we imposed abnormality on the historical data records until the regular PPCA based regression model shows a noticeable failure in its performance. That is where RPPCA based regression model would demonstrate its beneficial characteristics. Results of this case study are summarized in Table 2.2.

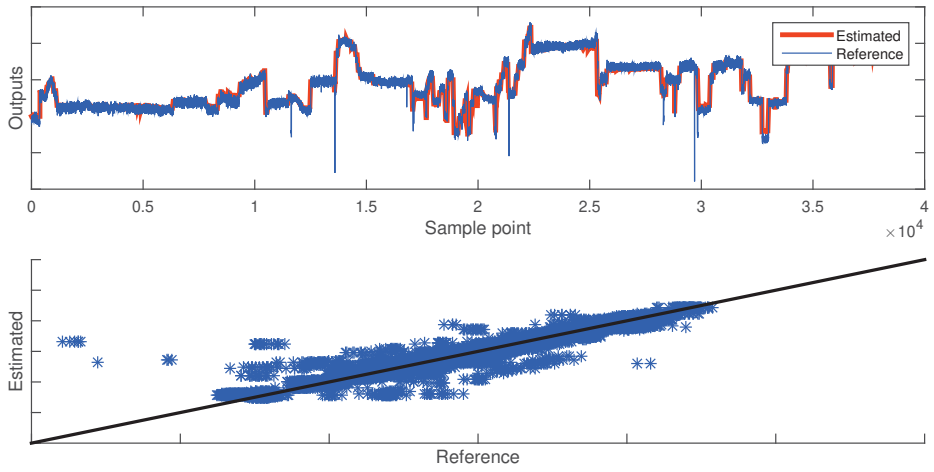


Figure 2.14: Prediction performance of RPPCA regression model on contaminated industrial data

In terms of performance measures, the PPCA based regression model has a weaker performance when applied to contaminated data set. The degree of performance reduction would be higher if the contamination is more severe. The RPPCA based model on the same data has superior performance as shown in Figure 2.14. Parameter convergence for RPPCA based model is shown in Figure 2.15.

Table 2.2: Prediction performance of regular and robust models in industrial plant case study

	Regular data	Contaminated data	
	PPCA	PPCA	RPPCA
R^2	0.9589	0.7665	0.9593
$RMSE$	0.6799	1.6889	0.6770

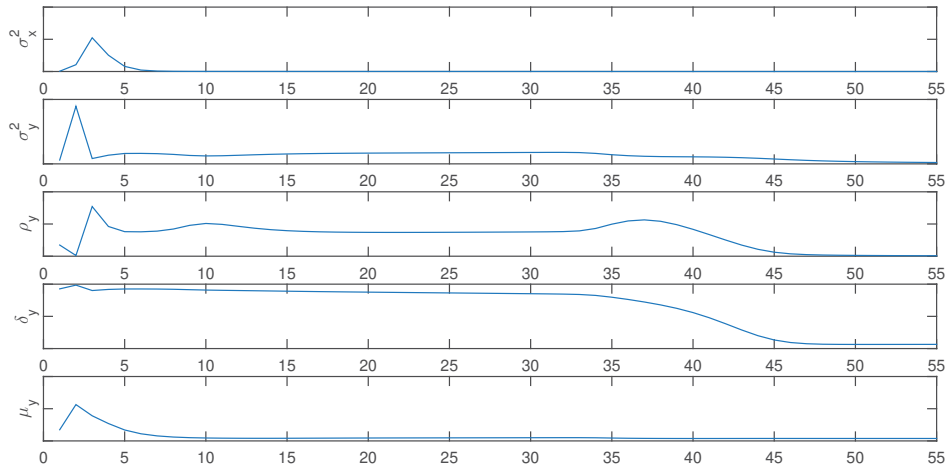


Figure 2.15: Parameter convergence for RPPCA model in industrial case study

2.6 Conclusions

A class of outlying observation problems was dealt with in this article. A contaminated noise assumption was considered for PPCA based regression models. Then a robust probabilistic model was developed, based on which a better prediction performance for the desired quality variable could be obtained in the presence of large random errors in data. Unlike the conventional PPCA based models with a single Gaussian noise model, the developed model downweighs the effect of scaled outlying observations in output prediction. The robust model development problem was formulated and solved with expectation maximization algorithm. Considering the contaminated Gaussian noise model allows us to get closed form solutions for model parameters, as well as downweighing the effect of the outlying noise in output prediction. Robustness and performance of the model were demonstrated through simulated data and industrial case studies.

Chapter 3

Robust Probabilistic Principal Component Analysis Based Process Modeling: Dealing with Simultaneous Contamination of Both Input and Output Data*

"It is not certain that
everything is uncertain."

Blaise Pascal (1623 - 1662)

*A version of this chapter is published as Anahita Sadeghian, Ouyang Wu, Biao Huang, 2018, "Robust probabilistic principal component analysis based process modeling: Dealing with simultaneous contamination of both input and output data", Journal of Process Control

3.1 Introduction

Predictive modeling, soft sensors and their necessity

Production rate and quality are important determinants of an industry's profitability. When setting higher goals for production rate and quality, industries must also account for safety and environmental factors [24]. To achieve these objectives advanced process monitoring and control techniques need to be developed. This task demands operational data to be recorded for further analysis. The need for recorded historical data for process control and monitoring purposes has led to heavy use of instrumentation in industries in the form of sensors and analyzers.

Despite the use of instrumentation, online and fast rate measurements of process variables are not always available. This is due to restricted availability and/or reliability of measurement devices and techniques in place which results in missing values, off values or delay in process data. In the best case when appropriate devices are available, there are still some key process variables that are determined offline or online which lack the two characteristics of availability and reliability. For example, a laboratory sample analysis lacks in terms of availability while samples from an online analyzer are often unreliable. On top of this, the conventional methods of data acquisition are time consuming and may introduce delay and discontinuity into their application [75].

Predictive modeling has received increasing attention and discussion in recent years. These models can provide the user with frequent estimates of quality variables of interest. These could also be the variables that are not measurable but can be inferred based on other available measurements. The subject is especially important in the area of chemical process operation and control. The importance lies in the fact that there could be losses in profitability and/or safety of a chemical plant in the case of failure of a measuring device which plays a critical role either in operation or in control, especially for some key process variables. There are considerable costs involved in installation and maintenance of hardware sensors to keep track of all variable measurements. To overcome the above mentioned issues predictive models have commonly been used to infer the key process variables. This leverages profitability in terms of reducing cost of new hardware sensor installation as well as in terms of

maintenance and backup for the available sensors. The leverage is also seen in terms of faster and more accurate measurements which come from synthesis of the existing sensors.

Predictive mathematical models, also known as soft sensors, inferential sensors or virtual sensors, provide online estimates of key process variables based on some other process records [24]. Taking advantage of these models could help us to enhance the system under study in terms of reliability and accuracy and develop better control and monitoring policies.

Among the three main categories of soft sensor design approaches discussed in [24], this thesis focuses on data-driven approach which is useful based on the fact that the plant records contain information about the operation and the factors which affect it. Since this information is deeply buried in data, one should be aware that the relevance and quality of the data at hand affects the performance of the model, along with the method of analysis used. For high fidelity modeling, it is important to reduce the adverse effect of noise and disturbances in data records. Although there are different filtering methods to handle this issue, a proper analysis of the data is a better way to cope with it. A proper analysis helps us to detect and deal with data issues such as outliers, missing data, redundancy, low accuracy, etc. Several ad hoc solutions to these potential issues have been addressed in literature [24, 42, 33, 10]. Presence of outlying observations or outliers, is one of the main factors that affect data quality [58]. The definition for these points will be given in Section 3.1.

One other important characteristic of data which is the base of the data-driven modeling approaches, is multi-modality. This characteristic has been studied in literature mainly for the multiple *process* models, such as in [11, 39, 102, 98] to deal with identification of complex systems. The authors in [3] dealt with identification of multiple local linear process models; and built a global process model by combining those local models. The same characteristic could exist in *noise* models. The focus of this chapter will be on considering multiple noise models that contribute to the process data. This situation could be observed when there are outlying observations among the process or laboratory data. These points are basically caused by random errors and are discussed further in Section 3.1.

Outlying observations and robustness

Conventional models are best suited when their assumptions are met. However, if the data does not satisfy those assumptions, results might be misleading. Specifically, presence of outliers is one of the cases that violate the assumption of normally distributed residuals (noises) in a regular PPCA based regression model. This could be a result of measurement error, a change in the experimental settings, or even due to that specific outlying sample belonging to another population rather than the one that data was supposed to represent. It should be reminded that an outlier does not necessarily imply a wrong or bad sample, however the terminology sometimes is used as such [36].

The main issues caused by outliers has been reviewed in Section 1.2. In this chapter the main focus will be on certain type of outliers and a specific robustness problem. In [4] outliers are defined as observations that are inconsistent with other observations, or as [36] refers to, they do not follow the model that fits the majority of the set. The authors in [4] believe these data points are sometimes hidden to the user since they might not show up in the residual plots. This category will be discussed below.

Outliers can cause an adverse effect on parameter estimation and also on the prediction while they may remain unnoticed [2]. Two general remedies exist. To detect the outliers, remove them and carry on the modeling without them, or to cope with them by feeding the contaminated data into a robust algorithm which is capable of handling them. One should be reminded that this handling is possible to some extent, that is, contamination could be tolerated to a specific threshold in terms of its density and location.

As the authors in [97] and [34] state, the purpose of diagnostics is to find and detect deviations from assumptions, while the purpose of robustness is to prevent deviations from assumptions. [34] categorizes outlier detection (same as what [97] proposes) under diagnostics, not robustness. The authors see robustness to be a procedure insensitive to outliers. Examining residuals gives us an idea about outliers, but this method on its own is not always sufficient. This is specifically true for those outliers which correspond to high leverage points [73].

Robust methods (a.k.a resistant methods) are commonly used when data is contaminated with outliers; these methods are developed such that they are not simply distressed by outliers. One of the effective performance statistics that compares the performance of a regular model to a robust model is coefficient of determination, also known as R-Squared which gives information on the goodness of fit [2]. Most robust models are efficient and insensitive to unusual values of data points, but markedly abnormal leverage points can diminish their performance [97, 65]. Therefore, it is important to do a robust design to address this issue.

There are two examples in [61] that illustrate high leverage data points in detail. Some data points might lie further away from the rest of the data. These points might be regression (or residual or vertical, as some references use) outlier, or X-space (or input) outliers. The points that have their X-space values to be unusual are called as high leverage points (points B, C, D and E, in Figure 3.1). The terminology comes from a lever and the fulcrum which is the balancing point. The balancing point stands for the mean value of X-space data. So a high leverage point is the one which sits far from the balancing point. Some of these high leverage points (X-space outliers) which sit along with the largest spread of data, in a 2-D set for instance, could not be seen as a residual outlier since they fall into the general pattern. These points are also called a good leverage point (point D, in Figure 3.1). Some of the high leverage points could also be a residual outlier at the same time (points B and C, in Figure 3.1). These are the points that sit far away from the mean in the X-space and are also having a big distance to the regression line passing through most of the data points. According to where in the 2-D space the point sits, it could have different combination of the outlying/inlying properties. For example, point E in Figure 3.1, is an inlier in Y-space but a residual outlier as well. Or point A is inlier in X-space, but a residual outlier. To be clear we remind that term *outlier* is mostly used in terms of residual space, whereas the term *Inlier* mentioned above, is used in terms of variable space.

To challenge a robust model thoroughly, considerations should be taken in the design of contamination for both percentage of outliers and the percentage of high leverage points [2]. As stated before, regressions are affected by the presence of regression (or vertical) outliers or bad leverage points. The importance of leverage

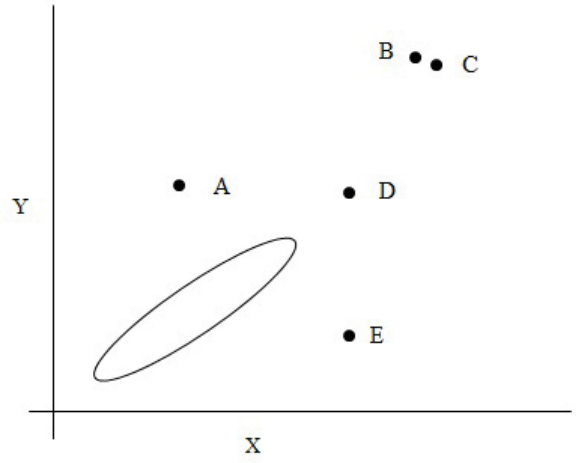


Figure 3.1: Scatter plot for different type of outlying observations-reproduced from [2]

point consideration could be seen by the effect of points A and B on the regression line. Both points have equal residuals to the regression line, but including/excluding B in the data will affect the regression slope (the model parameters) more than including/excluding A. This chapter is dedicated to the study of both contamination in the input space and in the response space. To do so, input and output noises of the PPCA based model are formulated as a Gaussian mixture with components which differ in location and in scale. For simplicity this is done in a symmetric layout. That means the side contaminating components of the Gaussian mixture noise are located at the same distance from the main component, but follow a different spread compared to the main one by an inflation factor. More details are being discussed in Section 3.3.

Solving all data issues in a single step is a highly complicated task. This chapter will address the problem of outliers through the development of a robust predictive model which is invulnerable to their presence. Outliers can belong to different families as discussed in [45]. One of the contributions of this chapter in extension to the previous work presented in Chapter 2, is on generalizing Gaussian mixture outliers, with special attention to symmetric Gaussian location outliers. This type of outliers represents a common problem such as a jammed instrument and is another distin-

guished type of outliers seen in processes in comparison with that considered in [75]. We propose a probabilistic modeling approach under a matching and more general noise formulation. Another contribution of this chapter is on handling the input outliers in addition to the output outliers. Existence of input outliers or leverage points in addition to output outliers significantly increases complexity of the problem. To derive the complex probabilistic robust principal component regression models, certain fundamentals are needed. The following sections of this chapter will concentrate on a brief review of basics of probabilistic principal component analysis (PPCA) and PPCA based regression models, as well as the type of outliers versus robustness. Next we present the formulation of our robust probabilistic model which is resistant to a common type of outlier in parameter estimation. The developed model is then solved for parameters using expectation maximization (EM) algorithm. Finally, the developed robust model is evaluated through discussion over two case studies which use a simulated and a real industrial data set.

3.2 Fundamentals

The preliminaries of the latent variables and the probabilistic modeling approach have been previously discussed in Section 1.3. PPCA-based regression model is also reviewed in Section 1.3.3. This section will review some preliminaries over the formulation of the aimed probabilistic approach and the method of solving the formulated problem.

In the current chapter, as a continuation of the work presented in Chapter 2 that focuses on scale outliers, we will consider a formation of Gaussian *location* mixture noise. Also, to make a broader exploration here, the location mixture noise appears in both input and output of the process, rather than a Gaussian *scaled* mixture that affects the output only, as the case of Chapter 2. As before, to estimate the unknown parameters of the model, loading matrices and the variances, maximization of likelihood of the complete data is performed. EM algorithm is used for this estimation because of presence of hidden variables [27].

3.2.1 Robust PPCA (RPPCA)-based Regression Model (with Gaussian Scaled Mixture Noise)

This type of model is a robust PPCA based regression model, which was discussed in Chapter 2, assumes a more representative noise distribution for output equation than the conventional PPCA based models. The noise is assumed to follow a mixture distribution consisting of two Gaussian components. One component of this noise model is a Gaussian distribution with a specific mean and variance, whereas the other one is a Gaussian with same mean as the first component but a different variance (often taking an extreme value). The variance of the latter one, which we call as contaminating part, is inflated with respect to that of the first component by a factor of ρ^{-1}, where $\rho \in (0, 1]$. In other words, this mixture output noise model is assumed to incorporate a scaled contaminating portion to acknowledge the outliers which originate from measurements violating the physical limitations of a process unit, such as a very large flow rate. Thus, the noise models would become as in (3.1); such a noise model will help downweighing the effect of outliers in the output prediction only.

$$\begin{aligned} e_i &\sim \mathcal{N}(0, \sigma_x^2 I) \\ f_i &\sim (1-\delta_y)\,\mathcal{N}(0, \sigma_y^2 I) + \delta_y\,\mathcal{N}(0, \rho_y^{-1}\sigma_y^2 I) \end{aligned} \tag{3.1}$$

Vector of variable $Q_Y = [q_{y_1}, \cdots, q_{y_n}]^T \in \Re^{n\times 1}$, $(q_{y_i} \in \Re^{1\times 1})$ is introduced as a vector of binary indicators which implies identity of each sample point. When this indicator $q_{y_i} = 1$, the output noise f_i is distributed as $\mathcal{N}(0, \sigma_y^2)$; and when $q_{y_i} = \rho_y$, f_i is distributed as $\mathcal{N}(0, \rho_y^{-1}\sigma_y^2)$. [45] represents a Bernoulli distribution for q_{y_i} as:

$$P(q_{y_i}|\delta_y, \rho_y) = \delta_y^{\,1-\frac{q_{y_i}-\rho_y}{1-q_{y_i}\rho_y}} (1-\delta_y)^{\frac{q_{y_i}-\rho_y}{1-q_{y_i}\rho_y}}, \tag{3.2}$$

where δ_y is the probability that the output observation y_i is generated by the contaminated Gaussian noise component; ρ_y is the variance inflation factor for that component which could vary between 0 and 1; smaller the value of ρ_y, the bigger would be the magnitude of noise (that leads to an outlying observation).

3.2.2 EM Algorithm

EM algorithm is an alternative iterative solution when a maximum likelihood problem is not tractable due to the absence of some data or variables. This data incompleteness [62] could arise from different scenarios such as missing observation data, truncated distributions, censored or grouped observations, or even from statistical models like random effects, mixtures, convolution and latent variable structures. This algorithm consists of two main steps as the name suggests, expectation and maximization. These iterative steps formulate a complete-data problem, given an incomplete-data problem. The complete-data resulted from this procedure is referred to as an augmented data. A succession of optimization tasks are performed in the EM formulation as opposed to one single complex optimization task performed in a maximum likelihood formulation based on an incomplete-data case [16]. The iterations in this algorithm are performed until a predefined convergence criterion is satisfied. This criterion could be the relative difference between the parameters or the complete-data likelihood function of two successive iterations. The latter is known to be faster; nevertheless, [94] in particular devoted effort to discussing whether the convergence of likelihood can automatically involve the convergence of parameters.

One of the advantages of this algorithm is its numerical stability, such that in each iteration the likelihood function is increased until the convergence occurs. As another beneficial properties of this algorithm, we can refer to easy implementation [50], reliability to provide the user with estimates for missing data and low computational cost per iteration as stated in [62]. This algorithm might demand the user to start over from different initial sets of parameters by using Monte Carlo simulation to achieve convergence in the case of multiple local optima for a multimodal complete-data likelihood function.

There have been discussions over the drawbacks of this algorithm as well. Mainly, the available literature refers to slow convergence, dependence of the solution on stopping criterion and on the selection of initial values for the parameters [64, 50, 6, 43]. [63] and [71] have dealt with the slow convergence issue to alleviate the problem by means of Aitken's approach or by creating different augmentations, respectively.

The concern on stopping criterion also has been addressed in many works such as

[80]. Main ideas for this concern are to base the criterion on the relative change in parameters as mentioned above and used in this chapter, or on the relative change of the log-likelihood function [57]. [43] suggests that when the algorithm gets stuck in the flat area of the likelihood function, one might continue the iterations hoping for reaching a global optimum after a large number of iterations. The authors in [43] prefer to start from various initial points and stop after a limited number of iterations and choose the best one.

The EM algorithm is also sensitive to the choice of initial values selected for parameters. Efficient initialization is an important step for the convergence of the algorithm to the best optimum point for the likelihood function and it also affects the speed of convergence. The authors in [64] also mention that among the substantial number of initialization methods they have reviewed, there is no method which could be considered as the best one in performance. They suggest to perform several methods and choose which works best for the problem. However, another option which has been used in the literature is to use the estimates which are obtained from other methods as an initial value set [26, 25]. As mentioned above, [43] have claimed that in practice it is best to start from multiple initial points to ensure the algorithm ends up with the global optimum. This variation in the selection of initial points is also known as compounding; and it is believed to increase the search area for a set of predefined iteration steps and it is claimed that it can save computation time [6].

In this chapter, to alleviate the effect of the initial values, we have used multiple random initialization sets for the regular PPCA based model and tried the algorithm for multiple simulations and chose the best answer. The other possible approach to the initialization problem, which is suggested by [50], is to combine EM algorithm with other global optimization methods such as genetic algorithm or particle swarm optimization. However, they indicate the heavy computation load demanded by this approach on the big data sets. In this chapter, for the robust PPCA based model, we have used the estimated parameters of the regular model as the initial values for the common parameters; and random initial values for the rest.

3.3 Robust Model Development

In this section, robust PPCA is modeled by considering a Gaussian location mixture noise for both input and output generative models. Then, the developed model is solved for its parameters through EM iterations. Final results are shown in Section 3.3.2. Posterior probabilities for hidden variables are presented in Section 3.3.3.

Probabilistic PPCA based regression model is introduced as shown below by [86].

$$\begin{cases} x_i = Pt_i + \mu_x + e_i \\ y_i = Ct_i + \mu_y + f_i \end{cases} \tag{3.3}$$

In this model, $t_i \in \Re^{q\times 1}$ is latent variable vector, $x_i \in \Re^{m\times 1}$ and $y_i \in \Re^{r\times 1}$ are observed input and output data vectors, respectively. Vector μ stands for mean values. The same model in matrix form of the above mentioned variables could be written as:

$$\begin{cases} X = TP^T + M_x + E \\ Y = TC^T + M_y + F \end{cases} \tag{3.4}$$

where similarly, $X = [x_1, \cdots, x_n]^T \in \Re^{n\times m}$, $Y = [y_1, \cdots, y_n]^T \in \Re^{n\times r}$ and $T = [t_1, \cdots, t_n]^T \in \Re^{n\times q}$. Besides the general assumptions for a PPCA based regression model as stated in [86, 75], here it is assumed that the noises for both generative models of input and output have the Gaussian mixture distribution. In this way, outliers for both input (high leverage points) and output measurements could be taken care of in the modeling process. Here, we assume a case of Gaussian location mixture which is capable of accounting for process measurements which do not conform to the technological extent of their measuring device and show off-values among the other measurements. These off-values are usually seen in both sides of the normal values. In other cases where the measurements are obtained through a jammed instrument, asymmetric location outliers could be observed [45, 75]. In this chapter a case of Gaussian location mixture with symmetric means in their distributions is studied for both input and output measurements. Note that symmetric means of the outliers distributions do not imply that the locations of each pair of outliers have to be symmetric. This formulation is a more general one compared to a single Gaussian noise. This will help in looking at outliers in the data by considering some large random errors which are added to data produced by the generative model (3.3). The

second term in the Gaussian mixture noise distribution in (3.5), stands for such errors.

$$\begin{aligned} e_i &\sim (1-\delta_x)\,\mathcal{N}(0,\sigma_x^2 I) + \delta_x\,[\,0.5\,\mathcal{N}(\Delta_x,\rho_x^{-1}\sigma_x^2 I) + 0.5\,\mathcal{N}(-\Delta_x,\rho_x^{-1}\sigma_x^2 I)] \\ f_i &\sim (1-\delta_y)\,\mathcal{N}(0,\sigma_y^2 I) + \delta_y\,[\,0.5\,\mathcal{N}(\Delta_y,\rho_y^{-1}\sigma_y^2 I) + 0.5\,\mathcal{N}(-\Delta_y,\rho_y^{-1}\sigma_y^2 I)] \end{aligned} \tag{3.5}$$

In this mixture Gaussian distribution of three Gaussian components, the second and the third components have a different mean and a different variance from the first one. This type is a more generalized case of a Gaussian location mixture where the second and third components not only have a different mean value from the first one, but also have a different variance from the first one. This could be termed as a combination case of location and scale Gaussian mixture that accounts for any sort of large random error scenarios mentioned in [45]. Vector of variables $Q_Y = [q_{y_1}, \cdots, q_{y_n}]^T \in \Re^{n\times 1}$, $(q_{y_i} \in \Re^{1\times 1})$ and $Q_X = [q_{x_1}, \cdots, q_{x_n}]^T \in \Re^{n\times 1}$, $(q_{x_i} \in \Re^{1\times 1})$ are introduced as vector of binary indicators which imply identity of each sample point in output and input data sets, respectively. When this indicator, for instance $q_{y_i} = 1$, the output noise f_i is distributed as $\mathcal{N}(0,\sigma_y^2)$; when $q_{y_i} = \Delta_y$, f_i is distributed as $\mathcal{N}(\Delta_y,\rho_y^{-1}\sigma_y^2 I)$; and when $q_{y_i} = -\Delta_y$, f_i is distributed as $\mathcal{N}(-\Delta_y,\rho_y^{-1}\sigma_y^2 I)$. This holds for inputs as well, where the two latter cases $q_i = \pm\Delta$ are showing the symmetric location-scaled counterparts of the main distribution for noise, that is case one $q_i = 1$. [45] represents a Bernoulli distribution for $|q_{y_i}|$ and $|q_{x_i}|$, which is the equivalent of a categorical distribution for these two values:

$$\begin{aligned} P(q_{y_i}|\theta) &= (0.5\delta_y)^{\frac{|q_{y_i}|+q_{y_i}}{2\Delta_y}}\,(0.5\delta_y)^{\frac{|q_{y_i}|-q_{y_i}}{2\Delta_y}}\,(1-\delta_y)^{1-\frac{|q_{y_i}|}{\Delta_y}}, \\ P(q_{x_i}|\theta) &= (0.5\delta_x)^{\frac{|q_{x_i}|+q_{x_i}}{2\Delta_x}}\,(0.5\delta_x)^{\frac{|q_{x_i}|-q_{x_i}}{2\Delta_x}}\,(1-\delta_x)^{1-\frac{|q_{x_i}|}{\Delta_x}}, \end{aligned} \tag{3.6}$$

where δ_y and δ_x are the probabilities that the output or input observation y_i or x_i is generated by the contaminating Gaussian noise component; ρ_y, ρ_x are the variance inflation factors for that component which could vary between 0 and 1, for output and input data respectively; the smaller the value of ρ is, the bigger would be the magnitude of noise (that leads to an outlying observation). Δ_y and Δ_x in this formulation are the shift in mean values for the contaminating counterpart in the Gaussian mixture. As it is seen in (2.3), the location counterpart has two terms which have

symmetric mean values on both sides of the mean value of the main component which is assumed to be zero.

