University of Alberta

Change Point Detection Using Expectation Maximization Approach

by

Marziyeh Keshavarz

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

Master of Science
in
Process Control

Department of Chemical and Materials Engineering

©Marziyeh Keshavarz

Fall 2013
Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

To my parents and my husband for their support and unconditional love

Abstract

Data analysis plays an important role in system modeling, monitoring and optimization. Among those data analysis techniques, change point detection has been widely applied in various areas including chemical process, climate monitoring, examination of gene expressions and quality control in the manufacturing industry, etc. In this thesis, an Expectation Maximization (EM) algorithm is proposed to detect the time instants at which data properties are subject to change. This method performs efficiently especially in missing data problem or when directly maximizing the likelihood is difficult. The change point detection problem is solved under various scenarios including univariate and multivariate data, known and unknown covariance. The problem is also extended to changing covariance in the case of multivariate data analysis. Moreover, using Bayesian inference method these problems are solved and the results are compared with EM. The results show that in terms of computation, due to some iterations involved in EM algorithm, it has higher computation but the convergence is fast. In the presence of uncertain hyperparameters of missing variables (in EM formulation) or priors (in Bayesian method), EM outperforms Bayesian method.

Besides, using change point models, different unknown properties of data such as mean and covariance can be estimated in the context of EM algorithm. In this thesis, assuming change points as missing variables, the mean vectors in every segment of data are estimated. This estimation is extended to constrained parameter space for both linear and nonlinear case studies. Using simulation examples, it is shown that EM performance is satisfactory leading to accurate estimation along with fast convergence.

Acknowledgements

First of all, I would like to thank Dr. Biao Huang for his support and advice. During my research, he always encouraged me at different stages of my work. Without his recommendations and support, these achievements were too far to reach. His outstanding knowledge, insight and patience always opened new horizons to me during my master program.

Next, I would also like to thank Da Zheng for his support and help during running the hybrid tank experiments in the lab. Besides, special thanks to Li Xie who broadened my insight about my research.

I also would like to acknowledge my friend in our EM sub-group meeting. Our weekly meeting and discussions with Tianbo Liu, Aditya Tusln, Lei Chen always were inspiring.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Accuracy of process measurements is critical in the sense of safe and continuous operations. To achieve this goal, timely detecting and repairing of faulty instruments plays a significant role in terms of instrument availability and reliability. These problems along with other factors such as a major process upset, a change in equipment performance, etc. may all lead to change such as mean shift in which a bias is introduced in steady state operating data.

Various techniques in literature have been developed to detect and estimate bias error in the instruments. These include node test, measurement test, likelihood ratio test, PCA, etc. Depending on the problem and available information of the process including models and constraints, one may take advantage of a certain methodology.

Basically, numerous factors can lead to abnormality of data such as miscalibration or malfunctioning of instruments, instrument biases or process leak. Thus, developing an efficient method to identify these types of systematic errors or gross errors is of interest for process industry. An excellent survey and review of single or multiple gross error detection methods with their applications can be found in [1]. In all of these methods, process constraints such as material balance, energy balance, etc. are taken into account.

On the other hand, due to problems raised in derivation of process models, data-driven methods have been widely used in recent years. For instance, data-based approach to detection of the time instant at which bias is introduced to measurements has received great attention especially in applications such as hydrology, signal processing, finance, economics, pharmacology, environmental studies, meteorology and etc [2]. In chemical engineering applications, there are a lot of sensors or instruments which may be subject to bias or drift. Detection of time instants where these biases or drifts are introduced is important in instrument fault identification. Detection of instants where system operating mode changes can be another application of change detection. These problems are often formulated as change point detection. At these change points, the mean of data shifts. As detection of these change points is performed, new mean and hence bias magnitude can also be determined accordingly.

Various methods have been developed in literature to tackle the problem of change point detection. A good review of these methods can be found in [2] [3] [4]. One of the elementary methods in change detection is random-size sliding window algorithm. In quality control, these methods are called finite or infinite moving average control charts [2] in which higher weights on recent observations and lower weights on past ones are used. Among various approaches, probabilistic frameworks, such as Bayesian inference, have been applied in various areas. These approaches are powerful in the sense that one can incorporate priori knowledge in estimation of unknown parameters. In [5] and [6], a Bayesian approach is used to detect gross errors based on process models. This method is applied sequentially over various time periods of data by updating the priors and posteriors at the end of each period. Computation of this method for medium to large problems is intensive despite the modification made by the authors of [6]. In [7], a Bayesian decision rule is developed to detect the change point in univariate data which needs selection of prior distributions for unknown parameters and then derivation of posterior probability of shift point given the data.

The multivariate version of Bayesian single change point detection can be found in [8] [9] [10] [11][12][13]. In [13], an empirical Bayes stopping time is studied for detection of a change in distribution of data when prior is not completely known. Multiple change points detection is investigated through methods such as hypothesis testing [14], maximum likelihood [15] and a clustering-based algorithm called product partition model (PPM)[16] [17]. In PPM technique, the prior probabilities for a random partition are determined. As a result, the posterior probability of the partition is of the same form. In essence, in this method, a large amount of computations result from Markov sampling to determine the estimates of means, derived by conditioning on the partition and summing over all possible partitions.

In [18], the change points are determined by minimizing a penalised contrast function which measures how the model, derived based on change point sequence, fits the observed data.