3.3.1 RPPCA-based Model with Gaussian Location Mixture Noises in Both Input and Output

RPPCA based regression model parameters consist of P, C, μ_x, μ_y and the hyperparameters of noise terms e and f which are the noise variances σ_x^2, σ_y^2 and the inflation factors ρ_y, ρ_x, as well as δ_y, δ_x which adjust the proportion of the two Gaussian components in the Gaussian mixture noise. For this chapter, two location parameters Δ_y and Δ_x are also added. Common assumptions for PPCA regression model, as stated in [86] hold for this model as well.

To model and solve the problem of Robust PPCA based regression with Gaussian mixture input and output noise, EM algorithm is adopted in this work. Available input and output data are shown by matrices X and Y, respectively. Unknown information of the problem is:

$$\{P, C, \sigma_x^2, \sigma_y^2, \mu_x, \mu_y, \delta_y, \delta_x, \rho_y, \rho_x, \Delta_y, \Delta_x, T, Q_Y, Q_X\},$$

from which $\{t_i\}_{i=1}^n$, $\{q_{y_i}\}_{i=1}^n$ and $\{q_{x_i}\}_{i=1}^n$ are treated as hidden variables of the model and θ is defined to be set of parameters of the model to be estimated in Section 3.3.2.

$$\theta \triangleq \{P, C, \sigma_x^2, \sigma_y^2, \mu_x, \mu_y, \delta_y, \delta_x, \rho_y, \rho_y, \Delta_y, \Delta_x\}$$

To employ EM algorithm, the $Q - function$ has to be built based on both observed and hidden variables. The general form of a $Q - function$ is as shown in (3.7).

$$\mathbb{Q} = \underset{T,Q_Y,Q_X|X,Y,\theta^{old}}{\mathbb{E}} \Bigg(logP(\underbrace{\overbrace{X,Y}^{Observed} \ , \ \overbrace{T,Q_Y,Q_X}^{Hidden}}_{CompleteData} |\theta) \Bigg), \tag{3.7}$$

where $P(X, Y, T, Q_Y, Q_X|\theta)$ is regarded as the complete data likelihood. The other common assumption for this model is that the noises in the input and output data are independent and identically distributed (i.e., i.i.d.). Consequently the i.i.d. property also holds for the latent variable t_i and the sample indicators q_{y_i} and q_{x_i}.

$$
\begin{aligned}
P(X,Y,T,Q_Y,Q_X|\theta) &= P(X,Y|T,Q_Y,Q_X,\theta)\ P(T,Q_Y,Q_X|\theta) \\
&= P(X|T,Q_Y,Q_X,\theta)\ P(Y|T,Q_Y,Q_X,\theta) \\
&\quad \times P(T|\theta)\ P(Q_Y|\theta)\ P(Q_X|\theta) \\
&= \prod_{i=1}^{n} \Big[P(x_i|t_i,q_{y_i},q_{x_i},\theta)\ P(y_i|t_i,q_{y_i},q_{x_i},\theta) \\
&\qquad \times P(t_i|\theta)\ P(q_{y_i}|\theta)\ P(q_{x_i}|\theta) \Big] \qquad (3.8)
\end{aligned}
$$

The complete data likelihood is obtained as described in (3.8) in order to be substituted into the $Q-function$ (3.7). By taking the logarithm of both sides of (3.8), log-likelihood of the complete data would be a summation consisting of five terms as follows in (3.9). Based on different combinations of the two indicator variables that take three possibilities each, every term in (3.9) could have nine combinatorial outcomes.

$$
\begin{aligned}
\mathbb{L} &= logP(X,Y,T,Q_Y,Q_X|\theta) \\
&= \sum_{i=1}^{n} \Big[logP(x_i|t_i,q_{y_i},q_{x_i},\theta) + logP(y_i|t_i,q_{y_i},q_{x_i},\theta) \\
&\qquad + logP(t_i|\theta) + logP(q_{x_i}|\theta) + logP(q_{y_i}|\theta) \Big] \\
&= \sum_{i=1}^{n} \Big[\textcircled{I} + \textcircled{II} + \textcircled{III} + \textcircled{IV} + \textcircled{V} \Big] \qquad (3.9)
\end{aligned}
$$

For example distributions for $P(x_i|t_i,q_{y_i},q_{x_i},\theta)$ and $P(y_i|t_i,q_{y_i},q_{x_i},\theta)$ in nine different combinations would be shown as in Table 3.1. Therefore, first term in (3.9) would become as shown in Table 3.2. In a similar way second term of the complete-data log-likelihood could be obtained. For the third term, we have a normal distribution as in (3.10). Fourth and fifth terms are presented in (3.11) and (3.12), respectively.

$$
\text{Term III} \quad = \tfrac{-1}{2} log((2\pi)^k|I|) - \tfrac{1}{2}\left(t_i^T t_i \right) \qquad (3.10)
$$

Table 3.1: Probabilities of data given hidden variables and parameters in different combinatorial values of indicator variables

q_{x_i}	q_{y_i}	$P(x_i\|t_i, q_{y_i}, q_{x_i}, \theta)$	$P(y_i\|t_i, q_{y_i}, q_{x_i}, \theta)$
0	0	$\mathcal{N}(Pt_i + \mu_x, \sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y, \sigma_y^2 I)$
	Δ_y	$\mathcal{N}(Pt_i + \mu_x, \sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y + \Delta_y, {\rho_y}^{-1}\sigma_y^2 I)$
	$-\Delta_y$	$\mathcal{N}(Pt_i + \mu_x, \sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y - \Delta_y, {\rho_y}^{-1}\sigma_y^2 I)$
Δ_x	0	$\mathcal{N}(Pt_i + \mu_x + \Delta_x, {\rho_x}^{-1}\sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y, \sigma_y^2 I)$
	Δ_y	$\mathcal{N}(Pt_i + \mu_x + \Delta_x, {\rho_x}^{-1}\sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y + \Delta_y, {\rho_y}^{-1}\sigma_y^2 I)$
	$-\Delta_y$	$\mathcal{N}(Pt_i + \mu_x + \Delta_x, {\rho_x}^{-1}\sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y - \Delta_y, {\rho_y}^{-1}\sigma_y^2 I)$
$-\Delta_x$	0	$\mathcal{N}(Pt_i + \mu_x - \Delta_x, {\rho_x}^{-1}\sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y, \sigma_y^2 I)$
	Δ_y	$\mathcal{N}(Pt_i + \mu_x - \Delta_x, {\rho_x}^{-1}\sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y + \Delta_y, {\rho_y}^{-1}\sigma_y^2 I)$
	$-\Delta_y$	$\mathcal{N}(Pt_i + \mu_x - \Delta_x, {\rho_x}^{-1}\sigma_x^2 I)$	$\mathcal{N}(Ct_i + \mu_y - \Delta_y, {\rho_y}^{-1}\sigma_y^2 I)$

$$\text{Term IV} \quad = (1 - \frac{|q_{x_i}|}{\Delta_x})log(1-\delta_x) + \frac{|q_{x_i}|}{\Delta_x}log(0.5\delta_x) \tag{3.11}$$

$$\begin{aligned}\text{Term V} &= \frac{|q_{y_i}| - q_{y_i}}{2\Delta_y}log(0.5\delta_y) + \frac{|q_{y_i}| + q_{y_i}}{2\Delta_y}log(0.5\delta_y) + (1 - \frac{|q_{y_i}|}{\Delta_y})log(1-\delta_y) \\ &= (1 - \frac{|q_{y_i}|}{\Delta_y})log(1-\delta_y) + \frac{|q_{y_i}|}{\Delta_y}log(0.5\delta_y)\end{aligned} \tag{3.12}$$

Having (3.9), $Q - function$ is built according to (3.7) by taking the conditional expectation on the hidden variables.

$$\mathbb{Q} = \mathop{\mathbb{E}}_{T,Q_Y,Q_X|X,Y,\theta^{old}} \Big(logP(X,Y,T,Q_Y,Q_X|\theta) \Big)$$

$$\mathbb{Q} = \mathop{\mathbb{E}}_{T,Q_Y,Q_X|X,Y,\theta^{old}} \Big(\sum_{i=1}^{n} \Big[\textcircled{\text{I}} + \textcircled{\text{II}} + \textcircled{\text{III}} + \textcircled{\text{IV}} + \textcircled{\text{V}} \Big] \Big)$$

$$= \int_{t_i} \sum_{q_{y_i}} \sum_{q_{x_i}} \Big(\sum_{i=1}^{n} \Big[\textcircled{\text{I}} + \textcircled{\text{II}} + \textcircled{\text{III}} + \textcircled{\text{IV}} + \textcircled{\text{V}} \Big] \times P(t_i, q_{y_i}, q_{x_i}|x_i, y_i, \theta^{old}) \Big)\, dt_i$$

$$\triangleq \mathbb{Q}_1 + \mathbb{Q}_2 + \mathbb{Q}_3 + \mathbb{Q}_4 + \mathbb{Q}_5 \tag{3.13}$$

Hidden variables of this problem are of two different types. Variable t_i is of continuous type and the indicators are of discrete type. This causes the expectation

to appear as in (3.13). The joint probability of hidden variables in (3.13) can be expanded as shown below in (3.14).

$$\begin{aligned} P(t_i, q_{y_i}, q_{x_i}|x_i, y_i, \theta^{old}) &= P(t_i|x_i, y_i, q_{y_i}, q_{x_i}, \theta^{old}) \\ &\quad \times P(q_{y_i}|y_i, \theta^{old}) \\ &\quad \times P(q_{x_i}|x_i, \theta^{old}) \end{aligned} \tag{3.14}$$

Therefore, based on the combinatorial values of q_{x_i} and q_{y_i} to take, there are nine combinations for each of the five terms of the $Q - function$. For a more detailed explanation about $Q - function$, readers are referred to Chapter 2. As an example, the first term of $Q - function$ will be elaborated below.

$$\begin{aligned} \mathbb{Q}_1 &= \int_{t_i} \sum_{q_{y_i}} \sum_{q_{x_i}} \left(\sum_{i=1}^{n} (\text{Term I}) \times P(t_i, q_{y_i}, q_{x_i}|x_i, y_i, \theta^{old}) \right) dt_i \\ &\triangleq \mathbb{Q}_1^{first} + \mathbb{Q}_1^{second} + \mathbb{Q}_1^{third} + \mathbb{Q}_1^{fourth} + \mathbb{Q}_1^{fifth} \\ &\quad + \mathbb{Q}_1^{sixth} + \mathbb{Q}_1^{seventh} + \mathbb{Q}_1^{eighth} + \mathbb{Q}_1^{ninth} \end{aligned} \tag{3.15}$$

where Term I, or $\textcircled{\text{I}}$, takes various expressions depending on the values of q_{x_i} and q_{y_i}, and is arranged in Table 3.2. For simplicity, the integration has been broken into nine terms as indicated in (3.15). For illustration, the first of these nine terms is given in (3.16). The remaining terms have similar expressions.

$$\begin{aligned} \mathbb{Q}_1^{first} &= \int_{t_i} \left(\sum_{i=1}^{n} (\text{Term I}) \right) \times P(t_i|x_i, y_i, q_{y_i} = 0, q_{x_i} = 0, \theta^{old}) \\ &\quad \times P(q_{y_i} = 0|y_i, \theta^{old}) \\ &\quad \times P(q_{x_i} = 0|x_i, \theta^{old})\, dt_i. \end{aligned} \tag{3.16}$$

Define $P_1 \triangleq P(q_{y_i} = 0, q_{x_i} = 0|x_i, y_i, \theta^{old})$, which is also equal to $P(q_{y_i} = 0|y_i, \theta^{old}) \times P(q_{x_i} = 0|x_i, \theta^{old})$. The proportions for each term of the $Q - function$ must add up to one, i.e., $\sum_{j=1}^{9} P_j = 1$.

3.3.2 Parameter Estimation

Finally, derivations of previous section lead to an expansion of the $Q-function$. Then the $Q - function$ should be maximized with respect to set of parameters (M-step).

Table 3.2: First term of complete data log-likelihood function in different combinatorial values of indicator variables

q_{x_i}	q_{y_i}	Term I
0	0 Δ_y $-\Delta_y$	$\frac{-1}{2}log((2\pi)^m\|\sigma_x^2 I\|) - \frac{1}{2}\Big((x_i - Pt_i - \mu_x)^T\sigma_x^{-2}(x_i - Pt_i - \mu_x)\Big)$
Δ_x	0 Δ_y $-\Delta_y$	$\frac{-1}{2}log((2\pi)^m\|{\rho_x}^{-1}\sigma_x^2 I\|) - \frac{1}{2}\Big((x_i - Pt_i - \mu_x - \Delta_x)^T\rho_x\sigma_x^{-2}(x_i - Pt_i - \mu_x - \Delta_x)\Big)$
$-\Delta_x$	0 Δ_y $-\Delta_y$	$\frac{-1}{2}log((2\pi)^m\|{\rho_x}^{-1}\sigma_x^2 I\|) - \frac{1}{2}\Big((x_i - Pt_i - \mu_x + \Delta_x)^T\rho_x\sigma_x^{-2}(x_i - Pt_i - \mu_x + \Delta_x)\Big)$

This optimization process is shown in (3.17).

$$\begin{aligned}\theta &= \underset{\theta}{\text{argmax}} \underset{T,Q_Y,Q_X|X,Y,\theta^{old}}{\mathbb{E}} \Big(logP(X,Y,T,Q_Y,Q_X|\theta)\Big) \\ &= \underset{\theta}{\text{argmax}}\ \mathbb{Q}\end{aligned} \tag{3.17}$$

This maximization problem can be broken into a set of equations as in (3.18). This is because terms of the $Q-function$ can be arranged in a way such that parameters are separated. This fact simplifies this optimization. Solving a series of equation in (3.18), yields the parameters update equations as shown in equations (3.19) to (3.26).

$$\begin{array}{llll}
\text{P} : & \frac{\partial\mathbb{Q}_1}{\partial P} = 0\,, & \text{C} : & \frac{\partial\mathbb{Q}_2}{\partial C} = 0 \\
\sigma_x^2 : & \frac{\partial\mathbb{Q}_1}{\partial\sigma_x^2} = 0\,, & \sigma_y^2 : & \frac{\partial\mathbb{Q}_2}{\partial\sigma_y^2} = 0 \\
\mu_x : & \frac{\partial\mathbb{Q}_1}{\partial\mu_x} = 0\,, & \mu_y : & \frac{\partial\mathbb{Q}_2}{\partial\mu_y} = 0 \\
\delta_x : & \frac{\partial\mathbb{Q}_4}{\partial\delta_x} = 0\,, & \delta_y : & \frac{\partial\mathbb{Q}_5}{\partial\delta_y} = 0 \\
\Delta_x : & \frac{\partial\mathbb{Q}_1}{\partial\delta_x} + \frac{\partial\mathbb{Q}_4}{\partial\delta_x} = 0\,, & \Delta_y : & \frac{\partial\mathbb{Q}_2}{\partial\delta_y} + \frac{\partial\mathbb{Q}_5}{\partial\delta_y} = 0 \\
\rho_x : & \frac{\partial\mathbb{Q}_1}{\partial\rho_x} = 0\,, & \rho_y : & \frac{\partial\mathbb{Q}_2}{\partial\rho_y} = 0
\end{array} \tag{3.18}$$

$$
\begin{aligned}
P^{new} \;=\; & \Bigg[\sum_{i=1}^{n}\Bigg(2(x_i-\mu_x)\big(P_1 \underset{00}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=0,\theta^{old})^T \\
& \qquad +P_2 \underset{0\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old})^T \\
& \qquad +P_3 \underset{0-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old})^T\big) \\
& +2\rho_x(x_i-\mu_x-\Delta_x)\big(P_4 \underset{\Delta 0}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=0,\theta^{old})^T \\
& \qquad +P_5 \underset{\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old})^T \\
& \qquad +P_6 \underset{\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})^T\big) \\
& +2\rho_x(x_i-\mu_x+\Delta_x)\big(P_7 \underset{-\Delta 0}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=0,\theta^{old})^T \\
& \qquad +P_8 \underset{-\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old})^T \\
& \qquad +P_9 \underset{-\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})^T\big)\Bigg)\Bigg] \\
& \times\Bigg[\sum_{i=1}^{n}\Bigg(P_1 \underset{00}{Coeff}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=0,\theta^{old}) \\
& \qquad +P_2 \underset{0\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old}) \\
& \qquad +P_3 \underset{0-\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old}) \\
& \qquad +\rho_x P_4 \underset{\Delta 0}{Coeff}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=0,\theta^{old}) \\
& \qquad +\rho_x P_5 \underset{\Delta\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old}) \\
& \qquad +\rho_x P_6 \underset{\Delta-\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old}) \\
& \qquad +\rho_x P_7 \underset{-\Delta 0}{Coeff}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=0,\theta^{old}) \\
& \qquad +\rho_x P_8 \underset{-\Delta\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old}) \\
& \qquad +\rho_x P_9 \underset{-\Delta-\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})\Bigg)\Bigg]^{-1}
\end{aligned}
\tag{3.19}
$$

where P_j defines probability of different cases of indicators from first to ninth case as

shown in Table 3.2, and

$$\begin{aligned}
\underset{*\$}{Coeff}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) &= \underset{*\$}{S}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
&\quad + \underset{*\$}{S}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T \\
&\quad + 2\,\underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
&\quad \times \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T, \qquad (3.20)
\end{aligned}$$

where $\underset{*\$}{S}$ is as shown below in which the asterisk or dollar sign subfixes indicate either of the nine cases for $q_{y_i} = *$ and $q_{y_i} = \$$.

$$\begin{aligned}
\underset{*\$}{S}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) &= \underset{*\$}{E}(t_i t_i^T|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
&\quad - \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
&\quad \times \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T, \qquad (3.21)
\end{aligned}$$

$$\begin{aligned}
\sigma_x^{2\,new} &= \frac{1}{n\,m}\sum_{i=1}^{n}\Bigg(P_1\,\underset{00}{A}^0(t_i|x_i, y_i, q_{x_i} = 0, q_{y_i} = 0, \theta^{old}) \qquad (3.22) \\
&\quad + P_2\,\underset{0\Delta}{A}^0(t_i|x_i, y_i, q_{x_i} = 0, q_{y_i} = \Delta_y, \theta^{old}) \\
&\quad + P_3\,\underset{0-\Delta}{A}^0(t_i|x_i, y_i, q_{x_i} = 0, q_{y_i} = -\Delta_y, \theta^{old}) \\
&\quad + \rho_x P_4\,\underset{\Delta 0}{A}^{\Delta}(t_i|x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = 0, \theta^{old}) \\
&\quad + \rho_x P_5\,\underset{\Delta\Delta}{A}^{\Delta}(t_i|x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = \Delta_y, \theta^{old}) \\
&\quad + \rho_x P_6\,\underset{\Delta-\Delta}{A}^{\Delta}(t_i|x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = -\Delta_y, \theta^{old}) \\
&\quad + \rho_x P_7\,\underset{-\Delta 0}{A}^{-\Delta}(t_i|x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = 0, \theta^{old}) \\
&\quad + \rho_x P_8\,\underset{-\Delta\Delta}{A}^{-\Delta}(t_i|x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = \Delta_y, \theta^{old}) \\
&\quad + \rho_x P_9\,\underset{-\Delta-\Delta}{A}^{-\Delta}(t_i|x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = -\Delta_y, \theta^{old})\Bigg).
\end{aligned}$$

where, $\underset{*\$}{A}$s for each case are as formulated in Appendix I.

$$\mu_x^{new} = \frac{1}{n}\sum_{i=1}^{n}\Big(- P_1(-x_i + P\underset{00}{E}(t_i|x_i, y_i, q_{x_i}=0, q_{y_i}=0, \theta^{old})) \tag{3.23}$$
$$-P_2(-x_i + P\underset{0\Delta}{E}(t_i|x_i, y_i, q_{x_i}=0, q_{y_i}=\Delta_y, \theta^{old}))$$
$$-P_3(-x_i + P\underset{0-\Delta}{E}(t_i|x_i, y_i, q_{x_i}=0, q_{y_i}=-\Delta_y, \theta^{old}))$$
$$-P_4\rho_x(\Delta_x - x_i + P\underset{\Delta 0}{E}(t_i|x_i, y_i, q_{x_i}=\Delta_x, q_{y_i}=0, \theta^{old}))$$
$$-P_5\rho_x(\Delta_x - x_i + P\underset{\Delta\Delta}{E}(t_i|x_i, y_i, q_{x_i}=\Delta_x, q_{y_i}=\Delta_y, \theta^{old}))$$
$$-P_6\rho_x(\Delta_x - x_i + P\underset{\Delta-\Delta}{E}(t_i|x_i, y_i, q_{x_i}=\Delta_x, q_{y_i}=-\Delta_y, \theta^{old}))$$
$$-P_7\rho_x(-\Delta_x - x_i + P\underset{-\Delta 0}{E}(t_i|x_i, y_i, q_{x_i}=-\Delta_x, q_{y_i}=0, \theta^{old}))$$
$$-P_8\rho_x(-\Delta_x - x_i + P\underset{-\Delta\Delta}{E}(t_i|x_i, y_i, q_{x_i}=-\Delta_x, q_{y_i}=\Delta_y, \theta^{old}))$$
$$-P_9\rho_x(-\Delta_x - x_i + P\underset{-\Delta-\Delta}{E}(t_i|x_i, y_i, q_{x_i}=-\Delta_x, q_{y_i}=-\Delta_y, \theta^{old}))\Big).$$

$$\delta_x^{new} = \frac{\sum_{i=1}^{n}(P_4 + P_5 + P_6 + P_7 + P_8 + P_9)}{n}, \tag{3.24}$$

$$\Delta_x^{new} = (\frac{1}{\sum_{i=1}^{n}(P_4 + P_5 + P_6 + P_7 + P_8 + P_9)}) \tag{3.25}$$
$$\times\sum_{i=1}^{n}\Big(P_4(x_i - \mu_x - P\underset{\Delta 0}{E}(t_i|x_i, y_i, q_{x_i}=\Delta_x, q_{y_i}=0, \theta^{old}))$$
$$+P_5(x_i - \mu_x - P\underset{\Delta\Delta}{E}(t_i|x_i, y_i, q_{x_i}=\Delta_x, q_{y_i}=\Delta_y, \theta^{old}))$$
$$+P_6(x_i - \mu_x - P\underset{\Delta-\Delta}{E}(t_i|x_i, y_i, q_{x_i}=\Delta_x, q_{y_i}=-\Delta_y, \theta^{old}))$$
$$-P_7(x_i - \mu_x - P\underset{-\Delta 0}{E}(t_i|x_i, y_i, q_{x_i}=-\Delta_x, q_{y_i}=0, \theta^{old}))$$
$$-P_8(x_i - \mu_x - P\underset{-\Delta\Delta}{E}(t_i|x_i, y_i, q_{x_i}=-\Delta_x, q_{y_i}=\Delta_y, \theta^{old}))$$
$$-P_9(x_i - \mu_x - P\underset{-\Delta-\Delta}{E}(t_i|x_i, y_i, q_{x_i}=-\Delta_x, q_{y_i}=-\Delta_y, \theta^{old}))\Big),$$

$$\rho_x^{new} = m\sigma_x^2\Big[\sum_{i=1}^{n}(P_4+P_5+P_6+P_7+P_8+P_9)\Big] \tag{3.26}$$
$$\times\Big[\sum_{i=1}^{n}\Big(P_4\,\underset{\Delta 0}{A^{\Delta}}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=0,\theta^{old})$$
$$+P_5\,\underset{\Delta\Delta}{A^{\Delta}}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old})$$
$$+P_6\,\underset{\Delta-\Delta}{A^{\Delta}}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})$$
$$+P_7\,\underset{-\Delta 0}{A^{-\Delta}}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=0,\theta^{old})$$
$$+P_8\,\underset{-\Delta\Delta}{A^{-\Delta}}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old})$$
$$+P_9\,\underset{-\Delta-\Delta}{A^{-\Delta}}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})\Big)\Big]^{-1}.$$

The remaining parameters that are related to output equation are derived similarly. Readers are referred to Appendix II for the final expressions.