Generally, among methods applied in change point detection, Bayesian approach has been most widely adopted. There are two approaches proposed in literature to solve the change point detection problem based on the Bayesian approach. One approach relies on finding the mode of posterior probability called Maximum a Posteriori (MAP) approach which is optimization-based. The other is to calculate means of various posterior probabilities, which leads to integration calculation. Basically, these integrations are difficult to solve analytically. As a result, Markov Chain Monte Carlo (MCMC) is often used which draws samples from posterior distributions. The sampling from posterior distribution is performed using various techniques such as Metropolis-Hasting or Gibbs sampler [10][19][20]. In [19], Stochastic Approximation Monte Carlo (SAMC) is applied to multiple change point detection problem and SAMC performance is compared with reversible jump Markov Chain

Monte Carlo approach (RJMCMC). It was shown that in change point estimation, SAMC outperforms RJMCMC for complex Bayesian model selection problem.

On the other hand, Expectation Maximization (EM) can be viewed as an iterative approach to find the maximum likelihood [21] [22]. EM can also be employed to detect the changes. Some researchers have already used EM to detect change points [23] [24]. In [23], a Sequential Monte Carlo (SMC) online EM algorithm is proposed to estimate the change point. In [24] an EM method is presented to estimate the distribution of change point. These traditional EM methods require complex calculation while EM algorithm proposed in this work does not require heavy and complex computation and it is relatively easy to implement as well due to availability of explicit solutions. This framework has the advantage of handling improper selection of hyperparameters compared with Bayesian approach.

In this thesis, a closed form solution to the Bayesian formulation of single and multiple change point detection problem is first considered for multivariate data and MAP is used for the estimation of the parameters. Moreover, considering the sensitivity of the Bayesian approach to prior selection, EM is adopted to solve both single or multiple change points detection problem. By comparison, it is shown that EM is more powerful when priors are highly uncertain while the Bayesian approach has its advantage of less computation demand.

In the following, a quick review on Bayesian inference and Expectation Maximization methods are given to provide a better insight for the following chapters.

1.1 A Review on Model-Based Bayesian Inference

In the context of parameter estimation, Bayesian interpretation of probabilities is one of the most widely used approaches. It updates the knowledge about unknowns, parameters based on the information observed from the data. In other words, in Bayesian inference, one can specify one's belief in a statement or a given evidence in terms of probability [25]. The basis for most Bayesian approaches in various applications is Bayes theorem which is equal to

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

where A and B are two arbitrary events. Replacing B with observed data, y, A with the parameter set, θ, yields

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{1.2}$$

where $P(\theta)$ is the prior distribution of unknown parameters before y is observed, $P(y|\theta)$ is the likelihood of observation, y, under a model and $P(\theta|y)$ is the joint posterior distribution or full posterior distribution. It expresses the uncertainty about parameter, θ, after considering the information of both prior and likelihood. The denominator of (1.2), can be derived as

$$P(y) = \int P(y|\theta)P(\theta)d\theta = c \tag{1.3}$$

where c is a constant. It is called "marginal likelihood" of y or the "prior predictive distribution" of y and may be set as an unknown constant, i.e. normalizing constant. Thus, by removing c from denominator of (1.2), we can express the joint posterior distribution as

$$P(\theta|y) \propto P(y|\theta)P(\theta) \tag{1.4}$$

This form is called unnormalized joint posterior distribution which can be used for inference.

On the other hand, the parameter set, θ, consists of multiple parameters as $\theta = (\theta_1, ..., \theta_j)^T$, but not all the parameters in the vector θ are of interest. These uninterested parameters are called nuisance parameters. In other words, a nuisance parameter is one that is part of the θ in the joint posterior distribution of a model but it is not of main interest. One way to omit these nuisance parameters is to integrate out or marginalize the joint posterior distribution with respect to them. Define $\theta = (\phi, \omega)$ where ϕ represents the main parameters of interest and ω indicates the nuisance parameters. The marginal posterior of ϕ can be derived as

$$P(\phi|y) = \int P(\phi, \omega|y)d\omega \tag{1.5}$$

Thus, in Bayesian Inference, the marginal unnormalized joint posterior distribution is taken into account for estimation of unknown parameters. For instance, one may utilize MAP method to find the mode of marginal unnormalized joint posterior distribution to estimate the parameters. In Chapters 2 and 3, we will use this method in Bayesian inference.

1.2 Introduction to Expectation Maximization (EM)

EM algorithm is based on maximum likelihood estimation. This method was first introduced by [26] in 1977. Numerous applications in various areas can be found in literature based on this method such as in machnine learning, computer vision, medical imaging, mixture models, speech recognition, etc. It is effective especially when it is not easy to find the maximum of $P(parameter|data)$ directly [21] [22]. This algorithm consists of iteration between two steps: expectation-step or E-step and maximization-step or M-step. In other words, EM can be formulated as (1) finding the conditional expectation with respect to missing variables given the data and the current estimate of the parameter and (2) maximizing the expectation derived in the previous step to estimate the parameters [21]. Convergence of EM algorithm is guaranteed because at each iteration, the likelihood function is non-decreasing [22]. In this framework, E-step can be formulated as

$$Q(\theta|\theta^{(k)}) = E_{Z|D,\theta^{(k)}}\{P(D, Z|\theta)\} \tag{1.6}$$

where Z is the missing data or variable, D is the observed data and θ is the parameter to be estimated. $\theta^{(k)}$ is the current estimate of parameter. In essence, in E-step, missing data are estimated given the observation and the current estimate of unknown parameters. In other words, in E-step, a conditional expectation is derived, where depending on the types of missing variables, continues or discrete variables, an integration or summation is computed. In the M-step, the new parameter, $\theta^{(k+1)}$, is chosen so that it maximizes $Q(\theta|\theta^{(k)})$; that is,

$$Q(\theta^{(k+1)}|\theta^{(k)}) \geq Q(\theta|\theta^{(k)}), \forall\theta \tag{1.7}$$

M-step can be expressed as

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta|\theta^{(k)}) \tag{1.8}$$

Starting EM with initial values for the parameters, E-step and M-step are repeated until a suitable stopping rule criterion is satisfied. EM has interesting properties such as increasing the likelihood of observed data at each iteration but it may converge to a local maximum of observed likelihood which makes it dependent on starting values in some problems [22]. In Chapter 3, a review of existing methods for proper initialization of EM is elaborated. Some other properties such as selection of stopping criteria is important depending on the parameter estimation problem.