3.3.3 Posterior Distributions of Hidden Variables

After solving the model for its parameters in Section 3.3.2, the parameter update equations are obtained. These update equations should go through a series of iterations until convergence. These equations contain the expected value with respect to the posterior distributions of the hidden variables. Therefore, the posterior distributions of the hidden variables must be derived.

In this section the above mentioned posteriors of hidden variables are derived. For the continuous hidden variable t_i based on the assumption that the posteriors of all combinations will have Gaussian distributions, and their sufficient statistics are mean and variances. These two statistics are seen in the parameter update equations as $\underset{*\$}{E}(t_i|x_i,y_i,q_{x_i}=*,q_{y_i}=\$,\theta^{old})$ or $\underset{*\$}{E}(t_it_i^T|x_i,y_i,q_{x_i}=*,q_{y_i}=\$,\theta^{old})$ [27]. For the discrete hidden variables, posterior probabilities, defined as P_js, are evaluated during the iterations by having their prior probability based on δ values in each case, which is the categorical distribution shown in (3.6). To derive the sufficient statistics, we start from the joint probability of the hidden variables. As shown in (3.14), the three probabilities constitute the joint probability. These three terms are given as follows

in (3.27), (3.28) and (3.29) using *Bayes' rule*.

$$P(t_i|q_{y_i}, q_{x_i}, x_i, y_i, \theta^{old}) \;\propto\; P(x_i, y_i|t_i, q_{y_i}, q_{x_i}, \theta^{old})\; P(t_i|q_{y_i}, q_{x_i}, \theta^{old}) \tag{3.27}$$

$$P(q_{x_i}|x_i, y_i, \theta^{old}) \;\propto\; P(x_i, y_i|q_{x_i}, \theta^{old})\; P(q_{x_i}|\theta^{old}) \tag{3.28}$$

$$P(q_{y_i}|x_i, y_i, \theta^{old}) \;\propto\; P(x_i, y_i|q_{y_i}, \theta^{old})\; P(q_{y_i}|\theta^{old}) \tag{3.29}$$

Feeding the prior distributions into (3.27), (3.28) and (3.29), considering different combinations according to q_{x_i} and q_{y_i}, equations for the conditional expected values of latent variable t_i and $t_i t_i^T$ are obtained for different combinations of indicator variables. Equations for the first case of both latent variables are shown below as an example and the other eight cases are provided in Appendices III and IV, respectively. Following the similar approach posteriors for the discrete hidden variables are obtained from (3.28) and (3.29). Since the indicators are independent, both of them can be brought into a joint posterior probability. For instance, in the first possible case, $P_1 = P(q_{x_i} = 0, q_{y_i} = 0|x_i, y_i, \theta^{old})$, and based on *Bayes' rule* $P_1 \propto P(x_i y_i|q_{x_i} = 0, q_{y_i} = 0)P(q_{y_i} = 0|\theta^{old})P(q_{x_i} = 0|\theta^{old})$, where $P(x_i y_i|q_{x_i} = 0, q_{y_i} = 0)$ is as shown in (3.32), and $P(q_{y_i} = 0|\theta^{old}) = 1 - \delta_y$, $P(q_{x_i} = 0|\theta^{old}) = 1 - \delta_x$. The same derivations are performed for the other eight cases.

$$\begin{aligned} \underset{00}{E}(t_i|x_i, y_i, q_{x_i} = 0, q_{y_i} = 0, \theta^{old}) \;=\; & (\sigma_x^{-2} P^T P + \sigma_y^{-2} C^T C + I)^{-1} \\ & \times(\sigma_x^{-2} P^T (x_i - \mu_x) + \sigma_y^{-2} C^T (y_i - \mu_y)) \end{aligned} \tag{3.30}$$

$$\begin{aligned} \underset{00}{E}(t_i t_i^T|x_i, y_i, q_{x_i} = 0, q_{y_i} = 0, \theta^{old}) \;=\; & (\sigma_x^{-2} P^T P + \sigma_y^{-2} C^T C + I)^{-1} \\ & + \underset{00}{E}(t_i|x_i, y_i, q_{x_i} = 0, q_{y_i} = 0, \theta^{old}) \\ & \times \underset{00}{E}(t_i|x_i, y_i, q_{x_i} = 0, q_{y_i} = 0, \theta^{old})^T \end{aligned} \tag{3.31}$$

$$P(x_i y_i|q_{x_i} = 0, q_{y_i} = 0) \sim \; \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} PP^T + \sigma_x^2 I & PC^T \\ CP^T & CC^T + \sigma_y^2 I \end{bmatrix} \right) \tag{3.32}$$

$$
\begin{aligned}
\hat{t}_i = \; & E(t_i|x_i,\theta^{old}) \qquad (3.33)\\
= \; & \sum_{q_{y_i}}\sum_{q_{x_i}} E(t_i|x_i,q_{y_i},q_{x_i},\theta^{old})\; P(q_{y_i}|x_i,\theta^{old})\; P(q_{x_i}|x_i,\theta^{old})\\
= \; & \underset{00}{E}(t_i|x_i,q_{y_i}=0,q_{x_i}=0,\theta^{old})\; P(q_{y_i}=0,q_{x_i}=0|x_i,\theta^{old})\\
& + \underset{0\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old})\; P(q_{x_i}=0,q_{y_i}=\Delta_y|x_i,\theta^{old})\\
& + \underset{0-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old})\; P(q_{x_i}=0,q_{y_i}=-\Delta_y|x_i,\theta^{old})\\
& + \underset{\Delta 0}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=0,\theta^{old})\; P(q_{x_i}=\Delta_x,q_{y_i}=0|x_i,\theta^{old})\\
& + \underset{\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old})\; P(q_{x_i}=\Delta_x,q_{y_i}=\Delta_y|x_i,\theta^{old})\\
& + \underset{\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})\; P(q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y|x_i,\theta^{old})\\
& + \underset{-\Delta 0}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=0,\theta^{old})\; P(q_{x_i}=-\Delta_x,q_{y_i}=0|x_i,\theta^{old})\\
& + \underset{-\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old})\; P(q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y|x_i,\theta^{old})\\
& + \underset{-\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})\; P(q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y|x_i,\theta^{old})
\end{aligned}
$$

3.3.4 Predictions

Inferring via a soft sensor or prediction of variables of interest as in [24], refers to inferring about a quality variable based on the developed model (soft sensor) incorporating some other variables, which are easier to be measured on-line. In this chapter, we use the proposed RPPCA based model developed in Section 2.4 to predict a variable of interest. This will be elaborated further in Section 3.4 with two examples. This probabilistic model needs posterior distribution for hidden variables for doing such inference, since these hidden variables are not known. Determination of the posterior distribution of hidden variables has been presented in Section 3.3.3. To take into account the uncertainty, the *law of total expectation* is applied as in (3.33).

Having the estimation of the latent variable and using the generative model as in (3.3), the output prediction can be obtained as shown in (3.34). The error for this prediction ϵ, which is the difference between predicted and real values of the output variable, is obtained from (3.35). Along with this measure, there are many other performance measures for evaluating the goodness of fit and thus the prediction. R-squared and mean squared error (MSE) are the other often used measures. Correlation between predicted and real values is also checked. Root mean square error (RMSE)

as in (3.36) is a well-known measure to give a quantitative sense in comparisons.

$$\hat{y}_i = C\hat{t}_i + \hat{\mu}_y \tag{3.34}$$

$$\epsilon = \hat{y} - y. \tag{3.35}$$

RMSE is defined as

$$RMSE \triangleq \sqrt{\frac{\sum_{i=1}^{n'} \|\hat{y}_i - y_i\|^2}{n'}}, \tag{3.36}$$

where, n' is the total number of test samples and $\hat{y}$ and y are predicted and real output values, respectively.

3.4 Case Studies

In this section, two examples are discussed to evaluate the performance of our developed algorithms. In the first case study, a data set is generated by a generative PPCA based regression model with fixed parameters; in the second case study, a data set from an industrial steam-assisted gravity drainage (SAGD) process plant is employed for evaluation. In both case studies, we illustrate the improved performance of the robust model presented in Section 3.3 compared to that of the regular PPCA based model which has the single Gaussian noise distribution. The results will demonstrate that the developed model is insensitive to outliers in terms of parameter estimation and thus results in better prediction performance.

The two important contributions of this chapter are to evaluate the role of high leverage points and assess the importance of large random errors sitting far from the center of noise distribution, i.e. location type of outliers. Two comparison studies are done on two data sets to investigate the importance of our new contributions to robust identification, respectively.

In the first comparison study, we will illustrate that considering outliers in both input and output is important. In the second comparison study, we demonstrate the performance of our proposed robust approach through comparison with the existing approach when outliers occur only in the output. By this study we will illustrate the

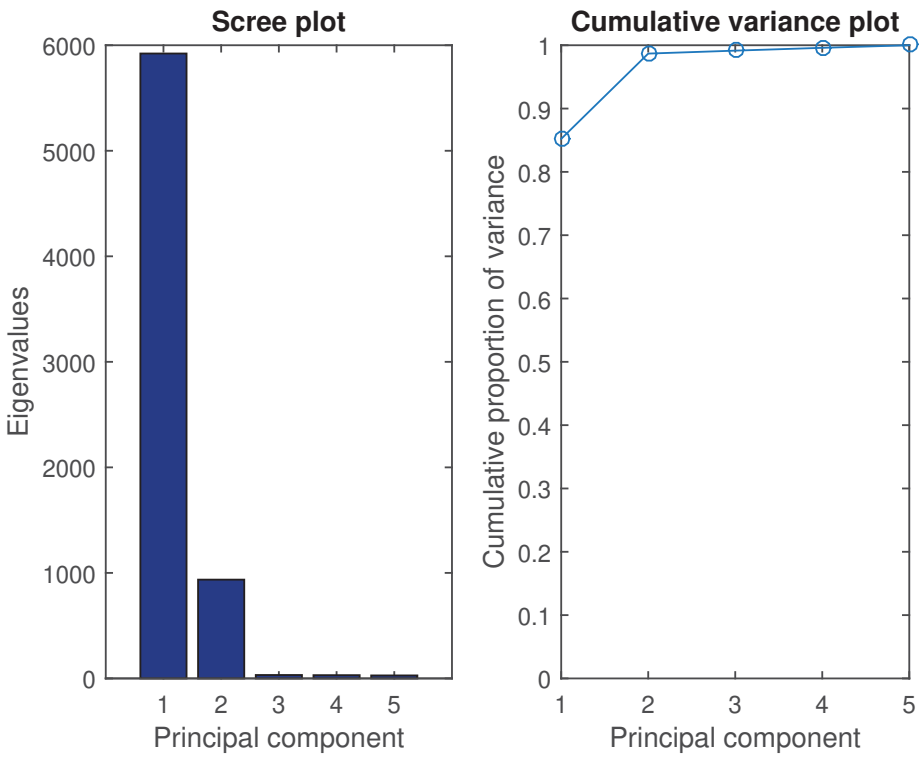


Figure 3.2: Scree plot (left) and cumulative variance explained plot (right) for generated data used in numerical example, obtained through PCA

importance of having appropriate outlier characterization. The robust approach in this chapter is designed to be insensitive to large random outliers sitting far from the center of the noise distribution. This is the case with a jammed instrument and is a common type of outlier seen in processes.

3.4.1 Case I: Numerical Example

For model performance assessment through a numerical example, a data set has been generated using the model (3.3). The set of data contains 1000 observations from 5 input variables and 1 output variable. Two principal components are considered for generating the data set. The loading matrix is set to be $P_{(m=5)\times(k=2)} = [40, 10; 20, 30; 15, 20; 20, 40; 40, 15]$, regression coefficient matrix is $C_{(r=1)\times(k=2)} = [10, 20]$ and input/output mean vectors are $\mu_x = [1, 2, 3, 4, 5]$ and $\mu_y = [0]$, respectively. Input/output noises are distributed as $e_i \sim \mathcal{N}(0, 30I)$, $f_i \sim \mathcal{N}(0, 30I)$, respectively. Each latent variable vector has the standard Gaussian distribution as $t_{i_{k\times 1}} \sim \mathcal{N}(0, I)$,

making latent variable matrix, $T_{n\times k} = [t_1, \cdots, t_n]^T$, to become a set of $i.i.d$ vectors.

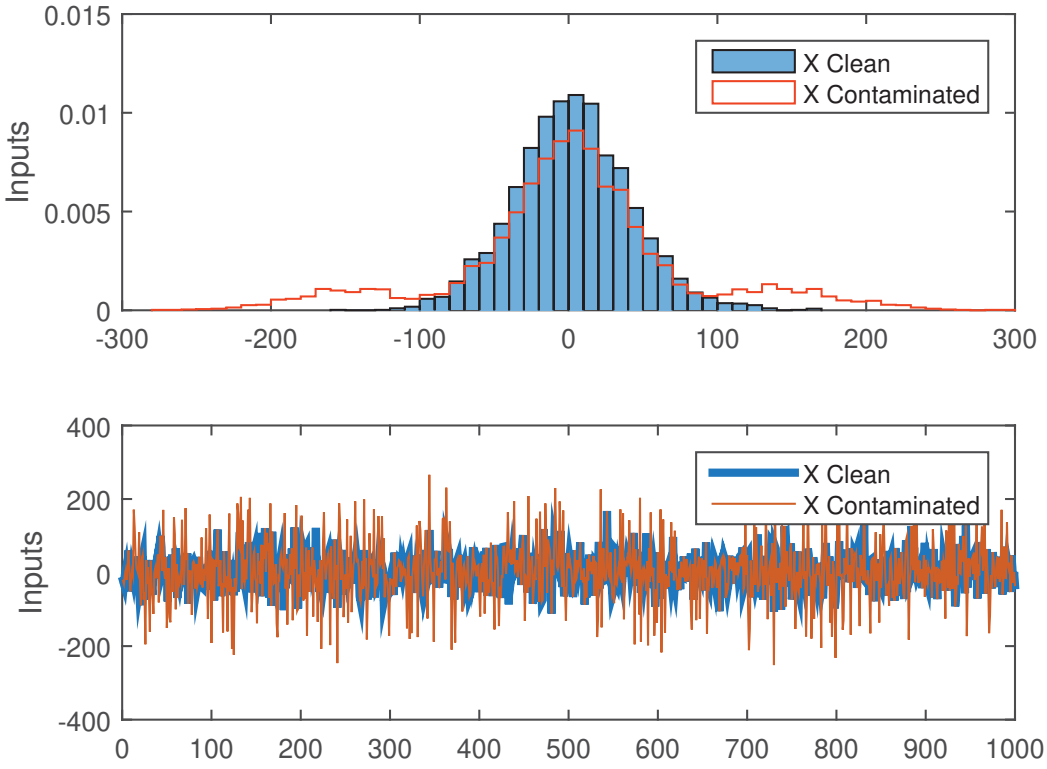


Figure 3.3: Input data before and after contamination (x_5)

To employ PPCA based regression, a rough estimate of the number of latent variables is needed. In this step, PCA is performed on input data. Based on the scree plot shown in Figure 3.2, k is chosen to be 2. To see how robust the proposed RPPCA based regression model is, a series of data points are replaced with outlying values by adding a large random error to the observed data points. The RPPCA based model is expected to show better predictions compared to a regular PPCA, which is not designed to be robust to observations with outliers. To make the data set contaminated by such random errors, a portion of the data set is randomly selected and replaced by outliers. This is done by means of adding location outliers to the data points. Here, 20% of the points in both input and output data sets are selected to be replaced with outliers. The locations of the two symmetric counterparts of the mixture distribution are chosen to be far away from the center component's mean, and these counterparts are set to have an inflated variance with an inflation factor of

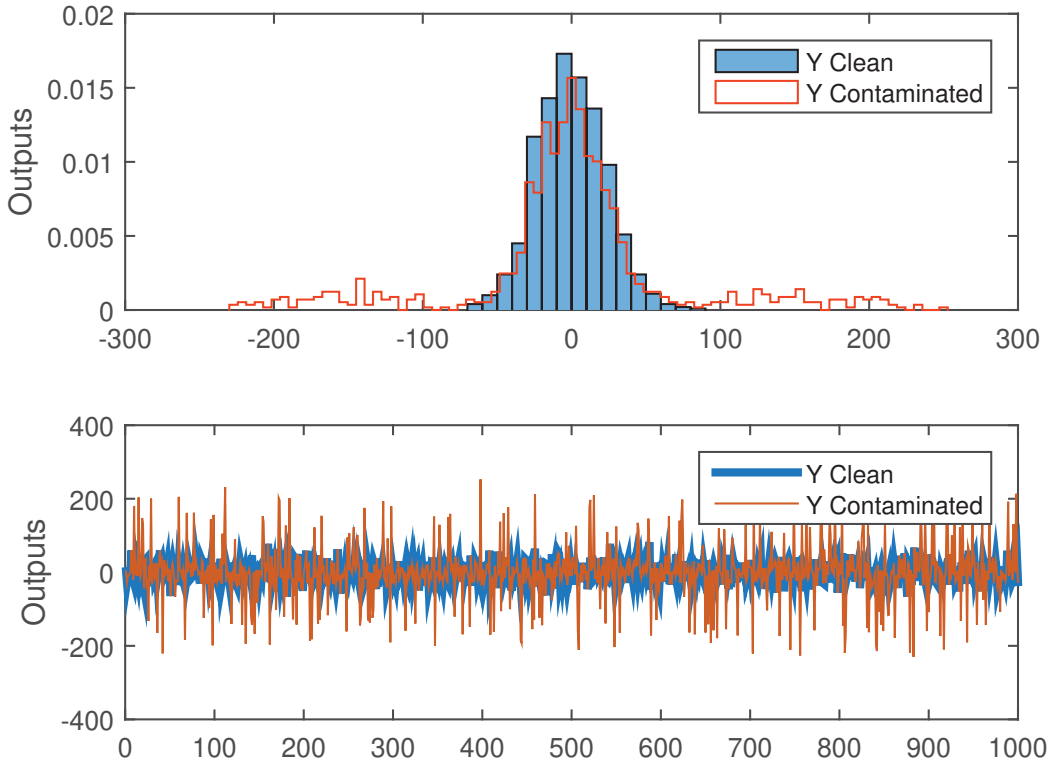


Figure 3.4: Output data before and after contamination

0.9. This setting is considering two far and distinguished Gaussian noise components with symmetric means of distributions for the outliers that do not overlap with the main component. This is applied to both input and output data, meaning that both data sets are considered to have this Gaussian general mixture random error. For simplicity of presentation, here only one of the inputs, x_5, is contaminated. Contaminated data are shown in Figures 3.3 and 3.4. The top panel shows the histograms before and after contamination, while the bottom panel shows the data time series before and after being contaminated.

Figures 3.5 and 3.6 show the performances of PPCA based model on the generated data without outliers and contaminated data with outliers, respectively. It is seen that the performance is affected by the contamination. However, the RPPCA based model of this chapter is insensitive to outliers. Results of this case study are summarized in Table 3.3. Figure 3.7 shows the performance of the RPPCA based model on contaminated data sets. The convergence of the robust model parameters is also

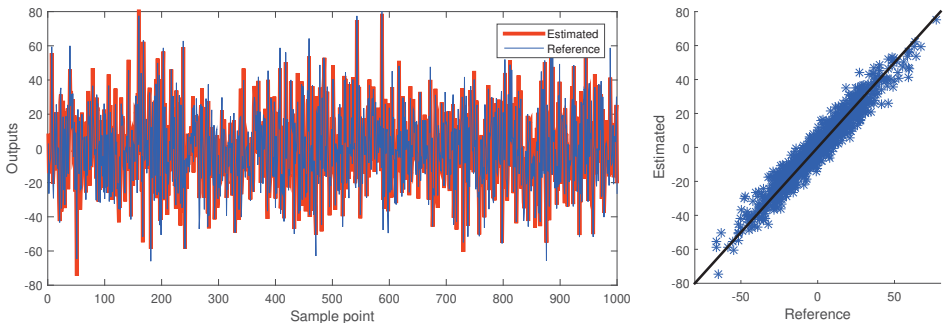


Figure 3.5: Prediction performance of PPCA regression model on generated data without outliers

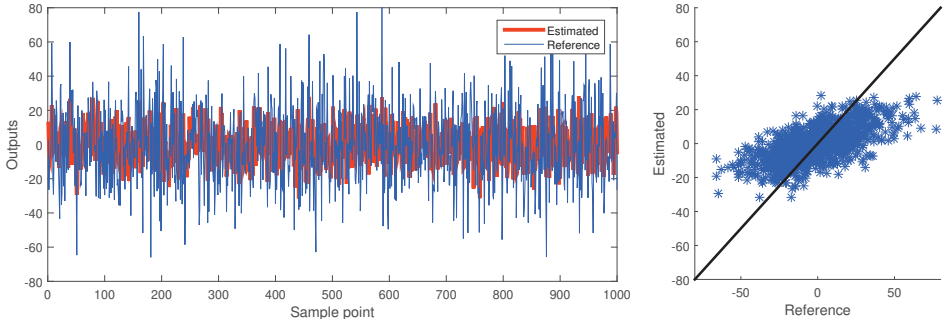


Figure 3.6: Prediction performance of PPCA regression model on contaminated generated data

Table 3.3: Prediction performance of regular and robust models in numerical example

	Regular data	Contaminated data	
	PPCA	PPCA	RPPCA
R^2	0.9399	0.4010	0.9398
$RMSE$	5.6960	17.9759	5.6981

shown in Figure 3.8.