1.3 Derivation of EM

Let X be a random vector and θ be unknown sets of parameters and assume that we wish to find the maximum likelihood estimate of for θ. Define the log likelihood function as

$$L(X|\theta) = lnP(X|\theta) \tag{1.9}$$

Since $ln(x)$ is strictly increasing, the value of θ that maximizes $p(X|\theta)$ also increases $L(\theta)$. EM is proceeded iteratively to maximize $L(\theta)$. Assume that the current estimate is θ^n, the procedure must be in way that the updated parameter estimate of θ such that

$$L(\theta) > L(\theta^n) \tag{1.10}$$

In problems in which there are some missing variables, EM can make the maximum likelihood estimation tractable. Denote the missing variable as Z. The total probability $P(X|\theta)$ can be expressed as

$$P(X|\theta) = \sum_z P(X|z,\theta)P(z|\theta) \tag{1.11}$$

The difference between likelihood function at current estimate and updated estimate can be written as

$$L(\theta) - L(\theta^n) = ln(\sum_z P(X|z,\theta)P(z|\theta)) - lnP(X|\theta^n) \tag{1.12}$$

Note that in this expression there is the logarithm of sum. One can apply Jensen's inequality to (1.12)

$$ln\sum_{i=1}^{n}\lambda_i x_i \geq \sum_{i=1}^{n}\lambda_i ln(x_i) \tag{1.13}$$

Thus, we can introduce the constant $P(z|X,\theta^n)$ as

$$\begin{aligned}
L(\theta) - L(\theta^n) &= ln(\sum_z P(X|z,\theta)P(z|\theta)) - lnP(X|\theta^n) \\
&= ln(\sum_z P(X|z,\theta)p(z|\theta).\frac{P(z|X,\theta^n)}{P(z|X,\theta^n)}) - lnP(X|\theta^n) \\
&= ln(\sum_z P(z|X,\theta^n)\frac{P(X|z,\theta)P(z|\theta)}{P(z|X,\theta^n)}) - lnP(X|\theta^n) \\
&\geq \sum_z P(z|X,\theta^n)ln(\frac{P(X|z,\theta)P(z|\theta)}{P(z|X,\theta^n)}) - lnP(X|\theta^n) \qquad (1.14) \\
&= \sum_z P(z|X,\theta^n)ln(\frac{P(X|z,\theta)P(z|\theta)}{P(z|X,\theta^n)P(X|\theta^n)}) \qquad (1.15) \\
&= \Delta(\theta|\theta^n) \qquad (1.16)
\end{aligned}$$

In equations (1.14) to (1.16), the fact that $\sum_z P(z|X,\theta^n) = 1$ is used so that $lnP(X|\theta^n) = \sum_z P(z|X,\theta^n)lnP(X|\theta^n)$. We can write

$$L(\theta) \geq L(\theta^n) + \Delta(\theta|\theta^n) \tag{1.17}$$

Thus

$$L(\theta) \geq l(\theta|\theta^n) \tag{1.18}$$

where

$$l(\theta|\theta^n) = L(\theta^n) + \Delta(\theta|\theta^n) \tag{1.19}$$

The function $l(\theta|\theta^n)$ is bounded above by the likelihood function $L(\theta)$. Also, we have

$$\begin{aligned}
l(\theta|\theta^n) &= L(\theta^n) + \Delta(\theta^n|\theta^n) \\
&= L(\theta^n) + \sum_z P(z|X,\theta^n)ln\frac{P(X|z,\theta^n)P(z|\theta^n)}{P(z|X,\theta^n)P(X|\theta^n)} \\
&= L(\theta^n) + \sum_z P(z|X,\theta^n)ln\frac{P(X,z|\theta^n)}{P(z,X,\theta^n)} \\
&= L(\theta^n) + \sum_z P(z|X,\theta^n)ln1 \\
&= L(\theta^n) \qquad (1.20)
\end{aligned}$$

Thus for $\theta = \theta^n$, the functions $l(\theta|\theta^n)$ and $L(\theta)$ are equal.

The value of θ is chosen so that $L(\theta)$ is maximized. It is shown that the function $l(\theta|\theta^n)$ is bounded above by the likelihood function $L(\theta)$. Thus any value of θ that increases $l(\theta|\theta^n)$ also increases $L(\theta)$. EM algorithm finds θ such that $l(\theta|\theta^n)$ is maximised. This value is called updated parameter and is defined as θ^{n+1}

$$\begin{aligned}\theta^{n+1} &= argmax_\theta\{l(\theta|\theta^n)\}\\ &= argmax_\theta\{L(\theta^n) + \sum_z P(z|X,\theta^n) ln\frac{P(X|z,\theta)P(z|\theta)}{P(X|\theta^n)P(z|X,\theta^n)}\}\end{aligned} \tag{1.21}$$

If we drop the terms that do not depend on θ

$$\begin{aligned} &= argmax_\theta\{\sum_z P(z|X,\theta^n) lnP(X|z,\theta)P(z|\theta)\}\\ &= argmax_\theta\{\sum_z P(z|X,\theta^n) lnP(X,z|\theta)\}\\ &= argmax_\theta\{\sum_z P(z|X,\theta^n) lnP(X,z|\theta)\}\\ &= argmax_\theta\{E_{Z|X,\theta^n}\{lnP(X,z|\theta)\}\}\end{aligned} \tag{1.22}$$

In (1.21), the expectation and maximisation steps are simultaneous. To sum up, EM algorithm can be expressed as two steps:
E-step: Compute the conditional expectation $E_{Z|X,\theta^n}\{lnP(X,z|\theta)\}$
M-step: Maximize this expression with respect to θ.

1.4 Rate of Convergence of EM

The rate of convergence of EM is of interest in many applications. This was first elaborated in [26]. It was shown that the rate of convergence of EM is linear and depends on the proportion of information of the observations. In other words, if a large proportion of data is missing, then the convergence can be very slow. The convergence properties of EM algorithm have been investigated in detail in [22]. Basically, EM defines a mapping $\theta \to M(\theta)$ from the parameter space Θ to itself such that

$$\theta^{(k+1)} = M(\theta^{(k)}) \quad (k = 0, 1, 2, ...) \tag{1.23}$$

where the function M is referred to as EM mapping. This function is used for many convergence theorems of EM algorithm. If $\theta^{(k)}$ converges to some point θ^* and $M(\theta)$ is continuous, then the fixed point of EM is θ^* which satisfies $\theta^* = M(\theta^*)$. Using Taylor series expansion, around $\theta^{(k)} = M(\theta^*)$, we can write

$$\theta^{(k+1)} - \theta^* \approx J(\theta^*)(\theta^{(k)} - \theta^*) \tag{1.24}$$

where $J(\theta)$ is $d \times d$ Jacobian matrix of $M(\theta) = (M_1(\theta), M_2(\theta), ..., M_d(\theta))^T$. These functions are called mapping functions associated with each parameter. EM algorithm is a linear iteration with rate matrix as $J(\theta^*)$ [22]. This matrix is called the matrix rate of convergence. For the parameters, θ, the global rate of convergence is defined as

$$r = \lim_{k\to\infty} \frac{\|\theta^{(k+1)} - \theta^*\|}{\|\theta^{(k)} - \theta^*\|} \tag{1.25}$$

where $\|.\|$ is any defined norm on d-dimensional Euclidean space. The larger the global rate of convergence, r, the slower the convergence. In addition, the global speed of convergence is defined as $s = 1 - r$ [27]. It can be shown that at each iteration of EM, the likelihood of observed data increases until a fixed point is reached. The stationary or fixed points of EM are Maximum Likelihood Estimates (MLE). Furthermore, the EM sequence is convergent if

$$\frac{\partial Q(\theta|\theta^{(k)})}{\partial\theta}|_{\theta=\theta^{k+1}} = 0 \tag{1.26}$$

and also the series $\{\theta^{(k)}\}$ converges to θ^* and $logP(z|y,\theta)$ is sufficiently smooth [28] [29]. It should be noted that there is no guarantee that EM converges to global maximum. If the likelihood function has multiple maxima, then it may converge to local maximum which makes it dependent on initial values, θ^0. There are several methods for proper initialization of EM algorithm which will be discussed in Chapter 3. In the presence of constrained parameter space, $\Theta \subset \Omega$, [30] provides a stricter condition for convergence. This condition assumes intersection of boundary of parameter Θ with the complement of boundary of Ω is a subset of θ ($\partial\Theta \backslash \partial\Omega \subseteq \Theta$).

1.5 Stopping Criteria

There are several criteria defined in order to terminate the EM algorithm [22]. They can be categorized in three groups:

- One criterion is based on absolute value of the likelihood change in two successive iterations. One can employ the relative change of likelihood since the values of likelihood depend on the sample size and therefore an absolute change in the order of 10^{-s} has different importance depending on the sample size. If $L(\theta^{(k)})$ is the value of the observed likelihood in kth iteration, and tol is a small number of the form 10^{-s}, then we can write

  $$\frac{|L(\theta^{(k+1)}) - L(\theta^{(k)})|}{|L(\theta^{(k)})|} \prec tol \tag{1.27}$$

  This criterion does not indicate the actual convergence [31].

- One criterion is based on the relative change of the parameters in two successive iterations. The maximum over all the parameters is considered as a measure of progress. This can be written as

  $$max_j \frac{|\theta_j^{(k+1)} - \theta_j^{(k)}|}{|\theta_j^{(k)}|} \prec tol \tag{1.28}$$

where j is the jth element in the vector of the parameters. Similarly, this criterion does not represent the actual convergence [31]. In [32], it was emphasized that the aforementioned methods are measures of lack of progress but not actual convergence. Thus, a new criterion is defined by [32] to solve the convergence problem as explained next.

- (Aitken Acceleration Criterion): In applications where the interest is the likelihood values rather than the sequence of estimated parameters, a new criterion is defined based on a projected likelihood [32]. Define

  $$L^k = L(\theta^{(k)}) \tag{1.29}$$

  The Aitken criterion, is expressed as

  $$|L_A^{k+1} - L_A^k| \prec tol \tag{1.30}$$

  where the projected likelihood is expressed as

  $$L_A^{k+1} = L^k + \frac{1}{(1-c^{(k)})}(L^{k+1} - L^k) \tag{1.31}$$

  where $c^{(k)} = \frac{(L^{k+1}-L^k)}{(L^k-L^{k-1})}$. According to [22], this criterion can be used for any log-likelihood sequence which is linearly convergent.