Effect of leverage points

In this section we illustrate two features of the developed model. First, we explore the impact of input outliers. The approach taken in this chapter is developing a robust model to deal with input outliers, or leverage points, as well as output outliers. For illustration, the input data is contaminated by adding large random values to

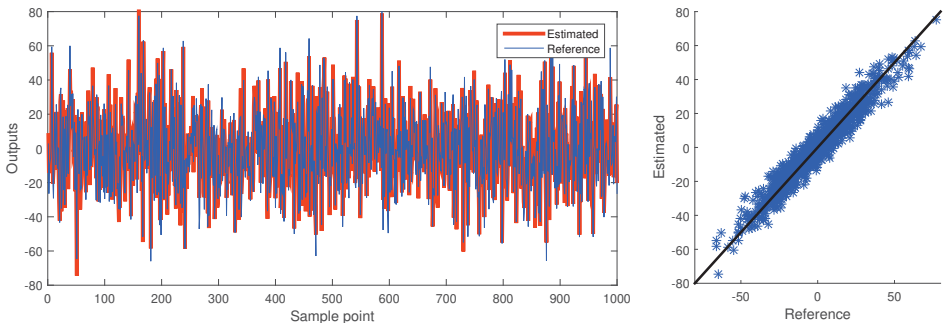


Figure 3.7: Prediction performance of RPPCA regression model on contaminated generated data

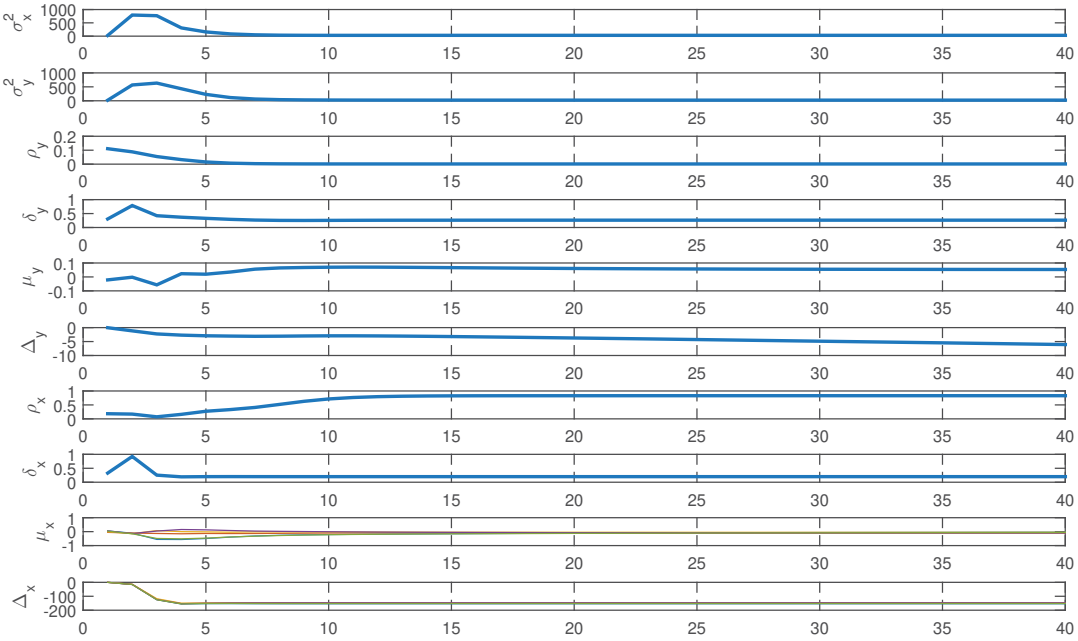


Figure 3.8: Parameter convergence for RPPCA model in the first case study

a portion of the data set. Output data is not contaminated to magnify the effect of input outliers on the model performance first. Contamination of inputs could be done in many different ways. We chose to contaminate x_5, as an example, with two distinguished side distributions. 35% of the data set is replaced by random outliers sitting far away from the center with inflation factor of 0.9. Refer to Figure 3.9. The performances of the regular PPCA and the RPPCA based model presented in this chapter are shown in Table 3.4 which again demonstrates superiority of the robust

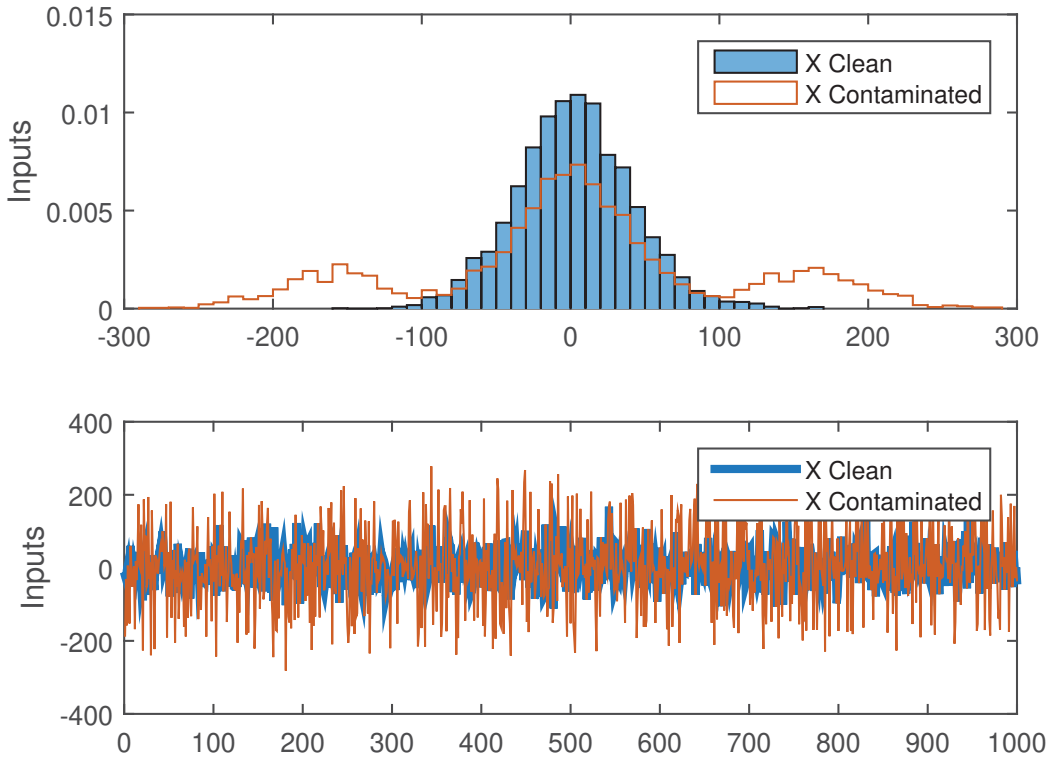


Figure 3.9: Input data before and after contamination (x_5) in comparison study I on generated data

Table 3.4: Prediction performance comparison for regular and developed robust model in presence of leverage points only

	Regular data	Contaminated data	
	PPCA	PPCA	RPPCA
R^2	0.9399	0.3196	0.9395
$RMSE$	5.6960	19.1572	5.7106

algorithm.

3.4.2 Case II: Industrial Application

This section is devoted to the application of the developed model to a set of data collected from a SAGD operation in Canada. This is a real industrial data set and has been normalized beforehand for propriety.

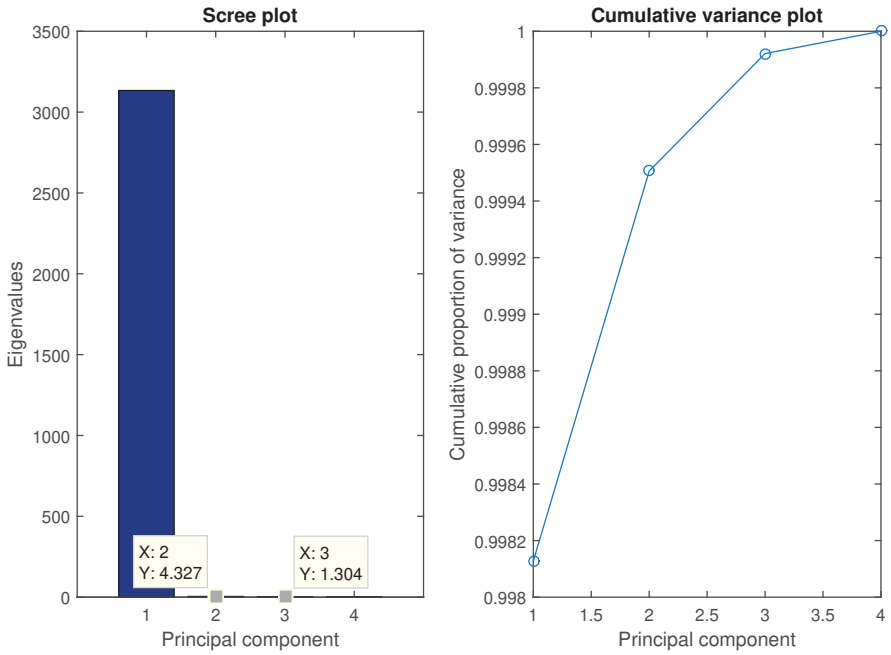


Figure 3.10: Scree plot (left) and cumulative variance explained plot (right) for industrial data used in second case study, obtained through PCA

SAGD process

SAGD is an enhanced in-situ oil recovery technology for bitumen and heavy crude oil production. In this technology two horizontal wells, known as well pairs, are drilled into the reservoir. Typically, these two wells have a vertical distance of about four to six meters. Injection of high pressure steam into the well located above will heat the oil which results in a reduction of viscosity. Heating via steam generates a steam chamber underground. When the viscosity of oil drops, it flows downward to the collector well due to gravity. Collected oil will need to be pumped up to the ground level. This technology is used extensively to extract oil from oil sands reservoirs which has majority of its deposits deep-seated underground making it infeasible to be mined by conventional open-pit mining. It is worth mentioning that much of the expected future production growth in the Canadian oil sands is predicted to be from this technology [30, 68].

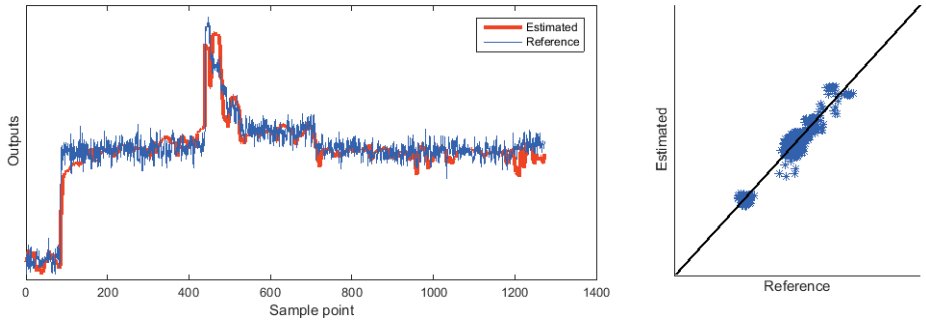


Figure 3.11: Prediction performance of PPCA regression model on industrial data

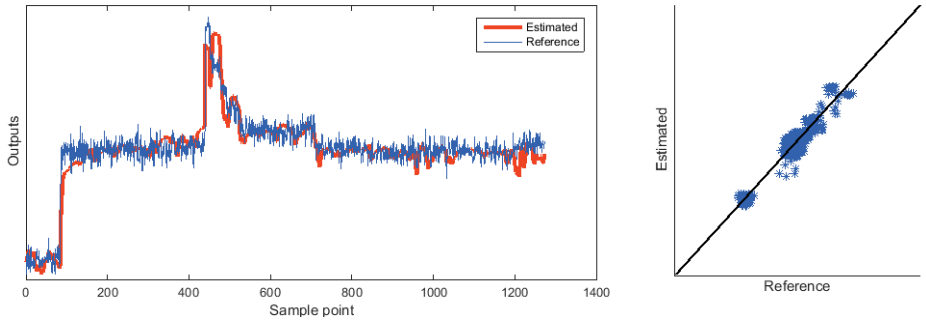


Figure 3.12: Prediction performance of RPPCA regression model on industrial data

Robust probabilistic modeling of SAGD process data

In this section, robust PPCA based soft sensor is used to predict the operation fluid flow rate as a quality variable of interest. Accurate measurement of this variable is required for better operation of downstream units. However, it is rarely available. A soft sensing approach can be used to provide a continuous and reasonably accurate estimate for this measurement based on some other available process variable measurements.

The available historical data from this process are a set of flow rates, pressures, and temperatures for the injection and production well pair, which are constructed by a $10 - min$ average data recorded for a season in year 2014. Four of the above mentioned variables which are highly correlated with the product oil flow rate are selected for further analysis; among them three principal components are chosen to represent the process. This data set consists of a total of 4 inputs and 1 output variables that have been recorded in about 52500 sample points from which a set of 10906 sample points have been selected for the analysis in which the process shows

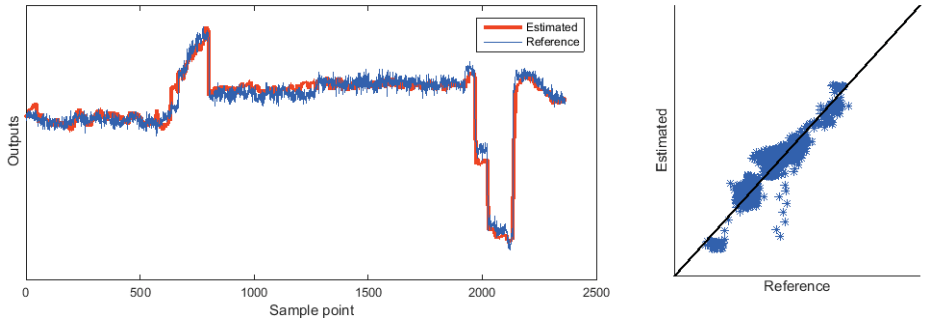


Figure 3.13: Self validation of PPCA regression model on industrial data

a linear and time-invariant behavior. This data set has been used in [75] for a scale Gaussian mixture based robust algorithm.

In this case study, the performance of the two models are compared as follows. Data has been divided to training and test sets by choosing the first 70% of samples as the training set, and the rest as the test set. Regular model is trained with the training data set from the raw data. The same is done for robust model to obtain the parameter set. Accordingly, the cross validation of the two models are checked. Cross validation for the regular model is done by the test data sets. Cross validation for the robust model is checked by feeding in the clean input test data and comparing its predictions with the clean output test data. The rational behind this comparison is that the robust model is expected to be trained for parameters such that the understanding of the process, which is reflected in model parameters, is robust. This is suggesting that the parameter estimation is not affected by the outlying observations in the input and output data set. Thus, the cross-validation of the robust model should be done as stated above.

To understand the effect of outliers on prediction, raw industrial data was used to train the models. Comparison of cross-validations shows that the robust model outperforms the regular one.The performance of the regular PPCA based model and the robust PPCA based model with three principal components on this industrial data set are shown in Figure 3.11 and Figure 3.12, respectively. The performances are based on cross-validation. For propitiatory reasons, all y-axis values have been removed.

As per the fit (self-validation), the regular model performs similar when compared

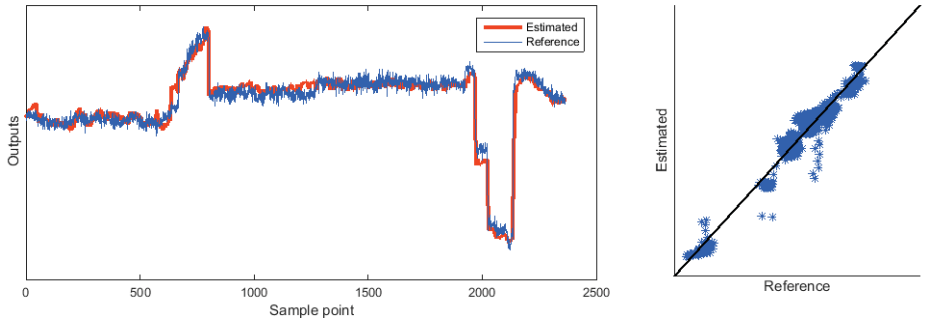


Figure 3.14: Self validation of RPPCA regression model on industrial data

to the robust one, as shown in Figure 3.13 and Figure 3.14, respectively. This is because of the fact that the industrial data in hand has not many severe outliers. Thus, we tend to manually impose some outliers on the raw data that we have and test how the robust model could catch up when the regular model is not able to perform as it did on the namely uncontaminated data. Table 3.5 summarizes this result.

Table 3.5: Prediction performance of regular and robust models in industrial example; Self = Self validation, Cross = Cross validation

	Regular model		Robust model	
	Self	Cross	Self	Cross
R^2	0.9489	0.8877	0.9489	0.8880
$RMSE$	0.5246	0.4359	0.5247	0.4354
$MAPE$	0.0155	0.0110	0.0155	0.0110

To carry out the contamination, we have tried a 5% and a 10% input and output contamination task separately to show the ability of the robust model in a kind of pressure test. In the contamination task, randomly selected data points are replaced with large outliers far from the center of the data distribution. To have visibly distinct Gaussian distributions at the two sides of the main noise distribution, we have chosen the locations to be far from the mean values. This ensures that the case of location outliers is imposed on the data. The values are calculated based on the estimated noise variances from the regular PPCA based regression to make sure the Gaussian bells are sitting far enough from each other so that they do not overlap. This

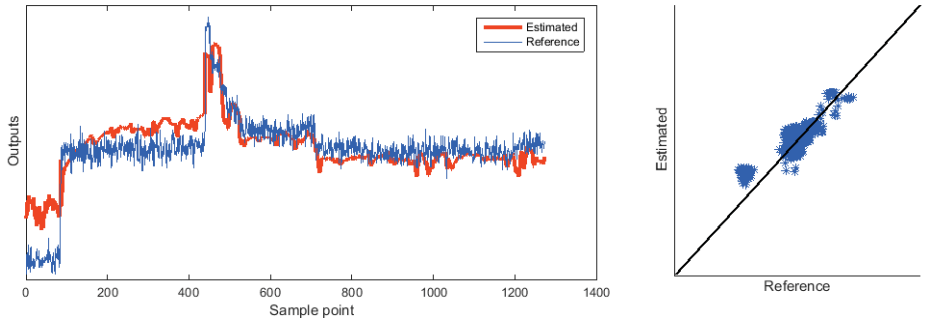


Figure 3.15: Prediction performance of PPCA regression model on 5% contaminated industrial data

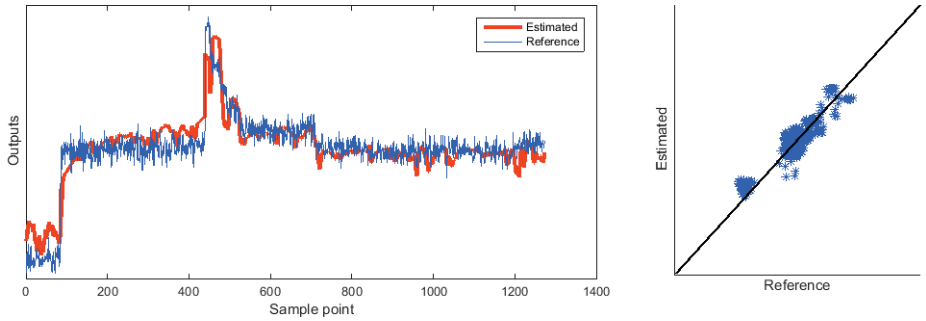


Figure 3.16: Prediction performance of RPPCA regression model on 5% contaminated industrial data

sort of outlier occurs in the presence of process measurements that do not conform to the technological extent of the measuring devices [45]. The symmetric location counterparts in the above-mentioned mixture have a variance inflation factor of 0.9. Prediction performances of the regular and robust model on the 5% contaminated sets are presented in Figure 3.15 and Figure 3.16, respectively. Table 3.6 summarizes this result.

Table 3.6: Prediction performance of regular and robust models in industrial example by 5% contamination in inputs and output

	Regular model	Robust model
R^2	0.6037	0.8234
$RMSE$	1.6115	0.5467
$MAPE$	0.0348	0.0142

To further test the capability of the robust model a 10% contamination pressure test

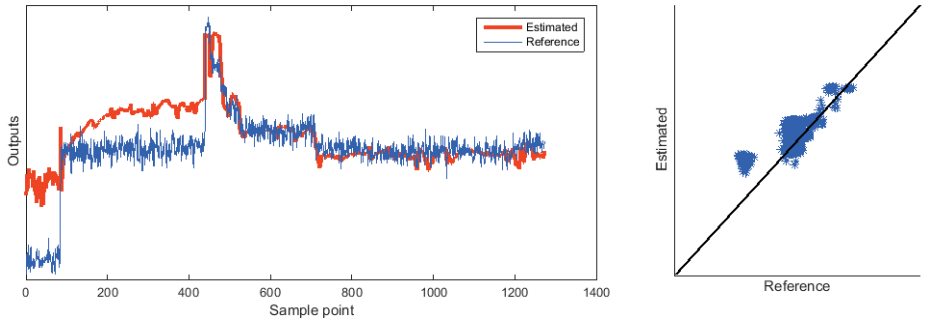


Figure 3.17: Prediction performance of PPCA regression model on 10% contaminated industrial data

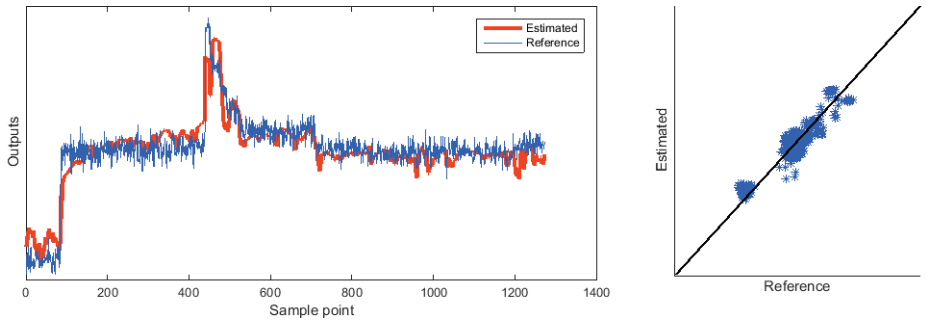


Figure 3.18: Prediction performance of RPPCA regression model on 10% contaminated industrial data

is done and the results are presented in Figure 3.17 and Figure 3.18, for the regular and robust model performances, respectively. Table 3.7 summarizes the results for this test. As it is observed, the extent of degradation of the regular model is closely related to the extent of contamination.

Table 3.7: Prediction performance of regular and robust models in industrial example by 10% contamination in inputs and output

	Regular model	Robust model
R^2	0.5057	0.8419
$RMSE$	2.0132	0.5173
$MAPE$	0.0545	0.0134

3.5 Conclusions

This chapter extends the main idea of authors' last work presented in [75], by considering the location outliers in both input and output data, which is very common in industry. The contaminating location outliers are described by a Gaussian mixture distribution. This is a more general class of distribution for outlying observations while still maintaining the analytical properties of a Gaussian distribution for obtaining closed form solutions. This model is considered for both input and output noises to account for both regression outliers and high leverage points, simultaneously.

The developed robust model under this noise assumption was solved for parameters using EM algorithm. The solution was evaluated using numerical and real data examples in the sense of prediction performance of the model and its error. Results show the robustness of the developed model while the performance of regular PPCA breaks down in the presence of outliers in the simulation case study. The robust model also showed its ability to overcome the problem of high leverage points. Then the developed model was applied to real process data from a SAGD operation to prove the robustness.

Chapter 4

Robust Probabilistic Principal Component Analysis for Process Modeling Subject to Switching Scaled Mixture Gaussian Noise*

"The fragile wants tranquility, the antifragile grows from disorder, and the robust doesn't care too much."