A great number of examples can be found in literture. [22] is a comprehensive review of EM theory and various applications of EM can be found there.

1.6 Problem Statement

As mentioned, there are many applications for change point problem in areas such as hydrology, DNA sequences, financial time series, signal processing, etc. Given a time series, change points split the data into segments which are disjoint. It is assumed that the data in every segments are independent of each other. In addition, the independence assumption holds for data within a segment. In every segment, the data can be fitted to a model. Thus, we can write

$$data = model + noise \tag{1.32}$$

In a time series, various scenarios may exist [33]:

- Different segments may have different models in terms of model types and model orders
- Different segments may have the same model structure with different parameters
- Different segments may have the same model structure with different noise levels

- Multivariate segments with different correlation coefficients

In this work, it is assumed that the data in every segment follow the same model structure but with different parameters. Four possible scenarios can be considered.

- The model in which no change point exists in the data. See Figure 1.1.a
- The model in which mean shifts at multiple points but variability is constant. See Figure 1.1.b
- The model in which mean is constant but variability changes at multiple points. See Figure 1.1.c
- The model in which mean and variability change simultaneously at multiple points. See Figure 1.1.d

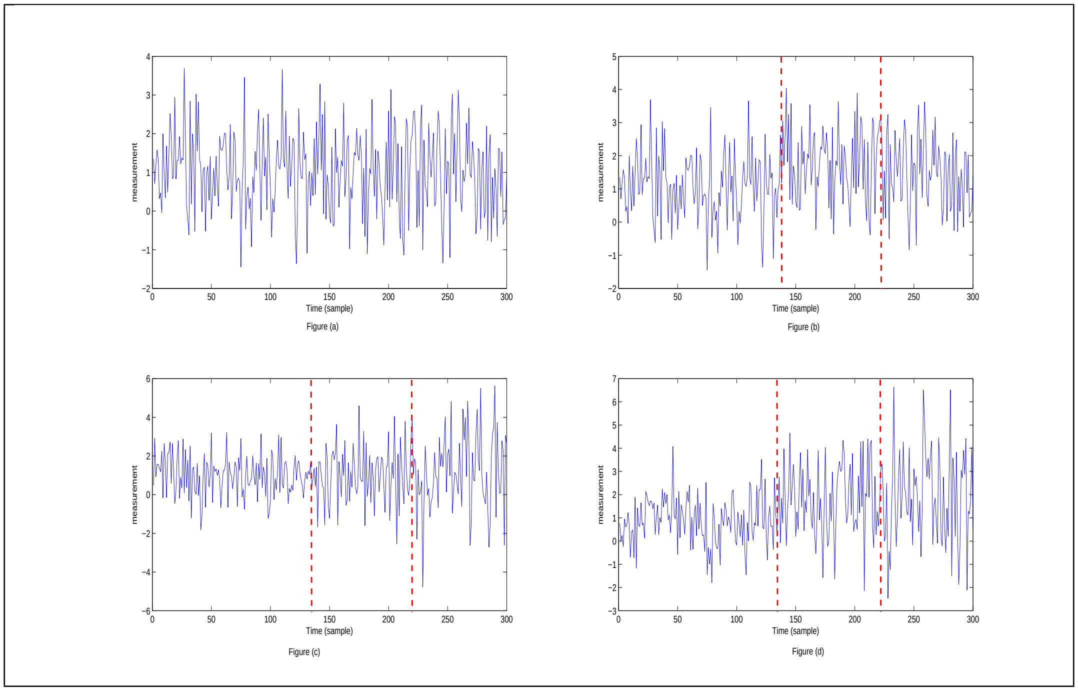

Figure (a)

Figure (b)

Figure (c)

Figure (d)

Figure 1.1: Figure (a): No change in Mean or Variability, Figure (b): Change in Mean Only, Figure (c): Change in Variability Only, Figure (d): Change in Both Mean and Variability

These scenarios are investigated in this work. It is assumed that the time instants where these changes occur are unknown. Thus, the main focus is derivation of a closed form solution for change point detection under various scenarios using EM and Bayesian methods.

1.7 Thesis Overview

1.7.1 Thesis Contribution

Various approaches have been reviewed for gross error detection, mean shift and change point detection. In terms of probabilistic frameworks, as mentioned earlier, there have been various techniques developed. Most of probabilistic approaches rely on Bayesian methods. Expectation Maximization method for change point detection was employed in [23] [24] but those methods require complex derivations. This thesis derives closed form solutions for the Q-functions which are critical functions in the EM algorithm and hence making the maximization step rather straightforward. The main contributions of this thesis are:

- Derivation of closed form solution using Bayesian and Expectation Maximization (EM) and Simplified EM (SEM) methods in mean shift detection problem in univariate data in the presence of known variance.

- Derivation of closed form solution using Bayesian and Expectation Maximization methods in mean shift detection problem in multivariate data in the presence of known covariance.

- Derivation of closed form solution using Expectation Maximization in change point detection problem in multivariate data in the presence of unknown and changing covariance.

- Mean estimation in change point model in the presence of process constraints using constrained EM method.

1.7.2 Thesis Outline

In this work, mean shift or change point detection is investigated under various scenarios. This thesis is organized as follows. In Chapter 2, the problem is solved using Bayesian inference as well as EM algorithm for univariate data with single change assuming known and constant variance. A simplified version of EM is also proposed to deal with this problem. The performance of these three methods is compared through simulated data.

In Chapter 3, the mean shift detection is solved for multivariate data using Bayesian and EM approaches in the presence of constant and known covariance. Bayesian framework is presented first for mean shift detection by generalizing the method employed in [7]. This formulation is different from [7] in the sense of selection of hyperparameters of priors. In addition, two novel EM algorithms are proposed for mean shift detection. Performance of the proposed algorithm is evaluated using simulated and pilot-scale experiment data.

In Chapter 4, the change point detection is solved using EM approaches in the presence of unknown covariance and then this solution is extended to the change point detection solution in the case of unknown and simultaneous changing mean and covariance. Using simulated and pilot-scale experiment case studies, the performance of the proposed methods are evaluated.

Chapter 5 covers the mean estimation in multiple change points model through EM algorithm. In addition, mean estimation solution is derived in the presence of process constraints using constrained EM method. The fast convergence of EM demonstrates the efficiency of the method in parameter estimation.

Finally, in Chapter 6, the thesis conclusions and future works are presented.

Chapter 2

Univariate Mean Shift Detection

2.1 Introduction

In this chapter, the univariate change point detection problem in the presence of known variance is investigated and three methods based on Bayesian, Expectation Maximization (EM) and Simplified Expectation Maximization (SEM) are derived and compared through simulation data.

2.2 Measurement Model

In this model, it is assumed that the process is operating under steady state condition. For each variable, say the ith variable, at time point$j$, the model can therefore be written as

$$y_{ij} = \mu_{1i} + \epsilon_{ij} \tag{2.1}$$

where μ_{1i} is the true value of variable i, ϵ_{ij} is the measurement noise of variable i at time j. Moreover, the measurement noise at different time instants is assumed to be independent. It is also assumed that for variables $i \neq k$, ϵ_i and ϵ_k are independent. For all j, ϵ_{ij}'s are assumed to follow a normal distribution with mean zero and constant and known variance, i.e.

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_{0i}^2) \tag{2.2}$$

Suppose that at time $j = m$, bias, δ_{ij}, occurs in variable i resulting in a change of mean value of process variable i as μ_{2i}. Assuming that the variance, σ_{0i}^2, remains constant, we have

$$\begin{aligned} y_{ij} &= \mu_{2i} + \epsilon_{ij}, \quad j = m+1, m+2, ..., n \\ \mu_{2i} &= \mu_{1i} + \delta_{ij} \end{aligned} \tag{2.3}$$

Thus, the measurements y_{ij} before and after bias shift can be written as

$$\begin{aligned} y_{ij} &\sim \mathcal{N}(\mu_{1i}, \sigma_{0i}^2) \;\; j = 1, 2, ...m \\ y_{ij} &\sim \mathcal{N}(\mu_{2i}, \sigma_{0i}^2) \;\; j = m+1, m+2, ...n \end{aligned} \tag{2.4}$$

Based on the model given in (2.1) with assumptions of (2.2), (2.3) and (2.4), the proposed methods can be applied.

2.3 Bayesian Method

As mentioned earlier, Bayesian inference is one of the most widely used method in parameter estimation. Under this framework, one first assigns prior probability distribution to unknown parameters. The unknown parameters in this problem are μ_{1i}, μ_{2i} and m. Consider a univariate process so that the subscript i can be dropped. Thus, $\mu_{1i} = \mu_1$, $\mu_{2i} = \mu_2$ and $\sigma_{0i}^2 = \sigma_0^2$. The priors are selected as

$$\begin{aligned} P(\mu_1) &= \mathcal{N}(\mu_1^0, \sigma_{10}^2) \\ P(\mu_2) &= \mathcal{N}(\mu_2^0, \sigma_{20}^2) \\ P(m) &= k_m (\text{Uniform Distribution}) \end{aligned} \tag{2.5}$$

where k_m is a constant, μ_1^0 and σ_{10}^2 are the hyperparameters of μ_1, and μ_2^0 and σ_{20}^2 are the hyperparameters of μ_2 and assumption of uniform distribution of m follows from [7]. For simplicity denote $\beta_0 = (\mu_1^0, \sigma_{10}^2, \mu_2^0, \sigma_{20}^2)$. Unlike [7], we assume a nonzero values for μ_1^0 and μ_2^0. The reason is that assigning a zero mean value to prior distribution of mean is equivalent to prior belief that the most probable value for mean of data is zero which is not true. These priors are assumed to be mutually independent. Independence of m and the mean values means that the time instant of change is independent of mean values before and after the change point. Assuming independent observations at different time instants, the likelihood function for n samples of data, $Y = (y_1, y_2, ..., y_n)$, is

$$P(Y|\mu_1, \mu_2, m, \sigma_0^2, \beta_0) = \prod_{n=1}^{m} \mathcal{N}(y_i|\mu_1, \sigma_0^2, \beta_0) \times \prod_{n=m+1}^{n} \mathcal{N}(y_i|\mu_2, \sigma_0^2, \beta_0) \tag{2.6}$$

Having obtained the likelihood and priors, the next step is to find the joint posterior probability of μ_1, μ_2 and m. According to Bayesian theorem:

$$P(\mu_1, \mu_2, m|Y, \sigma_0^2, \beta_0) \propto P(Y, \mu_1, \mu_2, m, \sigma_0^2, \beta_0) \tag{2.7}$$