Antifragile: Things That Gain from Disorder (N. N. Taleb, 2012)

*A version of this chapter is submitted to Chemometrics and Intelligent Laboratory Systems, as Anahita Sadeghian, Nabil Magbool Jan, Ouyang Wu, Biao Huang, "Robust Probabilistic Principal Component Analysis for Process Modeling Subject to Switching Scaled Mixture Gaussian Noise"

4.1 Introduction

To evaluate and measure the profitability of a process, it is essential to know product quality and/or rate of its production. On the other hand, compliance to safety and environmental constraints is also crucial. To meet these needs, advanced process control and monitoring techniques have an extensive role nowadays [24]. The first and one of the most important components in applying these techniques on the process at hand, is providing the appropriate data which are usually obtained from recorded observations showing the history of its behaviour. To accommodate for this need, instrumentation industry is heavily involved through using a wide range of sensors and analyzers.

The limitation of relying only on instruments is that some values needed to analyze the process are not measurable and/or are expensive to measure using traditional methods such as sending a sample for analysis in a laboratory. Such instances include providing the online value for a fast-rate process variable or an expensive lab analysis, expensive either time-wise or cost-wise. As an example, laboratory sample analysis might lack in terms of availability and an online analyzer might not be reliable at times. Other cases of shortfall might be due to restricted availability and/or reliability of available measurement techniques or devices. All these shortages might result in a data set that has a set of complications in dealing with process data including but not limited to missing, completely off and incorrect, biased and delayed values. This is aside from drawbacks of some conventional sampling methods that might need the process to be interrupted [75].

Predictive models are a solution to provide us with frequent estimates for a variable of interest which is either measurable but suffers from issues mentioned above, or a not measurable one that could still be inferred in some way based on other measured variables and the historical behaviour of the process. This is where advanced techniques and algorithms come into the picture to prevent loss of profitability and/or safety deprivations contingent upon a failed sensor or lack thereof. Thus, inference of key variables can potentially cut down the cost of hardware installation and maintenance. Predictive mathematical models, also known as soft sensors, inferential sensors or virtual sensors, provide online estimates of key process variables based on some

other process records [24].

This work is focusing on design of data-driven soft sensors. There are other categories mentioned in the literature [24]. Trivially the quality and relevance of the data used for this design has an important role in performance of designed soft sensor. Aside from data, choice of development methodology for this design also matters. One of the sought after avenues is high-fidelity modeling that is able to scale down the detrimental effect of disturbances and anomalies inherent in the data. An adequate choice of an analysis method increases the possibility of dealing with some of these issues inherent in the recorded data.

Although there are different filtering methods to handle this issue, a proper analysis of the data is a better way to cope with it. A proper analysis helps us to detect and deal with data issues such as outliers, missing data, redundancy, low accuracy, etc. Several ad-hoc solutions to these potential issues have been addressed in literature [74, 24, 42, 33, 10]. Presence of outliers, is one of the main concerns that affect data quality [58]. One other important characteristic of data which is the base of the data-driven modeling approaches, is multi-modality. This characteristic has been studied in literature mainly for the multiple *process* models, such as in [11, 39, 102, 98] to deal with identification of complex systems. The same characteristic could exist in *noise* models. This work will focus on considering multiple noise models that affect the process data. This situation could happen when there are outlying observations among the process or laboratory data. These points are basically caused by random errors and have been discussed in more detail in Chapter 2 and Chapter 3.

4.1.1 Hidden Markov Models

Hidden Markov Models (HMMs), as probabilistic generative models, are mathematical tools for temporal information implementation in process modeling and also in history-based fault diagnosis. Aside from temporal analysis, HMMs are also used in operating mode transition modeling because of their ability to model switching states. Through fusion of time-domain information, HMMs will greatly assist a model or a classifier in cases of fault diagnosis applications. For example, to distinguish process modes with significant overlaps, one can take advantage of HMMs. Such approach will consider some sort of prior knowledge about the process's temporal behavior in

the form of transition probability matrix [83].

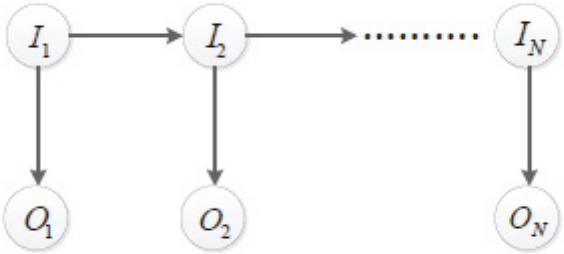


Figure 4.1: Typical HMM structure

HMMs, as probabilistic sequence models, are used to estimate the joint probability distribution of hidden states I and observations O. Under the framework of HMM, states corresponding to different operating modes can perform transitions and generate observations with different probability distributions. A typical structure is shown in Figure 4.1. For simplicity, two conditional independence assumptions are considered in HMMs, i.e., $P(I_t|I_{t-1}, ..., I_1) = P(I_t|I_{t-1})$ and $P(O_t|I_t, O_1, ..., O_{t-1}) = P(O_t|I_t)$. The former describes the transition probability between two hidden states and will turn into a transition probability matrix when considering all the possible transitions. As a result, in comparison with other multimodal algorithms such as Gaussian Mixure Models (GMM) and Mixture Probabilistic Principal Component Analysis (MPPCA), HMMs perform much better at modeling the transitions.

HMMs are simple, well-performed and extensible in modeling state transitions. In dynamic process modeling and monitoring, HMMs are effective frameworks in dealing with industrial data issues such as missing observations, outliers and time-varying transitions [77, 78]. In order to achieve satisfactory application performance, various feature extraction algorithms have been combined with HMMs. This improves model performance in process modeling and monitoring while considering industrial data dynamics.

Missing data problems usually arise from different sources such as sensor failures and other data collection errors. In order to deal with missing data problems, in HMM based operating mode diagnosis, [78] proposed an algorithm based on the expecta-

tion maximization (EM) approach for model estimation. [48] designed a semi-Markov model to estimate the statistical patterns of missing data. To solve the outlier issues, conventional normal threshold based outlier removal techniques will cause loss of information. In order to reduce the negative effect of outliers without information loss, different probability distributions have been used for data description. [38] applied a contaminated Gaussian distribution to describe the noises, but the Gaussian component corresponding to outliers has a fixed variance. More generally, the Student's t-distribution has been used for modeling with robustness to outliers [21]. The t-distribution is also embedded into the HMMs framework to describe the observation likelihood function given each hidden state [77, 89]. During process operation, due to external factors, the operation environment will be disturbed persistently, causing time-varying dynamics in data. For such situations, adaptive modeling has been proposed with an auxiliary scheduling variable. Accordingly, the parameter-invariant HMM has been extended to an adaptive framework with a time-varying transition matrix for dynamic modeling [78].

While using probabilistic generative models, HMMs require explicit probability distributions to model observations; to simplify the factorization, conditional independencies are assumed within HMMs [51]. To generalize, extended versions of HMMs have been proposed to relax these inherent assumptions of a conventional HMM. Some examples of these extensions are autoregressive HMMs [85], higher order HMMs [14] and the factorial HMM framework designed by [55] with three hidden layers, etc., although they are expensive in terms of computation. Since such extensions increase the computational load, model training and inference for these HMMs will become more complex. Moreover, consideration of appropriate distributions that capture the observations is still recommended in the extended versions. For an inaccurate probabilistic model, the accuracy of its application (for example, a fault diagnosis task) will be affected consequently.

It is worth noting that probabilistic discriminative models have been introduced to alleviate such vulnerability and to compensate for the potential shortcomings of generative models. In discriminative probabilistic models, instead of modeling the joint probability, the conditional probability is directly formulated and then optimized. On another note aside from HMMs popularity in research, they have some

limitations due to the common assumption of conditional independence. In [17], the state duration of HMMs follows an exponential distribution, which limits HMMs in providing adequate representation of temporal correlations.

HMMs have been used in conjunction with model-based approaches to improve fault detection. This way, instead of assuming that faulty data points are independent of each other, it is assumed that the faults are correlated over time periods. So, by the means of HMMs, previous information up to the current time is used as a prior for the more recent observations [83]. HMMs have also been adopted in qualitative trend analysis where significant events are extracted from process data. Then, major events are analyzed as a time sequence in order to provide critical information on the status of the process at hand. This approach has been used in online applications by considering a window of recent observations [92].

Recent studies have used probabilistic generative models such as PPCA with HMMs for fault diagnosis [100, 101]. PPCA considers the uncertainties and this is an advantage over using a regular PCA. On top of that, considering process dynamics and their transitions with HMMs builds a strong general structure. The goal would be inferring the true current operating mode upon the reception of a new observation. Other studies have improved PPCA-based algorithms by handling the outliers. For example, a more inclusive distribution, such as a Gaussian mixture of distributions with different centres and/or spreads, have been used as previously presented in Chapter 2 and Chapter 3. Similar consideration can be seen in [101] by use of a Student's t-distribution which translates as an ultimate case of Chapter 2, where the mixture distribution contains an infinite number of Gaussian distributions with the same centre.

Statistical classifiers are another example where HMMs come into the picture. In [37], the authors introduce fault diagnosis of a gear transmission system. They modeled a process considering a triple-state continuous-time homogeneous Markov process. Each of the different states corresponds to a process operating mode. They have targeted the inference of the current process mode in an online application. Many more works have been published incorporating these probabilistic sequence models [95, 28, 93, 76].

4.2 Fundamentals and Problem Formulation

This section is dedicated to a concise review of some preliminaries over the formulation of the aimed probabilistic algorithm, and over the methodology of solving it.

Given N samples of m-dimensional inputs and r-dimensional outputs, denote the input dataset as $X = [x_1, x_2, \cdots, x_N]^T \in \Re^{N\times m}$ and output data set as $Y = [y_1, y_2, \cdots, y_N]^T \in \Re^{N\times r}$. The PPCA based model is given by:

$$\begin{cases} x_n = Pt_n + \mu_x + e_n \\ y_n = Ct_n + \mu_y + f_n \end{cases} \tag{4.1}$$

where $t_n \in \Re^{q\times 1}$ is a latent variable vector having a dimension of $q < N$, such that $T = [t_1, t_2, \cdots, t_N]^T \in \Re^{q\times m}$. In a matrix form, the equation would be as

$$\begin{cases} X = TP^T + M_x + E \\ Y = TC^T + M_y + F \end{cases} \tag{4.2}$$

In this chapter, it is assumed that the measurement errors for inputs follow a Gaussian distribution such that,

$$e_n \sim \mathcal{N}(0, \sigma_x^2 I)$$

while measurement errors in the output are assumed to have switching noise modes I_{yn}. Vector of variable $I_Y = [I_{y_1}, \cdots, I_{y_N}]^T \in \Re^{N\times 1}$, $(I_{y_n} \in \Re^{1\times 1})$ is introduced as a vector of binary indicators which implies the identity of noise model. When this indicator $I_{y_n} = 1$, the output noise f_n is distributed as $\mathcal{N}(0, \sigma_y^2 I)$; and when $I_{y_n} = \rho_y$, f_n is distributed as $\mathcal{N}(0, \rho_y^{-1}\sigma_y^2 I)$, as shown below.

$$f_n | I_{y_n} \sim \begin{cases} \mathcal{N}(0, \sigma_y^2 I) & , \ for \quad I_{y_n} = 1 \\ \mathcal{N}(0, \rho^{-1}\sigma_y^2 I) & , \ for \quad I_{y_n} = 2 \end{cases}$$

The level of switching between different noise modes is parameterized in terms of transition probabilities matrix, $\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix}$, where α_{ij} signifies the transition probability of switching from noise mode i to the noise mode j. The prior probability of the occurrence of the first noise mode is denoted as π_{y_1}, and that of the second noise mode is given by $\pi_{y_2} = 1 - \pi_{y_1}$.

For the robust PPCA based algorithm to be developed, the available observations are X and Y. The unknown parameters to be estimated are collected in Θ. Since the

latent variables are unknown in the model presented in (4.1), we formulate the robust PPCA based problem in terms of a complete-data log-likelihood as follows:

$$\underset{T,I_Y,\Theta}{\arg\max} \quad log\ p\ (\underbrace{\overbrace{X,Y}^{Observed}\ ,\ \overbrace{T,I_Y}^{Hidden}}_{Complete\ \ Data}|\Theta), \tag{4.3}$$

where $\Theta = \{P, C, \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_y, \alpha, \Pi_{y_1}\}$.

Assumptions

- The input vectors, $\{x_n\}_{n=1}^N$, are assumed to be independent, and identically distributed (i.i.d) random variables.
- The output vector, $\{y_n\}_{n=1}^N$, is affected by the switching mode indicator I_{y_n}.
- The indicator variable, I_{y_n}, follows a first order Markov property.
- The latent vector, $\{t_n\}_{n=1}^N$, is assumed to be independent, and identically distributed.

So, the complete-data log-likelihood can be expressed as:

$$\begin{aligned} log\ p(X,Y,T,I_Y|\Theta) &= log\ \Big(p(X|T,\Theta)\ p(Y|T,I_Y,\Theta)\ p(T|I_Y,\Theta)\ p(I_Y|\Theta)\Big) \qquad (4.4)\\ &= log\ \Big(\prod_{n=1}^{N}\Big[p(x_n|t_n,\Theta)\ p(y_n|t_n,I_{y_n},\Theta)\ p(t_n|I_{y_n},\Theta)\ p(I_{y_n})\Big]\Big) \end{aligned}$$

Since I_Y follows a first order Markov property, $p(I_Y) = \Big[\prod_{n=2}^{N} p(I_{y_n}|I_{y_{n-1}})\Big]p(I_{y_1})$. Thus, (4.4) will be as

$$\begin{aligned} log\ p(X,Y,T,I_Y|\Theta) &= \sum_{n=1}^{N} log\ p(x_n|t_n,\Theta) + \sum_{n=1}^{N} log\ p(y_n|t_n,I_{y_n},\Theta)\\ &\quad + \sum_{n=1}^{N} log\ p(t_n|I_{y_n},\Theta) + \sum_{n=2}^{N} log\ p(I_{y_n}|I_{y_{n-1}})\\ &\quad + log\ p(I_{y_1}) \qquad (4.5) \end{aligned}$$

Owing to the presence of hidden variables, T, I_Y, the formulated robust algorithm is difficult to solve in general. Therefore, expectation maximization algorithm is used to

solve the formulated maximum likelihood problem above. Expectation Maximization (EM) has been previously discussed in Section 2.4.1.

4.3 Solution Methodology

Denote the set of observed variables as $C_{obs} = \{X, Y\}$, and the set of hidden variables as $C_{mis} = \{T, I_Y\}$. Then the complete data would be $\{C_{obs}, C_{mis}\}$. The unknown parameters, Θ, can be determined by solving the following Maximum Likelihood Estimation (MLE) problem:

$$\arg\max_{\Theta} \quad p(C_{obs}, C_{mis}|\Theta) \tag{4.6}$$

For problems involving hidden variables, EM algorithm is often used. The basic idea of this algorithm is to solve (4.6) in two steps, iteratively.

- Step 1 (Expectation): Evaluate the $Q - function$ defined as the conditional expectation of hidden variables,

$$Q(\Theta, \Theta^{old}) = \int p(C_{mis}|C_{obs}, \Theta^{old}) \Big[log\ p(C_{mis}, C_{obs}|\Theta) \Big]\ dC_{mis} \tag{4.7}$$

- Step 2 (Maximization): Optimize the parameters by maximizing the $Q-function$,

$$\Theta^{new} = \arg\max_{\Theta} \quad Q(\Theta, \Theta^{old}) \tag{4.8}$$

The current optimal values of parameters, Θ^{new}, are used to re-evaluate the $Q - function$ in the E-step, and the process is repeated until convergence. To derive the $Q - function$, we take the conditional expectation of the complete-data log-likelihood over the hidden variables. Thus, the $Q - function$ for the robust PPCA based algorithm under consideration is given by

$$
\begin{aligned}
Q &= \underset{T,I_y|X,Y,\Theta^{old}}{\mathbb{E}}\Big[log\ p(C_{mis}, C_{obs}|\Theta)\Big] = \underset{T,I_y|X,Y,\Theta^{old}}{\mathbb{E}}\Big[log\ p(X, Y, T, I_Y|\Theta)\Big] \\
&= \int_{t_n} \sum_{I_{y_n}} p(t_n, I_{y_n}|x_n, y_n, \Theta^{old})\Big[log\ p(C_{mis}, C_{obs}|\Theta)\Big]\ dt_n \\
&= \int_{t_n} \sum_{I_{y_n}} p(t_n, I_{y_n}|x_n, y_n, \Theta^{old})\Big[\sum_{n=1}^{N} log\ p(x_n|t_n, \Theta)\Big]\ dt_n \\
&+ \int_{t_n} \sum_{I_{y_n}} p(t_n, I_{y_n}|x_n, y_n, \Theta^{old})\Big[\sum_{n=1}^{N} log\ p(y_n|t_n, I_{y_n}, \Theta)\Big]\ dt_n \\
&+ \int_{t_n} \sum_{I_{y_n}} p(t_n, I_{y_n}|x_n, y_n, \Theta^{old})\Big[\sum_{n=1}^{N} log\ p(t_n|I_{y_n}, \Theta)\Big]\ dt_n \\
&+ \int_{t_n} \sum_{I_{y_n}} p(t_n, I_{y_n}|x_n, y_n, \Theta^{old})\Big[\sum_{n=2}^{N} log\ p(I_{y_n}|I_{y_{n-1}})\Big]\ dt_n \\
&+ \int_{t_n} \sum_{I_{y_n}} p(t_n, I_{y_n}|x_n, y_n, \Theta^{old})\Big[log\ p(I_{y1})\Big]\ dt_n \\
&= \mathbb{Q}_1 + \mathbb{Q}_2 + \mathbb{Q}_3 + \mathbb{Q}_4 + \mathbb{Q}_5
\end{aligned} \tag{4.9}
$$

In the above equation,

$$
\begin{aligned}
p(t_n, I_Y|x_n, y_n, \Theta^{old}) &= p(t_n|x_n, y_n, \Theta^{old})\ p(I_{y_1:y_n}|x_n, y_n, \Theta^{old}) \\
&= p(t_n|x_n, y_n, \Theta^{old})\Big[\prod_{j=2}^{n} p(I_{y_j}|I_{y_{j-1}}, x_n, y_n, \Theta^{old})\Big]\ p(I_{y1})
\end{aligned} \tag{4.10}
$$

4.3.1 Parameter Estimation

As previously mentioned, in RPPCA-based algorithm here the available observations are X and Y. The unknowns in this chapter and under the assumption of switching noise modes are

$$
Unknowns = \{P, C, \sigma_x^2, \sigma_y^2, \mu_x, \mu_y, \rho_y, T, I_Y, \alpha.\pi_y\}. \tag{4.11}
$$

From this set of unknowns, T, I_Y are treated as latent variables, and the rest are contained in the parameters set $\Theta = \{P, C, \sigma_x^2, \sigma_y^2, \mu_x, \mu_y, \rho_y\, \alpha, \pi_y\}$ to be estimated. The EM algorithm is deployed to estimate the parameters of the model, since there exist some hidden variables in addition to the unknown parameters. We need to

construct the $Q - function$ based on (4.7) first, and then maximize that function to obtain the estimation of parameters by means of an iterative procedure. To obtain the parameter update equations, (4.9) should be maximized with respect to the set of parameters. Simply,

$$\begin{aligned}\Theta &= \arg\max_{\Theta} \mathop{\mathbb{E}}_{T,I_Y|X,Y,\Theta^{old}} (logP(X,Y,T,I_Y|\Theta)) \\ &= \arg\max_{\Theta} \quad \mathbb{Q}_1 \ + \mathbb{Q}_2 \ + \mathbb{Q}_3 \ + \mathbb{Q}_4 \ + \mathbb{Q}_5 \end{aligned} \tag{4.12}$$

which is equivalent to solving a set of equations (4.13). For each parameter, a partial differentiation of a specific term of the $Q - function$, which contains the parameter of interest, is used. Results for parameters update equations, after solving (4.13), are given in (4.14) to (4.27).

$$\begin{aligned} &\text{P:} && \tfrac{\partial \mathbb{Q}_1}{\partial P} = 0 \\ &\text{C:} && \tfrac{\partial \mathbb{Q}_2}{\partial C} = 0 \\ &\sigma_x^2\text{:} && \tfrac{\partial \mathbb{Q}_1}{\partial \sigma_x^2} = 0 \\ &\sigma_y^2\text{:} && \tfrac{\partial \mathbb{Q}_2}{\partial \sigma_y^2} = 0 \\ &\mu_x\text{:} && \tfrac{\partial \mathbb{Q}_1}{\partial \mu_x} = 0 \\ &\mu_y\text{:} && \tfrac{\partial \mathbb{Q}_2}{\partial \mu_y} = 0 \\ &\rho_y\text{:} && \tfrac{\partial \mathbb{Q}_2}{\partial \rho_y} = 0 \\ &\alpha\text{:} && \tfrac{\partial \mathbb{Q}_4}{\partial \alpha} = 0 \\ &\pi\text{:} && \tfrac{\partial \mathbb{Q}_5}{\partial \pi_y} = 0 \end{aligned} \tag{4.13}$$

$$\begin{aligned} P^{new} \ = \ & \left[2 \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \ (x_n - \mu_x) \underset{k}{E}^T \right] \\ & \times \left[\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \ \underset{k}{Coeff} \right]^{-1}, \end{aligned} \tag{4.14}$$

where

$$\underset{k}{E} = \mathbb{E}(t_n|x_n, y_n, I_{y_n} = k, \Theta^{old}) \tag{4.15}$$

$$\underset{k}{Coeff} = \underset{k}{S}^T + \underset{k}{S} + 2 \underset{k}{E} \underset{k}{E}^T, \tag{4.16}$$

and

$$\underset{k}{S} = \underset{k}{E} - \underset{k}{E}\,\underset{k}{E}^T. \tag{4.17}$$

Similarly,

$$\begin{aligned} C^{new} &= \left[2\sum_{n=1}^{N}\left(\gamma_{n1}\ (y_n-\mu_y)\underset{1}{E}^T + \rho_y\gamma_{n2}\ (y_n-\mu_y)\underset{2}{E}^T\right)\right] \\ &\quad \times\left[\sum_{n=1}^{N}\left(\gamma_{n1}\ \underset{1}{Coeff} + \rho_y\gamma_{n2}\ \underset{2}{Coeff}\right)\right]^{-1} \end{aligned} \tag{4.18}$$

$$\mu_x^{new} = \frac{1}{N}\left[\sum_{n=1}^{N} x_n - P\Big(\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\,\underset{k}{E}\Big)\right] \tag{4.19}$$

$$\mu_y^{new} = \frac{\sum_{n=1}^{N}\left(\gamma_{n1}\ \left(y_n - C\,\underset{1}{E}\right) + \rho_y\gamma_{n2}\ \left(y_n - C\,\underset{2}{E}\right)\right)}{\sum_{n=1}^{N}\left(\gamma_{n1}+\rho_y\gamma_{n2}\right)} \tag{4.20}$$

$$\begin{aligned} \sigma_x^{2\,new} &= \frac{1}{m\,N}\left[\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\left((x_n-\mu_x)^T(x_n-\mu_x) - \underset{k}{E}^T P^T(x_n-\mu_x)\right.\right. \\ &\qquad \left.\left. -(x_n-\mu_x)^T P\,\underset{k}{E} + \underset{k}{E}(t_n^T P^T P t_n)\right)\right], \end{aligned} \tag{4.21}$$

in which

$$\begin{aligned} \underset{k}{E}(t_n^T P^T P t_n) &= tr\Bigg(P^T P\ \Big(\underset{k}{E}(t_n t_n^T|x_n,y_n,I_{y_n}=k,\Theta^{old}) \\ &\qquad - \underset{k}{E}(t_n|x_n,y_n,I_{y_n}=k,\Theta^{old})\underset{k}{E}(t_n|x_n,y_n,I_{y_n}=k,\Theta^{old})^T\Big)\Bigg) \\ &\quad + \underset{k}{E}(t_n|x_n,y_n,I_{y_n}=k,\Theta^{old})^T P^T P\,\underset{k}{E}(t_n|x_n,y_n,I_{y_n}=k,\Theta^{old}) \end{aligned} \tag{4.22}$$

$$\sigma_y^{2\,new} = \frac{1}{r\,N}\Bigg[\sum_{n=1}^{N}\Bigg(\gamma_{n1}\Big((y_n-\mu_y)^T(y_n-\mu_y) - \underset{1}{E}^T C^T(y_n-\mu_y) - (y_n-\mu_y)^T C\,\underset{1}{E} + \underset{1}{E}(t_n^T C^T C t_n)\Big) + \rho_y\gamma_{n2}\Big((y_n-\mu_y)^T(y_n-\mu_y) - \underset{2}{E}^T C^T(y_n-\mu_y) - (y_n-\mu_y)^T C\,\underset{2}{E} + \underset{2}{E}(t_n^T C^T C t_n)\Big)\Bigg)\Bigg] \tag{4.23}$$