Using chain rule, the joint probability distribution can be written as

$$\begin{aligned} P(Y, \mu_1, \mu_2, &\sigma_0^2, \beta_0, m) \\ &= P(Y|\mu_1, \mu_2, \sigma_0^2, \beta_0, m) P(\mu_1, \mu_2, \sigma_0^2, \beta_0, m) \\ &= P(Y|\mu_1, \mu_2, \sigma_0^2, \beta_0, m) P(\mu_1|\mu_2, \sigma_0^2, \beta_0, m) \times P(\mu_2, m, \sigma_0^2, \beta_0) \\ &= P(Y|\mu_1, \mu_2, \sigma_0^2, \beta_0, m) P(\mu_1|\mu_2, \sigma_0^2, \beta_0, m) \times P(\mu_2|m, \sigma_0^2, \beta_0) P(m) \\ &= P(Y|\mu_1, \mu_2, \sigma_0^2, \beta_0, m) P(\mu_1|\beta_0) P(\mu_2|\beta_0) P(m) \end{aligned} \tag{2.8}$$

The last equation in (2.8) signifies the independent assumptions made for priors. Therefore, the posterior probability in (2.7) is of the form

$$P(\mu_1, \mu_2, m|Y, \sigma_0^2, \beta_0) \propto P(Y|\mu_1, \mu_2, \sigma_{0i}^2, \beta_0, m) \times P(\mu_1|\beta_0)P(\mu_2|\beta_0)P(m) \tag{2.9}$$

Substituting (2.5) and (2.6) in (2.9), we have

$$\begin{aligned} P(\mu_1, \mu_2, &m|Y, \sigma_0^2, \beta_0) \propto exp(-\sum_{j=1}^{m}[\frac{(y_j-\mu_1)^2}{2\sigma_0^2}] - \frac{(\mu_1-\mu_1^0)^2}{2\sigma_{10}^2})\times \\ &exp(-\sum_{j=m+1}^{n}[\frac{(y_j-\mu_2)^2}{2\sigma_0^2}] - \frac{(\mu_2-\mu_2^0)^2}{2\sigma_{20}^2}) \end{aligned} \tag{2.10}$$

After some algebraic manipulations such as completing the square terms with respect to μ_1 and μ_2, the following result is derived:

$$P(\mu_1, \mu_2, m|Y, \sigma_0^2, \beta_0) \propto F_1 \times F_2 \tag{2.11}$$

where

$$\begin{aligned} F_1 = &exp\{\frac{-1}{2\sigma_0^2(mb)^{-1}}[\mu_1 - (mb)^{-1}(mh + \sum_{j=1}^{m} y_j)]^2\}\times \\ &exp\{\frac{-1}{2\sigma_0^2}[mh' + \sum_{j=1}^{m} y_j^2 - (mb)^{-1}[mh + \sum_{j=1}^{m} y_j]^2]\} \end{aligned} \tag{2.12}$$

and

$$\begin{aligned} F_2 = exp\{\frac{-1}{2\sigma_0^2(c)^{-1}}[\mu_2 - \frac{\sum_{j=m+1}^{n}(y_j) + (n-m)p}{c}]^2\}exp\{\frac{-1}{2\sigma_0^2}[\sum_{j=m+1}^{n}(y_j^2) + (n-m)p' - \\ c^{-1}(\sum_{j=m+1}^{n}(y_j) + (n-m)p)^2]\} \end{aligned} \tag{2.13}$$

Using notations similar to [7], we have

$$\begin{aligned} b &= 1 + \frac{\sigma_0^2}{m\sigma_{10}^2} \\ h &= \frac{\mu_1^0\sigma_0^2}{m\sigma_{10}^2} \\ h' &= \mu_1^0.h \end{aligned} \tag{2.14}$$

and the parameters of F_2 are

$$\begin{aligned} c &= n - m + \frac{\sigma_0^2}{\sigma_{20}^2} \\ p &= \frac{\mu_2^0\sigma_0^2}{(n-m)\sigma_{20}^2} \\ p' &= \mu_2^0.p \end{aligned} \tag{2.15}$$

In order to derive the marginal posterior $P(m|Y, \sigma_0^2, \beta_0)$ based on joint posterior probability, we can integrate out μ_1 and μ_2 from the joint posterior $P(\mu_1, \mu_2, m|Y, \sigma_0^2, \beta_0)$. Here, since F_1 only depends on μ_1 and F_2 only depends on μ_2, integrating F_1 with respect to μ_1 results in I_1 and integrating of F_2 with respect to μ_2 leads to I_2 as

$$I_1 = exp\{\frac{-1}{2\sigma_0^2}[mh' + \sum_{j=1}^{m} y_j^2 - (mb)^{-1}[mh + \sum_{j=1}^{m} y_j]^2]\}\sqrt{2\pi(mb)^{-1}} \times \sigma_0 \tag{2.16}$$

and

$$I_2 = exp\{\frac{-1}{2\sigma_0^2}[\sum_{j=m+1}^{n}(y_j^2) + (n-m)p' - c^{-1}(\sum_{j=m+1}^{n}(y_j) + (n-m)p)^2]\}\sqrt{2\pi c^{-1}} \times \sigma_0 \tag{2.17}$$

The marginal posterior is derived accordingly as

$$P(m|Y, \sigma_0^2, \beta_0) \propto I_1 \times I_2 \tag{2.18}$$

In order to find the most likely solution of m from the marginal posterior, one can maximize the marginal posterior with respect to m as

$$\hat{m} = \arg\max_m P(m|Y, \sigma_0^2, \beta_0) \tag{2.19}$$

Finally, according to [7], if m is found to be $1, 2, ..., n-1$, then there is gross error and if $m = n$, there is no gross error found.