, where

$$\underset{k}{E}(t_n^T C^T C t_n) = tr\Bigg(C^T C\Big(\underset{k}{E}(t_n t_n^T|x_n,y_n,I_{y_n}=k,\Theta^{old}) - \underset{k}{E}(t_n|x_n,y_n,I_{y_n}=k,\Theta^{old})\underset{k}{E}(t_n|x_n,y_n,I_{y_n}=k,\Theta^{old})^T\Big)\Bigg) + \underset{k}{E}(t_n|x_n,y_n,I_{y_n}=k,\Theta^{old})^T C^T C\,\underset{k}{E}(t_n|x_n,y_n,I_{y_n}=k,\Theta^{old}) \tag{4.24}$$

$$\rho_y^{new} = \Big[r\;\sigma_y^2\sum_{n=1}^{N}\gamma_{n2}\Big]\times\Big[\sum_{n=1}^{N}(y_n-\mu_y)^T(y_n-\mu_y) - \underset{2}{E}^T C^T(y_n-\mu_y) - (y_n-\mu_y)^T C\,\underset{2}{E} + \underset{2}{E}(t_n^T C^T C t_n)\Big]^{-1} \tag{4.25}$$

$$\alpha_{jk}^{new} = \Big[\sum_{n=2}^{N}\xi_{n-1,n}^{(jk)}\Big]\times\Big[\sum_{k=1}^{K}\sum_{n=2}^{N}\xi_{n-1,n}^{(jk)}\Big]^{-1} \tag{4.26}$$

$$\pi_k^{new} = \frac{\gamma_{1k}}{\sum_{k=1}^{K}\gamma_{1k}} \tag{4.27}$$

where,

$$\gamma_{nk} = p\,(I_{y_n}=k) \tag{4.28}$$

$$\xi_{n-1,n}^{(jk)} = p\,(I_{y_{n-1}}=j \quad , \quad I_{y_n}=k) \tag{4.29}$$

4.3.2 Latent Variables' A-posteriori Calculation

As shown in Section 4.3, I_Y and T in (4.2) were treated as hidden variables in order to use EM algorithm for the parameter estimation. For using the developed algorithm in order to predict the outputs corresponding to new inputs, as will be shown in Section 4.4, and to finalize the obtained solution of EM algorithm, we need to calculate the a-posteriori probabilities of the hidden variables. Also for the output prediction, the expectation of hidden variable T is needed. Basically, the goal of the expectation step in EM algorithm is to bring the updated parameters of the previous maximization step for re-calculating the a-posteriori probability distribution of latent variables and then updating their expected values. This is achieved via *Bayes' rule* and the *chain rule of probability*.

$$p(t_n, I_{y_n}, I_{y_{n-1}}|x_n, y_n, \Theta^{old}) = \frac{p(I_{y_n}, I_{y_{n-1}}|x_n, y_n, t_n, \Theta^{old})\, p(t_n|x_n, y_n, \Theta^{old})}{\sum_{I_{y_n}} \sum_{I_{y_{n-1}}} p(I_{y_n}, I_{y_{n-1}}|x_n, y_n, t_n, \Theta^{old})\, p(t_n|x_n, y_n, \Theta^{old})} \tag{4.30}$$

$$p(I_{y_n}, I_{y_{n-1}}|x_n, y_n, t_n, \Theta^{old}) = \frac{p(x_n, y_n, t_n|I_{y_n}, \Theta^{old})\, p(I_{y_n}|I_{y_{n-1}}, \Theta^{old})\, p(I_{y_{n-1}}|\Theta^{old})}{\sum_{I_{y_n}} \sum_{I_{y_{n-1}}} p(x_n, y_n, t_n|I_{y_n}, \Theta^{old})\, p(I_{y_n}|I_{y_{n-1}}, \Theta^{old})\, p(I_{y_{n-1}}|\Theta^{old})} \tag{4.31}$$

$$p(x_n, y_n, t_n|I_{y_n}, \Theta^{old}) = \frac{p(x_n, y_n|t_n, I_{y_n}, \Theta^{old})\, p(t_n|I_{y_n}, \Theta^{old})}{\sum_{I_{y_n}} \sum_{I_{y_{n-1}}} p(x_n, y_n|t_n, I_{y_n}, \Theta^{old})\, p(t_n|I_{y_n}, \Theta^{old})}$$

$$= \frac{p(x_n|t_n, I_{y_n}, \Theta^{old})\, p(y_n|t_n, I_{y_n}, \Theta^{old})\, p(t_n|I_{y_n}, \Theta^{old})}{\sum_{I_{y_n}} \sum_{I_{y_{n-1}}} p(x_n, y_n|t_n, I_{y_n}, \Theta^{old})\, p(t_n|I_{y_n}, \Theta^{old})} \tag{4.32}$$

$$\begin{aligned} x_n|t_n, I_{y_n}, \Theta^{old} &\sim \mathcal{N}(Pt_n + \mu_x, \sigma_x^2 I) \\ t_n|I_{y_n}, \Theta^{old} &\sim \mathcal{N}(0, I) \end{aligned} \tag{4.33}$$

$$y_n|t_n, I_{y_n}, \Theta^{old} \sim \begin{cases} \mathcal{N}(Ct_n + \mu_y, \sigma_y^2 I) & , \; for \quad I_{y_n} = 1 \\ \mathcal{N}(Ct_n + \mu_y, \rho_y^{-1}\sigma_y^2 I) & , \; for \quad I_{y_n} = 2 \end{cases} \tag{4.34}$$

Using the completion of squares formula, a joint distribution of input and output given the latent variables, can be obtained as follows

$$x_n, y_n, t_n | I_{y_n}, \Theta^{old} \sim \mathcal{N}(Joint\ mean, Joint\ variance) \tag{4.35}$$

where,

$$\begin{aligned} Joint\ mean = (\sigma_x^{-2} P^T P + \sigma_y^{-2} C^T C + I)^{-1} \cdots \\ \cdots \times (\sigma_x^{-2} P^T (x_n - \mu_x) + \sigma_y^{-2} C^T (y_n - \mu_y)) \end{aligned} \tag{4.36}$$

$$Joint\ variance = \begin{cases} (\sigma_x^{-2} P^T P + \sigma_y^{-2} C^T C + I)^{-1} & , \ for \quad I_{y_n} = 1 \\ (\sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} & , \ for \quad I_{y_n} = 2 \end{cases} \tag{4.37}$$

Expected mean and variance related terms of the a-posteriori distribution (4.10) are given in (4.38) to (4.41) for different values of the discrete hidden variable I_Y.

$$\begin{aligned} \underset{1}{E}(t_n | x_n, y_n, I_{y_n} = 1, \Theta^{old}) = (\sigma_x^{-2} P^T P + \sigma_y^{-2} C^T C + I)^{-1} \cdots \\ \cdots \times (\sigma_x^{-2} P^T (x_n - \mu_x) + \sigma_y^{-2} C^T (y_n - \mu_y)) \end{aligned} \tag{4.38}$$

$$\begin{aligned} \underset{1}{E}(t_n t_n^T | x_n, y_n, I_{y_n} = 1, \Theta^{old}) = (\sigma_x^{-2} P^T P + \sigma_y^{-2} C^T C + I)^{-1} \\ + \underset{1}{E}(t_n | x_i, y_n, q_{y_n} = 1, \Theta^{old}) \cdots \\ \cdots \times \underset{1}{E}(t_n | x_n, y_n, I_{y_n} = 1, \Theta^{old})^T \end{aligned} \tag{4.39}$$

$$\begin{aligned} \underset{2}{E}(t_n | x_n, y_n, I_{y_n} = 2, \Theta^{old}) = (\sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} \cdots \\ \cdots \times (\sigma_x^{-2} P^T (x_n - \mu_x) + \rho_y \sigma_y^{-2} C^T (y_n - \mu_y)) \end{aligned} \tag{4.40}$$

$$\begin{aligned} \underset{2}{E}(t_n t_n^T | x_n, y_n, I_{y_n} = 2, \Theta^{old}) = (\sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} \\ + \underset{2}{E}(t_n | x_n, y_n, I_{y_n} = 2, \Theta^{old}) \cdots \\ \cdots \times \underset{2}{E}(t_n | x_n, y_n, I_{y_n} = 2, \Theta^{old})^T \end{aligned} \tag{4.41}$$

4.4 Predictions

Inferring via a model (a.k.a. soft sensor), or prediction of the variables of interest as discussed in [24], refers to determining a quality variable in real-time based on the developed model by incorporating some other variables, which are easier to be measured in real-time. In this chapter, we use the proposed RPPCA based model developed in Section 2.4 to predict a variable of interest. This will be elaborated further in Section 4.5 with examples. Since the values of hidden variables of the probabilistic model are not known, posterior distribution for hidden variables is needed for performing such inference. Obtaining the posterior distribution of hidden variables has been presented in Section 4.3.2. To take uncertainty into account, the *law of total expectation* is applied as in (4.42).

$$\begin{aligned}\hat{t}_n &= E(t_n|x_n,\Theta^{old}) \\ &= \sum_{I_{y_n}}\sum_{I_{y_{n-1}}} E(t_n, I_{y_n}, I_{y_{n-1}}|x_n,\Theta^{old})\end{aligned} \tag{4.42}$$

The expectations of the hidden variable for each noise state are as follows:

$$\underset{1}{E}(t_n|x_n, I_{y_n}=1,\Theta^{old}) = (\sigma_x^{-2}P^TP+I)^{-1}(\sigma_x^{-2}P^T(x_n-\mu_x)) \tag{4.43}$$

$$\begin{aligned}\underset{1}{E}(t_nt_n^T|x_n, I_{y_n}=1,\Theta^{old}) = (\sigma_x^{-2}P^TP+I)^{-1} \\ + \underset{1}{E}(t_n|x_n, I_{y_n}=1,\Theta^{old})\cdots \\ \cdots \times \underset{1}{E}(t_n|x_n, I_{y_n}=1,\Theta^{old})^T\end{aligned} \tag{4.44}$$

Similarly,

$$\underset{2}{E}(t_n|x_n, I_{y_n}=2,\Theta^{old}) = (\sigma_x^{-2}P^TP+I)^{-1}(\sigma_x^{-2}P^T(x_n-\mu_x)) \tag{4.45}$$

$$\begin{aligned}\underset{2}{E}(t_nt_n^T|x_n, I_{y_n}=2,\Theta^{old}) = (\sigma_x^{-2}P^TP+I)^{-1} \\ + \underset{2}{E}(t_n|x_n, I_{y_n}=2,\Theta^{old})\cdots \\ \cdots \times \underset{2}{E}(t_n|x_n, I_{y_n}=2,\Theta^{old})^T\end{aligned} \tag{4.46}$$

Having the prediction for the latent variable, and using the generative model as in (4.2), the output prediction can be obtained as shown in (4.47). The error for this prediction, ϵ, which is the difference between predicted and real values of output variable, is obtained from (4.48).

$$\hat{y}_n = C\hat{t_n} + \hat{\mu}_y \tag{4.47}$$

$$\epsilon = \hat{y} - y \tag{4.48}$$

There are many standard performance measures for evaluating the goodness of fit and thus the prediction. R-squared and mean squared error (MSE) are used more often. The mean absolute percentage error (MAPE) is also a measure of how accurate a forecast is. It measures this accuracy as a percentage, and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values. Correlation between predicted and real values is also useful. Root mean squared error (RMSE) as in (4.49) is a well-known measure to give a quantitative sense in comparisons and is defined as

$$RMSE \triangleq \sqrt{\frac{\sum_{n=1}^{N'} \|\hat{y}_n - y_n\|^2}{N'}}, \tag{4.49}$$

where, N' is the total number of ***test*** samples and $\hat{y}$ and y are predicted and real output values, respectively.

4.5 Case Studies

In this section, the developed algorithm in this chapter is put into test for validation in a simulated example and then in an example with real industrial dataset.

4.5.1 Case I: Numerical Study

Here, the prediction performance of the developed algorithm in this chapter is verified and its robustness in presence of a switching noise is evaluated. To perform this analysis, a dataset is generated based on a known model. Output noise is simulated with an HMM that takes two states, one for the normal situation and

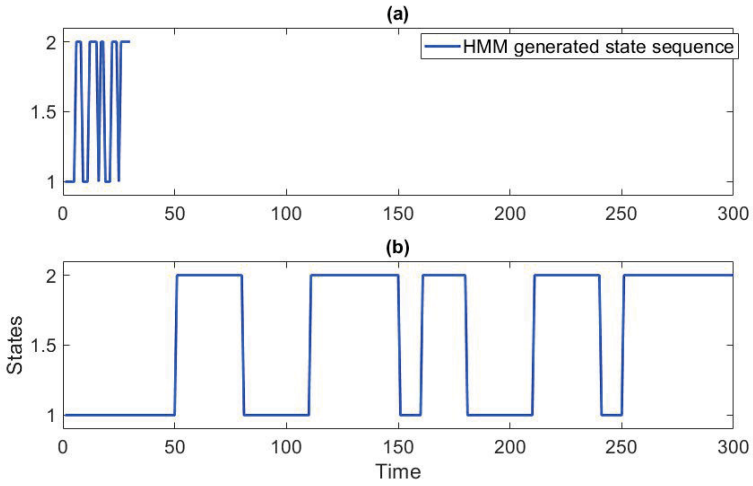


Figure 4.2: Generated state sequence for numerical case study using HMM

the other for the switched (abnormal) one. Data were generated including 300 observations from 5 input variables and 1 output variable. Two principal components are considered for generating the dataset. The loading matrix is set to $P_{(m=5)\times(k=2)} = \frac{1}{50}[40, 10; 20, 30; 15, 20; 15, 35; 40, 10]$, regression coefficient matrix is $C_{(r=1)\times(k=2)} = \frac{1}{50}[10, 40]$ and input/output mean vectors are $\mu_x = [0, 0, 0, 0, 0]$ and $\mu_y = [0]$, respectively. Input/output noises are Gaussian random variables. Input noise is a simple Gaussian variable with zero mean and output noise e_n has a switching model that will be discussed later in this section. Both input noise and the normal counterpart of the output noise, follow a Gaussian distribution with zero mean and a variance that is an averaged scaled version of original data variance (namely $e_n \sim \mathcal{N}(0, 30I)$, $f_n \sim \mathcal{N}(0, 30I)$, respectively). As a conventional assumption, latent variable vectors have a standard Gaussian distribution and are i.i.d. To get a hint on the number of latent variables, PCA can be applied to the data. Based on the scree plot, the number of latent variables can be chosen to capture the variability of this generated data. Generated state sequence is shown in Figure 4.2. Hidden state in this example has two modes, shown in the figure, "1" and "2", respectively. States are generated in different frequencies to simulate the mode changes. Figure 4.2(a) shows states that change in every time step, which represents a fast switching dynamic. Since chemical processes are not naturally fast in changes, Figure 4.2(b) is used to represent the state switching behaviour. To demonstrate the efficiency of

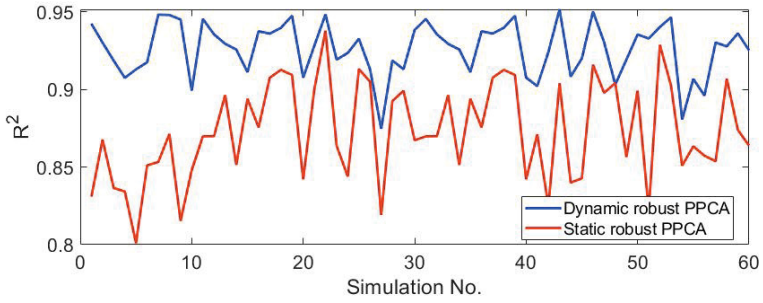


Figure 4.3: Fit statistics of Robust PPCA models with and without switching noise model

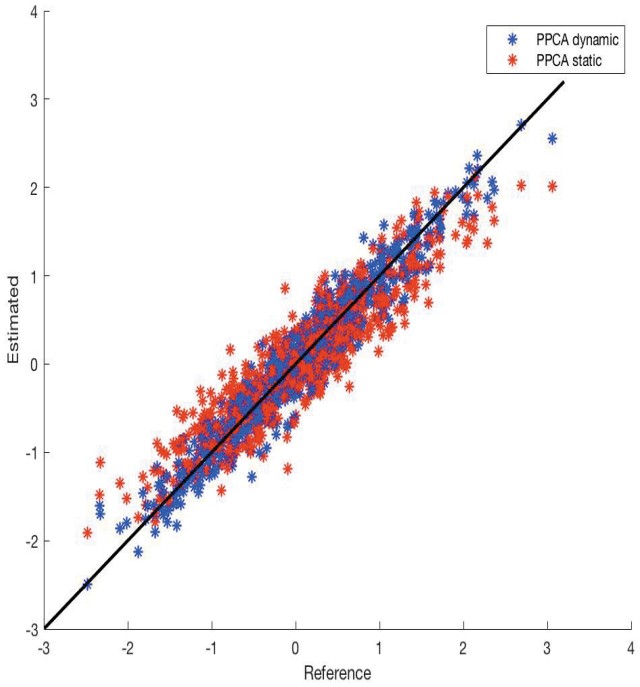


Figure 4.4: Scatter plot for predictions of two Robust PPCA-based models with and without switching noise model

our developed algorithm, we compared its performance with a robust PPCA based model that does not consider switching behaviour in the states of noise. This happens while both models are applied to a dataset that intentionally has switching modes for the noise model. Fit statistics R^2 for both models and in different runs are reported in Figure 4.3. It verifies the results of considering dynamic switching noise model as expected, when the underlying noise has a switching behaviour. Typical trend pre-

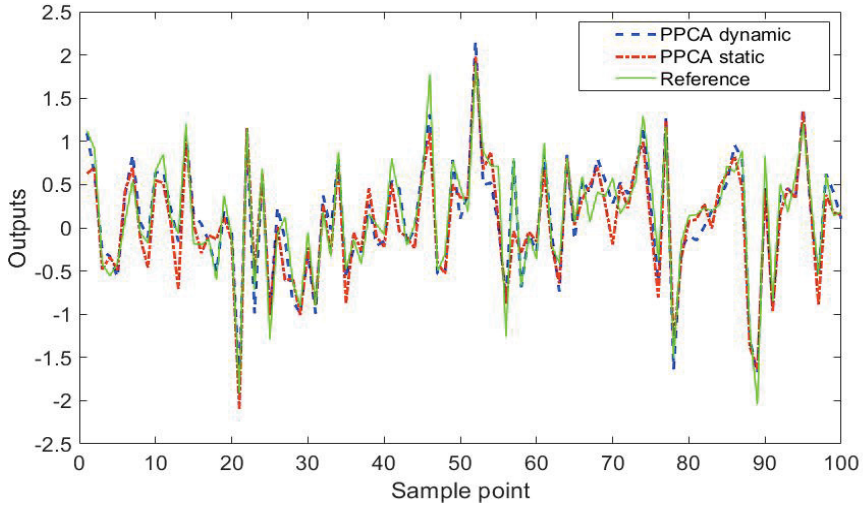


Figure 4.5: Trend plot for predictions of two Robust PPCA-based models with and without switching noise modes

diction performance of the two robust models are shown in Figure 4.4 and Figure 4.5 on a dataset that has noise switching. As the figures show the robust PPCA-based algorithm that considers noise mode switch, acts more robust compared to the previous algorithm developed in Chapter 2 which assumed independent and identically distributed (i.i.d.) noise attributes.

To conclude this case study, a set of contaminated data for two different noise state transition patterns, with the inflation factor of $\rho = 0.01$, are used and the two RPPCA based algorithms are compared in terms of their performances. Prediction error measures are compared in Table 4.1. *Static* and *Dynamic* in the table headings refer to the sequence in noise model.

Table 4.1: Prediction performance in robust models for two transitions, $\rho = 0.01$

	RPPCA-Static			RPPCA-Dynamic		
Transition matrix	R^2	RMSE	MAPE	R^2	RMSE	MAPE
$[0.9, 0.1; 0.1, 0.9]$	0.8405	0.372	1.144	0.9255	0.2543	0.8306
$[0.6, 0.4; 0.4, 0.6]$	0.8341	0.3675	1.3357	0.9244	0.2481	0.9714

4.5.2 Case II: Industrial Application

In this section, the robust algorithms are used on a real dataset to predict a variable of interest. SAGD operation is chosen for data acquisition. The operation details have been discussed in Section 2.5.2 and Section 3.4.2.

Robust probabilistic model with switching noise modes

Here, the developed robust PPCA based algorithm, which now has the capability of handling switching noise modes, is used for process modeling. The model performance is compared to that of a robust counterpart previously showcased in Chapter 2, which has only considered random noise mode switching. In both cases a scaled Gaussian mixture noise is considered for the model development, but each with different assumptions about the mode switching dynamics; one considers the correlation of the switching sequence and the other does not. A correlated switching sequence is added in the scaled Gaussian mixture measurement noise. In reality, the measurement noise is not solely originated from the same distribution; this could be seen in terms of distribution family and also in terms of hyperparameters of the mixed counterparts. For representing normal noise and outliers, two switching modes are considered.

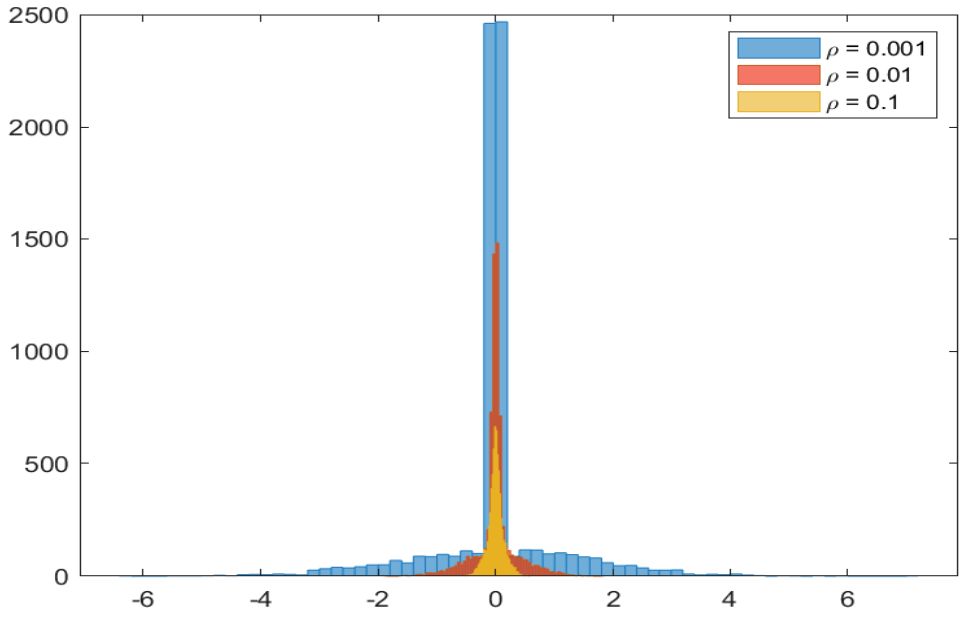


Figure 4.6: Mixture noise distribution for two different values of spread factor

The existing historical data from this process are a set of flow rates, pressures, and temperatures for injection and production well pair which are constructed by a 10 min average data recorded for a season in year 2014. This dataset consists of a total of 4 inputs and 1 output variables that have been recorded in about 52500 sample points, from which a set of 10906 sample points have been considered for the analysis in which the process shows a linear and time-invariant behavior. This dataset has been used in Chapter 2 for a scaled Gaussian mixture based robust algorithm without considering the dynamics of the switching sequence. In this study roughly two third of the data is used for training and the rest is left out for validation. Both robust algorithms are applied for obtaining a set of model parameters and then are compared in terms of their prediction performances. Noise model switching sequence is generated by a random sequence of states generated from an HMM. Figure 4.6 shows the noise distribution for two different inflation factors, representing the extent of contamination.

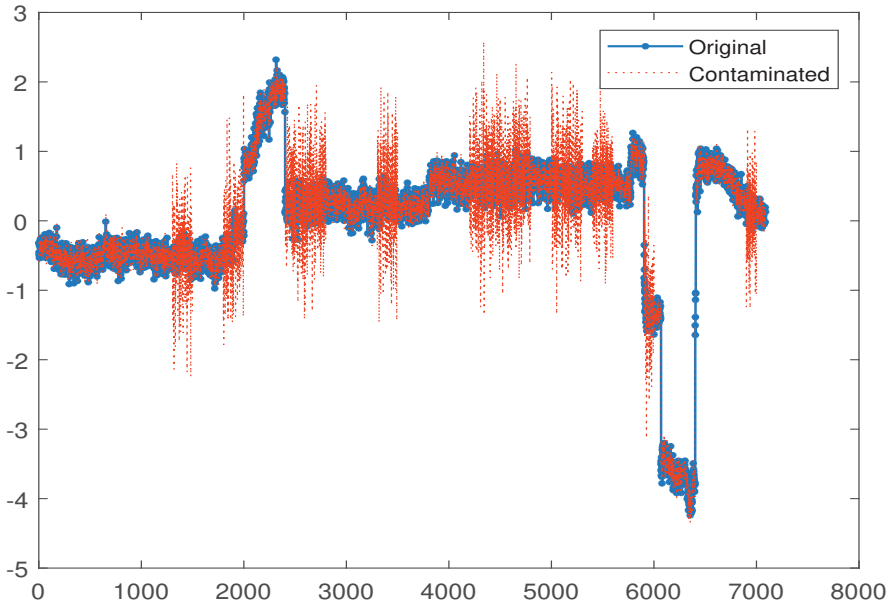
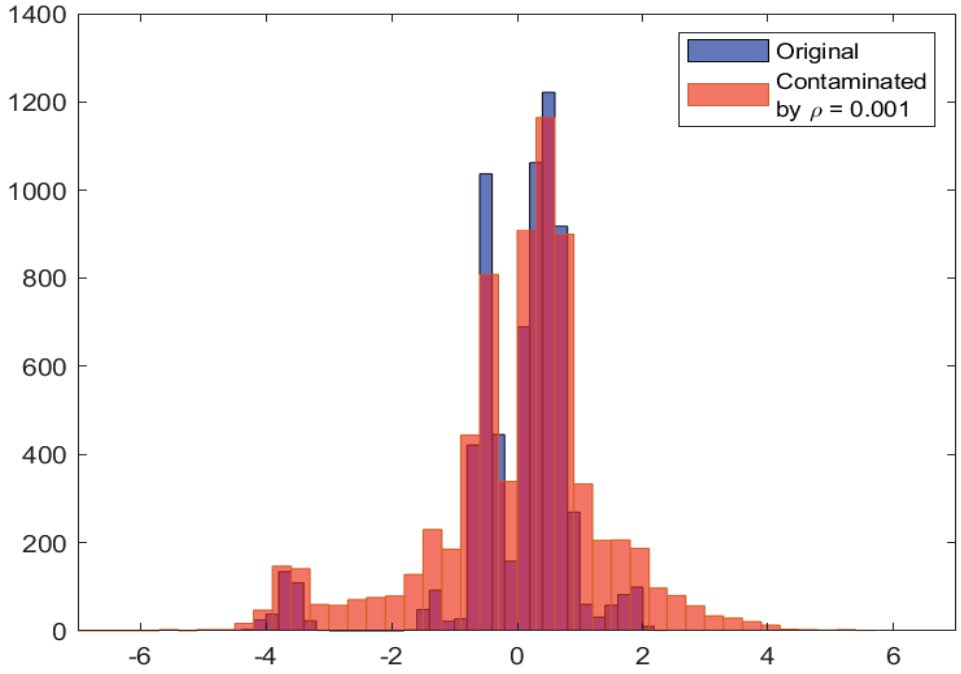
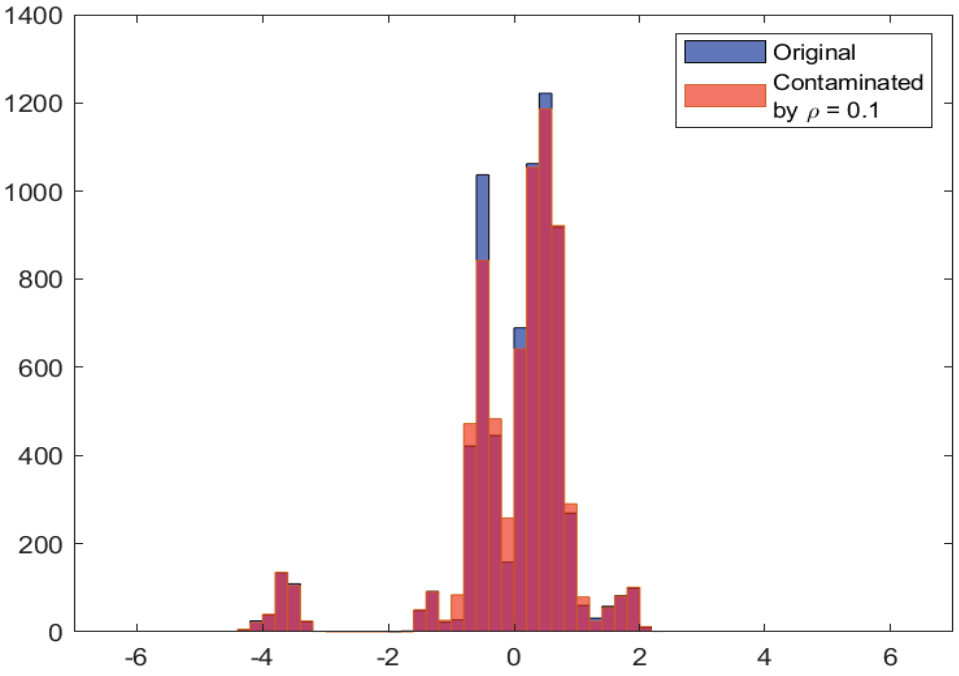


Figure 4.7: SAGD product oil flowrate and its contaminated counterpart



(a) Original data vs contaminated data with $\rho = 0.001$



(b) Original data vs contaminated data with $\rho = 0.1$

Figure 4.8: Data contamination

In this study, the inflation factor of the scaled noise mixture distribution is set to 0.001, meaning that one noise mode has a spread of 1000 times of the other to represent the outliers. For transitioning between the two noise modes (normal and outliers), transition probabilities of 30 percent and 70 percent are considered, indicating that during the time, each of the states would repeat itself most of the time (70 percent) and will have a transition to the other mode with a probability of 30 percent. The possible transition happens in intervals of every 100 samples. As for the emission probability, 0.9 and 0.1 are set, respectively, for the normal noise mode and the outlier mode. Figure 4.7 shows the data before and after contamination with a switching noise in the above mentioned setting.

Comparison of the data distributions before and after contamination is shown in Figure 4.8, for two different scenarios. Figure 4.8(b) shows the case when ρ is bigger and the spread of contamination is more narrow since its inverse determines the inflation of the standard deviation for the contaminating noise mode. As expected, when there is more severe contamination as in Figure 4.8(a), there would be more data lying farther from the center. Figure 4.10 shows the prediction performance for product flowrate. To compare the effect of the robust model in this chapter and that of a robust model that was developed in Chapter 2 without considering the correlation of the switching sequence, see Figure 4.9. Comparison of Chapter 3 and this chapter might be an interesting topic to investigate. Although that might not be a fair comparison on all datasets since the contamination is assumed to be of a specific nature. Instead, advancement of the idea of Chapter 3 with a mixture switching noise model might be of interest for future researchers as a forward movement on this topic.

Table 4.2: Prediction performance of two robust models in Industrial case example

	Original data		Contaminated data ($\rho = 0.001$)	
	RPPCA-Static	RPPCA-Dynamic	RPPCA-Static	RPPCA-Dynamic
R^2	0.8865	0.8872	0.6817	0.8740
$MAPE$	0.0111	0.112	0.0205	0.0116
$RMSE$	0.4395	0.4381	0.7358	0.4631

Table 4.2 summarizes the results for the comparison of the two RPPCA-based algorithms. As the outcome of EM algorithm depends on the initial points, the parameters from a regular PPCA based model are used as initial points for the robust model parameter optimization. To make sure the results are representative, a Monte Carlo simulation was carried out and the average of all runs are reported. As Table 4.2 shows, when the original data is used, the performances are similar for both robust algorithms, regardless of the noise model assumption.

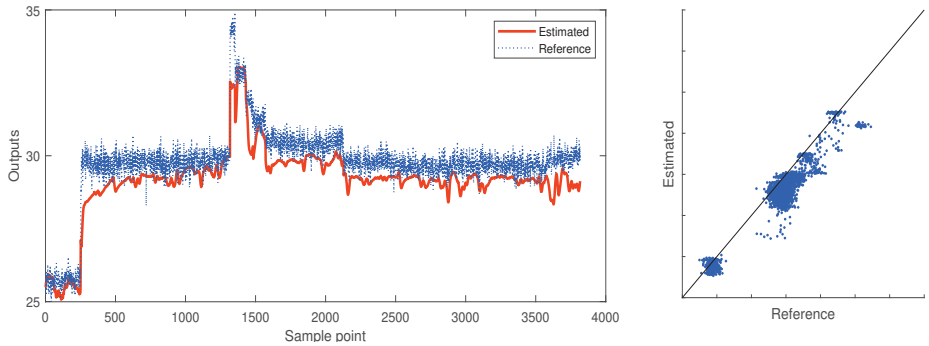


Figure 4.9: SAGD product oil flowrate prediction for previously developed RPPCA-based model

When feeding the algorithms by contaminated data with switching noise modes, the results indicate that the RPPCA proposed in this chapter can perform better while the RPPCA proposed in Chapter 2, has a poorer performance due to the fact that it does not consider the correlation of switching sequence. Nevertheless, the prediction performance and the extent of robustness will depend on the intensity and level of contamination.

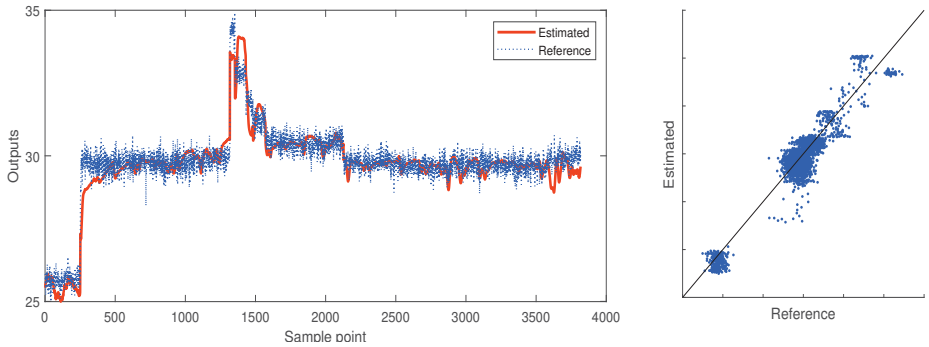


Figure 4.10: SAGD product oil flowrate prediction for proposed RPPCA-based model

4.6 Conclusion

This chapter advances the work presented in Chapter 2, by considering a more realistic scenario. A correlated switching behaviour is considered in the noise model of process measurements. Nonetheless, in real processes the number of noise modes could be more, depending on the complexity of the process and measurement devices. To describe normal noise and outliers, two noise modes are usually sufficient. The proposed robust model under this noise assumption was solved for parameters using EM algorithm. The solution was evaluated using a set of simulated data as well as a set of industrial data.

Chapter 5

Conclusions and Future Directions

"Resilience is accepting your new reality, even if it's less good than the one you had before. You can fight it, you can do nothing but scream about what you've lost, or you can accept that and try to put together something that's good."

Mary Elizabeth Anania Edwards (1949 - 2010)

5.1 Concluding Remarks

Different categories of outlying observation problems have been discussed. A Gaussian mixture distribution and a contaminated noise assumption were considered for PPCA based process models. This represents a more general class of distribution for explaining outlying observations while still maintaining the analytical properties of a Gaussian distribution for a closed-form solution. Accordingly, for different categories a robust probabilistic model was developed, based on which a better prediction performance for the desired quality variable could be obtained in the presence of large random errors in data. Unlike the conventional PPCA based models with a single Gaussian noise model, the developed robust models downweigh the effect of different styles of outlying observations in output prediction. The developed robust algorithms under these noise assumptions were solved for the model parameters using Expectation Maximization (EM) algorithm. Consideration of the Gaussian mixture noise model eases the process of getting closed-form solutions for model parameters, as well as downweighing the effect of the outlying noise in output prediction. Robustness and performance of the models were demonstrated through numerical and industrial case studies. Results confirm the robustness of the developed algorithms while the performance of other models breaks down in the presence of outlying observations.

5.2 Suggestions for Future Studies

Finally, there are more comprehensive directions of a robust PPCA-based algorithm development to be explored. A more general and inclusive assumption for the noise model which accommodates for other possible scenarios of outlying observations will expand the capability of the developed robust model. Some suggestions could be the consideration of a simultaneous switching dynamics for both input and output. Comparison of the algorithm introduced in Chapter 3 with a robust version of the algorithm that is developed under a switching noise assumption in Chapter 4 might be of interest for future contributions around this topic.

The switching itself can consist of more states to encompass a broader category of dynamic behaviour. States could be sourced from a scaled or location Gaussian

mixture or even a combination of both. This choice could be handed over to the user of the model for making the decision based on their knowledge of the process and its nature. Student's t-distribution assumption for the noise model is also common as its heavy tails are a representative of outlying process records. This thesis studies a mixture of Gaussian distributions for the sake of obtaining closed form solutions through this family of distributions. The ultimate case would be a Student's t-distribution which could be seen as an infinite number of Gaussian distributions in a mixture noise model. Laplace distribution can also be assumed as the noise model for its representative shape and properties.

Bibliography

[1] Hamza Albazzaz and Xue Z Wang. Historical data analysis based on plots of independent and parallel coordinates and statistical control limits. *Journal of Process Control*, 16(2):103–114, 2006.

[2] Özlem Gürünlü Alma. Comparison of robust regression methods in linear regression. *International Journal of Contemporary Mathematical Sciences*, 6(9):409–421, 2011.

[3] A Banerjee, Y Arkun, B Ogunnaike, and R Pearson. Estimation of nonlinear systems using linear multiple models. *AIChE Journal*, 43(5):1204–1226, 1997.

[4] Vic Barnett and Toby Lewis. *Outliers in statistical data*, volume 3. Wiley New York, 1994.

[5] Richard E Bellman. *Adaptive control processes*. Princeton university press, 2015.

[6] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

[7] SA Billings, HB Jamaluddin, and S Chen. Properties of neural networks with applications to modelling non-linear dynamical systems. *International Journal of Control*, 55(1):193–224, 1992.

[8] SA Blllings and WSF Voon. Correlation based model validity tests for non-linear models. *International journal of Control*, 44(1):235–244, 1986.

[9] John D Bomberger and Dale E Seborg. Determination of model order for narx models directly from input-output data. *Journal of Process Control*, 8(5-6):459–468, 1998.

[10] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[11] Lei Chen, Aditya Tulsyan, Biao Huang, and Fei Liu. Multiple model approach to nonlinear system identification with an uncertain scheduling variable using em algorithm. *Journal of Process Control*, 23(10):1480–1496, 2013.

[12] Sheng Chen, Stephen A Billings, and PM Grant. Non-linear system identification using neural networks. *International journal of control*, 51(6):1191–1214, 1990.

[13] Leo H Chiang, Randy J Pell, and Mary Beth Seasholtz. Exploring process data with the use of robust outlier detection algorithms. *Journal of Process Control*, 13(5):437–449, 2003.

[14] Wai Ki Ching, Eric S Fung, and Michael K Ng. Higher-order hidden markov models with applications to dna sequences. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 535–539. Springer, 2003.

[15] P Laurie Davies. Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, pages 1269–1292, 1987.

[16] Jing Deng, Li Xie, Lei Chen, Shima Khatibisepehr, Biao Huang, Fangwei Xu, and Aris Espejo. Development and industrial application of soft sensors with online bayesian model updating strategy. *Journal of Process Control*, 23(3):317–325, 2013.

[17] Ming Dong and David He. Hidden semi-markov model-based methodology for multi-sensor equipment health diagnosis and prognosis. *European Journal of Operational Research*, 178(3):858–878, 2007.

[18] David L Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford , 1982.

[19] Cristofer Englund and Antanas Verikas. A hybrid approach to outlier detection in the offset lithographic printing process. *Engineering Applications of Artificial Intelligence*, 18(6):759–768, 2005.

[20] B Everett. *An introduction to latent variable models*. Springer Science & Business Media, 2013.

[21] Yi Fang and Myong K Jeong. Robust probabilistic multivariate calibration model. *Technometrics*, 50(3):305–316, 2008.

[22] Balazs Feil, Janos Abonyi, Peter Pach, Sandor Nemeth, Peter Arva, Miklos Nemeth, and Gabor Nagy. Semi-mechanistic models for state-estimation–soft sensor for polymer melt index prediction. In *International Conference on Artificial Intelligence and Soft Computing*, pages 1111–1117. Springer, 2004.

[23] D Flynn, J Ritchie, and M Cregan. Data mining techniques applied to power plant performance monitoring. *IFAC Proceedings Volumes*, 38(1):369–374, 2005.

[24] L. Fortuna, S. Graziani, A. Rizzo, and M.G. Xibilia. *Soft Sensors for Monitoring and Control of Industrial Processes*. Advances in Industrial Control. Springer, 2007.

[25] W David Furman and Bruce G Lindsay. Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Computational statistics & data analysis*, 17(5):493–507, 1994.

[26] W David Furman and Bruce G Lindsay. Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics & Data Analysis*, 17(5):473–492, 1994.

[27] Zhiqiang Ge, Biao Huang, and Zhihuan Song. Mixture semisupervised principal component regression model and soft sensor application. *AIChE Journal*, 60(2):533–545, 2014.

[28] Alireza Ghasemi, Soumaya Yacout, and M-Salah Ouali. Parameter estimation methods for condition-based maintenance with indirect observations. *IEEE Transactions on reliability*, 59(2):426–439, 2010.

[29] Bonnie Ghosh-Dastidar and Joseph L Schafer. Outlier detection and editing procedures for continuous multivariate data. *Journal of Official Statistics*, 22(3):487, 2006.

[30] Diana Glassman, Michele Wucker, Tanushree Isaacman, and Corinne Champilou. The water-energy nexus. *Adding Water to the Energy Agenda, A World Policy Paper, EBG Capital, Environmental Investments*, 2011.

[31] Frank E Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[32] Roberto Guidorzi. *Multivariable system identification: from observations to models*. Bononia University Press, 2003.

[33] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[34] Peter J Huber. Between robustness and diagnostics. *Institute for Mathematics and Its Applications*, 33:121, 1991.

[35] Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[36] Mia Hubert and Sabine Verboven. A robust pcr method for high-dimensional regressors. *Journal of Chemometrics*, 17(8-9):438–452, 2003.

[37] Rui Jiang, Jing Yu, and Viliam Makis. Optimal bayesian estimation and control scheme for gear shaft fault detection. *Computers & Industrial Engineering*, 63(4):754–762, 2012.

[38] Xing Jin and Biao Huang. Robust identification of piecewise/switching autoregressive exogenous process. *AIChE journal*, 56(7):1829–1844, 2010.

[39] Xing Jin, Biao Huang, and David S Shook. Multiple model lpv approach to nonlinear process identification with em algorithm. *Journal of Process Control*, 21(1):182–193, 2011.

[40] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[41] Anatoli Juditsky, Håkan Hjalmarsson, Albert Benveniste, Bernard Delyon, Lennart Ljung, Jonas Sjöberg, and Qinghua Zhang. Nonlinear black-box models in system identification: Mathematical foundations. *Automatica*, 31(12):1725–1750, 1995.

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

[43] Dimitris Karlis and Evdokia Xekalaki. Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3):577–590, 2003.

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

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

[46] Minjin Kim, Young-Hak Lee, In-Su Han, and Chonghun Han. Clustering-based hybrid soft sensor for an industrial polypropylene process with grade changeover operation. *Industrial & engineering chemistry research*, 44(2):334–342, 2005.

[47] Tiina Komulainen, Mauri Sourander, and Sirkka-Liisa Jämsä-Jounela. An online application of dynamic pls to a dearomatization process. *Computers & Chemical Engineering*, 28(12):2611–2619, 2004.

[48] Farinaz Koushanfar and Miodrag Potkonjak. Markov chain-based models for missing and faulty data in mica2 sensor motes. In *SENSORS, 2005 IEEE*, pages 4–pp. IEEE, 2005.

[49] Mark A Kramer. Model-based monitoring. In *Proceedings of 2nd Conference on Foundations of Computer Aided Process Operations*, pages 1–24. Crested Butte, 1993.

[50] Wojciech Kwedlo. A new method for random initialization of the em algorithm for multivariate gaussian mixture learning. In *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*, pages 81–90. Springer, 2013.

[51] Machine Learning. An introduction to conditional random fields. *Mach. Learn*, 4(4):267–373, 2011.

[52] Jaeshin Lee, Bokyoung Kang, and Suk-Ho Kang. Integrating independent component analysis and local outlier factor for plant-wide process monitoring. *Journal of Process Control*, 21(7):1011–1021, 2011.

[53] Jong-Min Lee, ChangKyoo Yoo, and In-Beum Lee. Statistical process monitoring with independent component analysis. *Journal of process control*, 14(5):467–485, 2004.

[54] B Li. Project review: Soft sensors. Technical report, Technical Report, University of Alberta, Alberta, Canada, 2005.

[55] Zhinong Li, Yongyong He, Fulei Chu, Jie Han, and Wei Hao. Fault recognition method for speed-up and speed-down process of rotating machinery based on independent component analysis and factorial hidden markov model. *Journal of Sound and Vibration*, 291(1-2):60–71, 2006.

[56] Bao Lin, Bodil Recke, Philippe Renaudat, Jørgen Knudsen, and Sten Bay Jørgensen. Robust statistics for soft sensor development in cement kiln. *IFAC Proceedings Volumes*, 38(1):241–246, 2005.

[57] Mary J Lindstrom and Douglas M Bates. NewtonÂraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

[58] Hancong Liu, Sirish Shah, and Wei Jiang. On-line outlier detection and data cleaning. *Computers & chemical engineering*, 28(9):1635–1647, 2004.

[59] L Ljung. System identification-theory for the user 2nd edition ptr prentice-hall. *Upper Saddle River, NJ*, 1999.

[60] Hendrik P Lopuhaa. On the relation between s-estimators and m-estimators of multivariate location and covariance. *The Annals of Statistics*, pages 1662–1683, 1989.

[61] Richard J Martin. Leverage, influence and residuals in regression models when observations are correlated. *Communications in statistics-theory and methods*, 21(5):1183–1212, 1992.

[62] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[63] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[64] Volodymyr Melnykov and Igor Melnykov. Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, 56(6):1381–1395, 2012.

[65] Ana F Militino and M Dolores Ugarte. Bounded influence estimation in a spatial linear mixed model. In *Modelling Longitudinal and Spatially Correlated Data*, pages 211–220. Springer, 1997.

[66] S Frosch Møller, Jürgen von Frese, and Rasmus Bro. Robust methods for multivariate data analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 19(10):549–563, 2005.

[67] Philip RC Nelson, Paul A Taylor, and John F MacGregor. Missing data methods in pca and pls: Score calculations with incomplete observations. *Chemometrics and intelligent laboratory systems*, 35(1):45–65, 1996.

[68] Alberta Energy-Government of Alberta. *Talk About SAGD. Oil sands Facts Sheet*. Available at http://www.energy.alberta.ca/OilSands/pdfs, 2013.

[69] Karla Patricia Oliveira-Esquerre, Dale E Seborg, Roy E Bruns, and Milton Mori. Application of steady-state and dynamic modeling for the prediction of the bod of an aerated lagoon at a pulp and paper mill: Part i. linear approaches. *Chemical Engineering Journal*, 104(1-3):73–81, 2004.

[70] Ronald K Pearson. Outliers in process modeling and identification. *IEEE Transactions on control systems technology*, 10(1):55–63, 2002.

[71] Ramani S Pilla and Bruce G Lindsay. Alternative em methods for nonparametric finite mixture models. *Biometrika*, 88(2):535–550, 2001.

[72] Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

[73] Peter J Rousseeuw, Stefan Van Aelst, Katrien Van Driessen, and Jose A Gulló. Robust multivariate regression. *Technometrics*, 46(3), 2004.

[74] A Sadeghian, O Wu, and B Huang. Robust probabilistic principal component analysis based process modeling: Dealing with simultaneous contamination of both input and output data. *Journal of Process Control*, 67:94–111, 2018.

[75] Anahita Sadeghian and Biao Huang. Robust probabilistic principal component analysis for process modeling subject to scaled mixture gaussian noise. *Computers & Chemical Engineering*, 90:62–78, 2016.

[76] Nima Sammaknejad. Fault detection and isolation based on hidden markov models. 2015.

[77] Nima Sammaknejad, Biao Huang, and Yaojie Lu. Robust diagnosis of operating mode based on time-varying hidden markov models. *IEEE Transactions on Industrial Electronics*, 63(2):1142–1152, 2015.

[78] Nima Sammaknejad, Biao Huang, Weili Xiong, Alireza Fatehi, Fangwei Xu, and Aris Espejo. Operating condition diagnosis based on hmm with adaptive transition probabilities in presence of missing observations. *AIChE Journal*, 61(2):477–493, 2015.

[79] Shabnam Sedghi, Anahita Sadeghian, and Biao Huang. Mixture semisupervised probabilistic principal component regression model with missing inputs. *Computers & Chemical Engineering*, 103:176–187, 2017.

[80] Wilfried Seidel, Karl Mosler, and Manfred Alker. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52(3):481–487, 2000.

[81] Jonas Sjöberg, Håkan Hjalmarsson, and Lennart Ljung. Neural networks in system identification. *IFAC Proceedings Volumes*, 27(8):359–382, 1994.

[82] Jonas Sjöberg, Qinghua Zhang, Lennart Ljung, Albert Benveniste, Bernard Delyon, Pierre-Yves Glorennec, Håkan Hjalmarsson, and Anatoli Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691–1724, 1995.

[83] Padhraic Smyth. Hidden markov models for fault detection in dynamic systems. *Pattern recognition*, 27(1):149–164, 1994.

[84] WA Stahel. Robust estimation: Infinitesimal optimality and covariance matrix estimators. *Unpublished doctoral dissertation, ETH, Zurich, Switzerland*, 1981.

[85] Ioan Stanculescu, Christopher KI Williams, and Yvonne Freer. Autoregressive hidden markov models for the early detection of neonatal sepsis. *IEEE journal of biomedical and health informatics*, 18(5):1560–1570, 2013.

[86] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[87] John W Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, pages 1–67, 1962.

[88] Jannie SJ van Deventer, Kiew M Kam, and Tjaart J van der Walt. Dynamic modelling of a carbon-in-leach process with the regression network. *Chemical engineering science*, 59(21):4575–4589, 2004.

[89] Athanasios Voulodimos, Helmut Grabner, Dimitrios Kosmopoulos, Luc Van Gool, and Theodora Varvarigou. Robust workflow recognition using holistic features and outlier-tolerant fused hidden markov models. In *International Conference on Artificial Neural Networks*, pages 551–560. Springer, 2010.

[90] Karol Warne, Girijesh Prasad, Sina Rezvani, and L Maguire. Statistical and computational intelligence techniques for inferential model development: a comparative evaluation and a novel proposition for fusion. *Engineering applications of artificial intelligence*, 17(8):871–885, 2004.

[91] Svante Wold, Anders Berglund, and Nouna Kettaneh. New and old trends in chemometrics. how to deal with the increasing data volumes in r&d&p (research, development and production)Âwith examples from pharmaceutical research and process modeling. *Journal of chemometrics*, 16(8-10):377–386, 2002.

[92] James C Wong, Karen A McDonald, and Ahmet Palazoglu. Classification of process trends based on fuzzified symbolic representation and hidden markov models. *Journal of Process Control*, 8(5-6):395–408, 1998.

[93] Wee Chin Wong and Jay H Lee. Fault detection and diagnosis using hidden markov disturbance models. *Industrial & Engineering Chemistry Research*, 49(17):7901–7908, 2010.

[94] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[95] Jie Ying, Thia Kirubarajan, Krishna R Pattipati, and Ann Patterson-Hine. A hidden markov model-based algorithm for fault diagnosis with partial and imperfect tests. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):463–473, 2000.

[96] Jiu-sun Zeng and Chuan-hou Gao. Improvement of identification of blast furnace ironmaking process by outlier detection and missing value imputation. *Journal of Process Control*, 19(9):1519–1528, 2009.

[97] Temesgen Zewotir and Jacky S Galpin. A unified approach on residuals, leverages and outliers in the linear mixed model. *Test*, 16(1):58–75, 2007.

[98] Yu Zhao, Biao Huang, Hongye Su, and Jian Chu. Prediction error method for identification of lpv models. *Journal of process control*, 22(1):180–193, 2012.

[99] Le Zhou, Junghui Chen, Zhihuan Song, Zhiqiang Ge, and Aimin Miao. Probabilistic latent variable regression model for process-quality monitoring. *Chemical Engineering Science*, 116:296–305, 2014.

[100] Jinlin Zhu, Zhiqiang Ge, and Zhihuan Song. Dynamic mixture probabilistic pca classifier modeling and application for fault classification. *Journal of Chemometrics*, 29(6):361–370, 2015.

[101] Jinlin Zhu, Zhiqiang Ge, and Zhihuan Song. Hmm-driven robust probabilistic principal component analyzer for dynamic process fault classification. *IEEE Transactions on Industrial Electronics*, 62(6):3814–3821, 2015.

[102] Yucai Zhu and Zuhua Xu. A method of lpv model identification for control. *IFAC Proceedings Volumes*, 41(2):5018–5023, 2008.

Appendices

Appendix I

$$
\begin{aligned}
\underset{*\$}{A^0}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \;=\; & (x_i - \mu_x)^T (x_i - \mu_x) \\
& - \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T P^T (x_i - \mu_x) \\
& - (x_i - \mu_x)^T P \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
& + tr\left(P^T P \underset{*\$}{S}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \right) \\
& + \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T P^T \cdots \\
& \cdots \times P \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \,,
\end{aligned}
$$

$$
\begin{aligned}
\underset{*\$}{A^{\Delta}}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \;=\; & (x_i - \mu_x - \Delta_x)^T (x_i - \mu_x - \Delta_x) \\
& - \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T P^T (x_i - \mu_x - \Delta_x) \\
& - (x_i - \mu_x - \Delta_x)^T P \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
& + tr\left(P^T P \underset{*\$}{S}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \right) \\
& + \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T P^T \cdots \\
& \cdots \times P \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \,,
\end{aligned}
$$

$$
\begin{aligned}
\underset{*\$}{A^{-\Delta}}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \;=\; & (x_i - \mu_x + \Delta_x)^T (x_i - \mu_x + \Delta_x) \\
& - \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T P^T (x_i - \mu_x + \Delta_x) \\
& - (x_i - \mu_x + \Delta_x)^T P \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
& + tr\left(P^T P \underset{*\$}{S}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \right) \\
& + \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T P^T \cdots \\
& \cdots \times P \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \,.
\end{aligned}
$$

Appendix II

$$
\begin{aligned}
C^{new} \;=\; & \Bigg[\sum_{i=1}^{n}\Bigg(2(y_i-\mu_y)\big(P_1 \underset{00}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=0,\theta^{old})^T \\
& \qquad +P_4 \underset{\Delta 0}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=0,\theta^{old})^T \\
& \qquad +P_7 \underset{-\Delta 0}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=0,\theta^{old})^T\big) \\
& +2\rho_y(y_i-\mu_y-\Delta_y)\big(P_2 \underset{0\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old})^T \\
& \qquad +P_5 \underset{\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old})^T \\
& \qquad +P_8 \underset{-\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old})^T\big) \\
& +2\rho_y(y_i-\mu_y+\Delta_y)\big(P_3 \underset{0-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old})^T \\
& \qquad +P_6 \underset{\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})^T \\
& \qquad +P_9 \underset{-\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})^T\big)\Bigg)\Bigg] \\
& \times\Bigg[\sum_{i=1}^{n}\Bigg(P_1 \underset{00}{Coeff}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=0,\theta^{old}) \\
& \qquad +P_4 \underset{\Delta 0}{Coeff}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=0,\theta^{old}) \\
& \qquad +P_7 \underset{-\Delta 0}{Coeff}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=0,\theta^{old}) \\
& \qquad +\rho_y P_2 \underset{0\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old}) \\
& \qquad +\rho_y P_5 \underset{\Delta\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old}) \\
& \qquad +\rho_y P_8 \underset{-\Delta\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old}) \\
& \qquad +\rho_y P_3 \underset{0-\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old}) \\
& \qquad +\rho_y P_6 \underset{\Delta-\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old}) \\
& \qquad +\rho_y P_9 \underset{-\Delta-\Delta}{Coeff}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})\Bigg)\Bigg]^{-1}.
\end{aligned}
$$

$$
\begin{aligned}
{\sigma_y^2}^{new} \;=\; & \frac{1}{n\,r}\sum_{i=1}^{n}\Bigg(P_1\,\underset{00}{B}^{0}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=0,\theta^{old}) \\
& +P_4\,\underset{\Delta 0}{B}^{0}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=0,\theta^{old}) \\
& +P_7\,\underset{-\Delta 0}{B}^{0}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=0,\theta^{old}) \\
& +\rho_y P_2\,\underset{0\Delta}{B}^{\Delta}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old}) \\
& +\rho_y P_5\,\underset{\Delta\Delta}{B}^{\Delta}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old}) \\
& +\rho_y P_8\,\underset{-\Delta\Delta}{B}^{\Delta}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old}) \\
& +\rho_y P_3\,\underset{0-\Delta}{B}^{-\Delta}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old}) \\
& +\rho_y P_6\,\underset{\Delta-\Delta}{B}^{-\Delta}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old}) \\
& +\rho_y P_9\,\underset{-\Delta-\Delta}{B}^{-\Delta}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})\Bigg),
\end{aligned}
$$

where, $\underset{*\$}{B}$s for each case are as formulated at the end of this section.

$$
\begin{aligned}
\mu_y^{new} \;=\; & \frac{1}{n}\sum_{i=1}^{n}\Bigg(-P_1(-y_i+C\,\underset{00}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=0,\theta^{old})) \\
& -P_4(-y_i+C\,\underset{\Delta 0}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=0,\theta^{old})) \\
& -P_7(-y_i+C\,\underset{-\Delta 0}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=0,\theta^{old})) \\
& -P_2\rho_y(\Delta_y-y_i+C\,\underset{0\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old})) \\
& -P_5\rho_y(\Delta_y-y_i+C\,\underset{\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old})) \\
& -P_8\rho_y(\Delta_y-y_i+C\,\underset{-\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old})) \\
& -P_3\rho_y(-\Delta_y-y_i+C\,\underset{0-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old})) \\
& -P_6\rho_y(-\Delta_y-y_i+C\,\underset{\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})) \\
& -P_9\rho_y(-\Delta_y-y_i+C\,\underset{-\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old}))\Bigg).
\end{aligned}
$$

$$
\delta_y^{new}=\frac{\sum_{i=1}^{n}(P_2+P_5+P_8+P_3+P_6+P_9)}{n},
$$

$$
\begin{aligned}
\Delta_y^{new} &= (\frac{1}{\sum_{i=1}^{n}(P_2+P_5+P_8+P_3+P_6+P_9)}) \\
&\times \sum_{i=1}^{n}\Bigg(P_2(y_i-\mu_y-C\underset{0\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old})) \\
&+P_5(y_i-\mu_y-C\underset{\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old})) \\
&+P_8(y_i-\mu_y-C\underset{-\Delta\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old})) \\
&-P_3(y_i-\mu_y-C\underset{0-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old})) \\
&-P_6(y_i-\mu_y-C\underset{\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})) \\
&-P_9(y_i-\mu_y-C\underset{-\Delta-\Delta}{E}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old}))\Bigg),
\end{aligned}
$$

$$
\begin{aligned}
\rho_y^{new} &= r\sigma_y^2\Bigg[\sum_{i=1}^{n}(P_2+P_5+P_8+P_3+P_6+P_9)\Bigg] \\
&\times\Bigg[\sum_{i=1}^{n}\Bigg(P_2\,\underset{0\Delta}{B^{\Delta}}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=\Delta_y,\theta^{old}) \\
&\quad+P_5\,\underset{\Delta\Delta}{B^{\Delta}}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=\Delta_y,\theta^{old}) \\
&\quad+P_8\,\underset{-\Delta\Delta}{B^{\Delta}}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=\Delta_y,\theta^{old}) \\
&\quad+P_3\,\underset{0-\Delta}{B^{-\Delta}}(t_i|x_i,y_i,q_{x_i}=0,q_{y_i}=-\Delta_y,\theta^{old}) \\
&\quad+P_6\,\underset{\Delta-\Delta}{B^{-\Delta}}(t_i|x_i,y_i,q_{x_i}=\Delta_x,q_{y_i}=-\Delta_y,\theta^{old}) \\
&\quad+P_9\,\underset{-\Delta-\Delta}{B^{-\Delta}}(t_i|x_i,y_i,q_{x_i}=-\Delta_x,q_{y_i}=-\Delta_y,\theta^{old})\Bigg)\Bigg]^{-1},
\end{aligned}
$$

$$
\begin{aligned}
\underset{*\$}{B^0}(t_i|x_i,y_i,q_{x_i}=*,q_{y_i}=\$,\theta^{old}) &= (y_i-\mu_y)^T(y_i-\mu_y) \\
&\quad-\underset{*\$}{E}(t_i|x_i,y_i,q_{x_i}=*,q_{y_i}=\$,\theta^{old})^TC^T(y_i-\mu_y) \\
&\quad-(y_i-\mu_y)^TC\,\underset{*\$}{E}(t_i|x_i,y_i,q_{x_i}=*,q_{y_i}=\$,\theta^{old}) \\
&\quad+tr\Bigg(C^TC\,\underset{*\$}{S}(t_i|x_i,y_i,q_{x_i}=*,q_{y_i}=\$,\theta^{old})\Bigg) \\
&\quad+\underset{*\$}{E}(t_i|x_i,y_i,q_{x_i}=*,q_{y_i}=\$,\theta^{old})^TC^T\cdots \\
&\quad\cdots C\,\underset{*\$}{E}(t_i|x_i,y_i,q_{x_i}=*,q_{y_i}=\$,\theta^{old}),
\end{aligned}
$$

$$
\begin{aligned}
B^{\Delta}_{*\$}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \;=\; & (y_i - \mu_y - \Delta_y)^T (y_i - \mu_y - \Delta_y) \\
& - \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T C^T (y_i - \mu_y - \Delta_y) \\
& - (y_i - \mu_y - \Delta_y)^T C \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
& + tr\left(C^T C \underset{*\$}{S}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \right) \\
& + \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T C^T \cdots \\
& \cdots C \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \,,
\end{aligned}
$$

$$
\begin{aligned}
B^{-\Delta}_{*\$}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \;=\; & (y_i - \mu_y + \Delta_y)^T (y_i - \mu_y + \Delta_y) \\
& - \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T C^T (y_i - \mu_y + \Delta_y) \\
& - (y_i - \mu_y + \Delta_y)^T C \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \\
& + tr\left(C^T C \underset{*\$}{S}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \right) \\
& + \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old})^T C^T \cdots \\
& \cdots \times C \underset{*\$}{E}(t_i|x_i, y_i, q_{x_i} = *, q_{y_i} = \$, \theta^{old}) \,.
\end{aligned}
$$

Appendix III

$$\underset{0\Delta}{E}(t_i|x_i, y_i, q_{x_i} = 0, q_{y_i} = \Delta_y, \theta^{old}) = (\sigma_x^{-2}P^TP + \rho_y\sigma_y^{-2}C^TC + I)^{-1} (\sigma_x^{-2}P^T(x_i - \mu_x) + \rho_y\sigma_y^{-2}C^T(y_i - (\mu_y + \Delta_y)))$$

$$\underset{0-\Delta}{E}(t_i|x_i, y_i, q_{x_i} = 0, q_{y_i} = -\Delta_y, \theta^{old}) = (\sigma_x^{-2}P^TP + \rho_y\sigma_y^{-2}C^TC + I)^{-1} (\sigma_x^{-2}P^T(x_i - \mu_x) + \rho_y\sigma_y^{-2}C^T(y_i - (\mu_y - \Delta_y)))$$

$$\underset{\Delta 0}{E}(t_i|x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = 0, \theta^{old}) = (\rho_x\sigma_x^{-2}P^TP + \sigma_y^{-2}C^TC + I)^{-1} (\rho_x\sigma_x^{-2}P^T(x_i - (\mu_x + \Delta_x)) + \sigma_y^{-2}C^T(y_i - \mu_y))$$

$$\underset{\Delta\Delta}{E}(t_i|x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = \Delta_y, \theta^{old}) = (\rho_x\sigma_x^{-2}P^TP + \rho_y\sigma_y^{-2}C^TC + I)^{-1} (\rho_x\sigma_x^{-2}P^T(x_i - (\mu_x + \Delta_x)) \cdots \cdots + \rho_y\sigma_y^{-2}C^T(y_i - (\mu_y + \Delta_y)))$$

$$\underset{\Delta-\Delta}{E}(t_i|x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = -\Delta_y, \theta^{old}) = (\rho_x\sigma_x^{-2}P^TP + \rho_y\sigma_y^{-2}C^TC + I)^{-1} (\rho_x\sigma_x^{-2}P^T(x_i - (\mu_x + \Delta_x)) \cdots \cdots + \rho_y\sigma_y^{-2}C^T(y_i - (\mu_y - \Delta_y)))$$

$$\underset{-\Delta 0}{E}(t_i|x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = 0, \theta^{old}) = (\rho_x\sigma_x^{-2}P^TP + \sigma_y^{-2}C^TC + I)^{-1} (\rho_x\sigma_x^{-2}P^T(x_i - (\mu_x - \Delta_x)) + \sigma_y^{-2}C^T(y_i - \mu_y))$$

$$\underset{-\Delta\Delta}{E}(t_i|x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = \Delta_y, \theta^{old}) = (\rho_x\sigma_x^{-2}P^TP + \rho_y\sigma_y^{-2}C^TC + I)^{-1} (\rho_x\sigma_x^{-2}P^T(x_i - (\mu_x - \Delta_x)) \cdots \cdots + \rho_y\sigma_y^{-2}C^T(y_i - (\mu_y + \Delta_y)))$$

$$\underset{-\Delta-\Delta}{E}(t_i|x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = -\Delta_y, \theta^{old}) = (\rho_x\sigma_x^{-2}P^TP + \rho_y\sigma_y^{-2}C^TC + I)^{-1} (\rho_x\sigma_x^{-2}P^T(x_i - (\mu_x - \Delta_x)) \cdots \cdots + \rho_y\sigma_y^{-2}C^T(y_i - (\mu_y - \Delta_y))) .$$

Appendix IV

$$\underset{0\Delta}{E}(t_i t_i^T | x_i, y_i, q_{x_i} = 0, q_{y_i} = \Delta_y, \theta^{old}) = (\sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} + \underset{0\Delta}{E}(t_i | x_i, y_i, q_{x_i} = 0, q_{y_i} = \Delta_y, \theta^{old}) \cdots \cdots \underset{0\Delta}{E}(t_i | x_i, y_i, q_{x_i} = 0, q_{y_i} = \Delta_y, \theta^{old})^T$$

$$\underset{0-\Delta}{E}(t_i t_i^T | x_i, y_i, q_{x_i} = 0, q_{y_i} = -\Delta_y, \theta^{old}) = (\sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} + \underset{0-\Delta}{E}(t_i | x_i, y_i, q_{x_i} = 0, q_{y_i} = -\Delta_y, \theta^{old}) \cdots \cdots \times \underset{0-\Delta}{E}(t_i | x_i, y_i, q_{x_i} = 0, q_{y_i} = -\Delta_y, \theta^{old})^T$$

$$\underset{\Delta 0}{E}(t_i t_i^T | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = 0, \theta^{old}) = (\rho_x \sigma_x^{-2} P^T P + \sigma_y^{-2} C^T C + I)^{-1} + \underset{\Delta 0}{E}(t_i | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = 0, \theta^{old}) \cdots \cdots \underset{\Delta 0}{E}(t_i | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = 0, \theta^{old})^T$$

$$\underset{\Delta\Delta}{E}(t_i t_i^T | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = \Delta_y, \theta^{old}) = (\rho_x \sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} + \underset{\Delta\Delta}{E}(t_i | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = \Delta_y, \theta^{old}) \cdots \cdots \times \underset{\Delta\Delta}{E}(t_i | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = \Delta_y, \theta^{old})^T$$

$$\underset{\Delta-\Delta}{E}(t_i t_i^T | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = -\Delta_y, \theta^{old}) = (\rho_x \sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} + \underset{\Delta-\Delta}{E}(t_i | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = -\Delta_y, \theta^{old}) \cdots \cdots \times \underset{\Delta-\Delta}{E}(t_i | x_i, y_i, q_{x_i} = \Delta_x, q_{y_i} = -\Delta_y, \theta^{old})^T$$

$$\underset{-\Delta 0}{E}(t_i t_i^T | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = 0, \theta^{old}) = (\rho_x \sigma_x^{-2} P^T P + \sigma_y^{-2} C^T C + I)^{-1} + \underset{-\Delta 0}{E}(t_i | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = 0, \theta^{old}) \cdots \cdots \times \underset{-\Delta 0}{E}(t_i | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = 0, \theta^{old})^T \quad (10)$$

$$\underset{-\Delta\Delta}{E}(t_i t_i^T | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = \Delta_y, \theta^{old}) = (\rho_x \sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} + \underset{-\Delta\Delta}{E}(t_i | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = \Delta_y, \theta^{old}) \cdots \cdots \times \underset{-\Delta\Delta}{E}(t_i | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = \Delta_y, \theta^{old})^T$$

$$\begin{aligned}
\underset{-\Delta-\Delta}{E}(t_i t_i^T | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = -\Delta_y, \theta^{old}) &= (\rho_x \sigma_x^{-2} P^T P + \rho_y \sigma_y^{-2} C^T C + I)^{-1} \\
&\quad + \underset{-\Delta-\Delta}{E}(t_i | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = -\Delta_y, \theta^{old}) \cdots \\
&\quad \cdots \times \underset{-\Delta-\Delta}{E}(t_i | x_i, y_i, q_{x_i} = -\Delta_x, q_{y_i} = -\Delta_y, \theta^{old})^T
\end{aligned}$$