In [7], the mean values of the priors are set as zero for both μ_1 and μ_2. As discussed earlier, assigning zero as mean values of the priors can lead to false detection of change point. The reason is that the joint distribution depends on μ_1^0 and μ_2^0, as evident in I_1 and I_2 of (2.16) and (2.17), and poor selection of the priors will lead to deteriorated performance and could result in false detection. In next section, we propose EM algorithm to overcome this deficiency and to take advantage of use of improved priors based on the data.

2.3.1 Mean Shift Formulation Using EM

In this section, EM algorithm is proposed to solve single change point detection problem. Here, μ_1 and μ_2 are treated as the missing or hidden variables, $Y = (Y_1, Y_2, ..., Y_n)$ are the observed data and m, the change point, is the parameter of the interest to be determined. Thus, E-step can be expressed as

$$Q(m|m^{(k)}) = E_{\mu_1,\mu_2|Y,m^{(k)}}\{p(Y, \mu_1, \mu_2|m)\} \tag{2.20}$$

In M-step, the maximization is performed with respect to all possible values of the parameter, m, as

$$m^{(k+1)} = \arg\max_m Q(m|m^{(k)}) \tag{2.21}$$

In order to derive the E-step, we first write

$$\begin{aligned} P(Y, \mu_1, \mu_2|m) &= P(Y|\mu_1, \mu_2, m)P(\mu_1, \mu_2|m) \\ &= P(Y|\mu_1, \mu_2, m)P(\mu_1|\mu_2, m)P(\mu_2|m) \end{aligned} \tag{2.22}$$

Since m, the shift point, is independent of the mean values μ_1 and μ_2 and also these values are independent of each other, (2.22) can be written as

$$P(Y, \mu_1, \mu_2|m) = P(Y|\mu_1, \mu_2, m)P(\mu_1)P(\mu_2) \tag{2.23}$$

Thus, (2.20) yields

$$\begin{aligned} Q(m|m^{(k)}) &= E_{\mu_1,\mu_2|Y,m^{(k)}}\{P(Y|\mu_1, \mu_2, m)P(\mu_1)P(\mu_2)\} \\ &= \int\int P(Y|\mu_1, \mu_2, m)P(\mu_1)P(\mu_2)P(\mu_1, \mu_2|Y, m^{(k)})d\mu_1 d\mu_2 \end{aligned} \tag{2.24}$$

Since μ_1 and μ_2 are independent, we can write

$$Q(m|m^{(k)}) = \int\int P(Y|\mu_1, \mu_2, m)P(\mu_1)P(\mu_2)P(\mu_1|Y, m^{(k)})P(\mu_2|Y, m^{(k)})d\mu_1 d\mu_2 \tag{2.25}$$

In the integrand, there are two terms $P(\mu_1|Y, m^{(k)})$ and $P(\mu_2|Y, m^{(k)})$ which can be determined as

$$P(\mu_1|Y, m^{(k)}) = \frac{P(Y|\mu_1, m^{(k)})P(\mu_1|m^{(k)})}{p(Y|m^{(k)})} \tag{2.26}$$

$$= \frac{P(y_1, y_2, ..., y_{m^{(k)}}|\mu_1, m^{(k)})P(y_{m^{(k)}+1}, y_{m^{(k)}+2}, ..., y_n)P(\mu_1)}{P(Y|m^{(k)})} \tag{2.27}$$

$$= k_1 P(y_1, y_2, ..., y_{m^{(k)}}|\mu_1, m^{(k)})P(\mu_1) \tag{2.28}$$

In the nominator of (2.26), $P(\mu_1|m^{(k)}) = P(\mu_1)$ because μ_1 and $m^{(k)}$ are independent. In the denominator of (2.26), $P(Y|m^{(k)})$ is the distribution of Y that only depends on μ_1, μ_2 and the variance. These parameters are independent of $m^{(k)}$; therefore this probability is a constant value as

$$\frac{P(y_{m^{(k)}+1}, y_{m^{(k)}+2}, ..., y_n)}{P(Y|m^{(k)})} = k_1 \tag{2.29}$$

Using the same approach for $P(\mu_2|Y, m)$, we have

$$P(\mu_2|Y, m^{(k)}) = \frac{P(Y|\mu_2, m^{(k)})P(\mu_2|m^{(k)})}{p(Y|m^{(k)})} \tag{2.30}$$

$$= \frac{P(y_{m^{(k)}+1}, y_{m^{(k)}+2}, ..., y_n|\mu_2, m^{(k)})P(y_1, y_2, ..., y_{m^{(k)}})P(\mu_2)}{P(Y|m^{(k)})} \tag{2.31}$$

$$= k_2 P(y_{m^{(k)}+1}, y_{m^{(k)}+2}, ..., y_n|\mu_2, m^{(k)})P(\mu_2) \tag{2.32}$$

where

$$\frac{P(y_1, y_2, ..., y_{m^{(k)}})}{P(Y|m^{(k)})} = k_2 \tag{2.33}$$

Thus (2.25) can be rewritten as

$$\begin{aligned}
&Q(m|m^{(k)}) = k_1k_2 \int\int P(Y|\mu_1,\mu_2,m)P(\mu_1)P(\mu_2)\times \\
&P(y_{1:m^{(k)}}|\mu_1,m^{(k)})P(\mu_1)P(y_{m^{(k)}+1:n}|\mu_2,m^{(k)})P(\mu_2)d\mu_1 d\mu_2 \\
&= k_3 \int P(y_{1:m}|\mu_1,m)P(y_{1:m^{(k)}}|\mu_1,m^{(k)})(P(\mu_1))^2 d\mu_1\times \\
&\int P(y_{m+1:n}|\mu_2,m)P(y_{m^{(k)}+1:n}|\mu_2,m^{(k)})(P(\mu_2))^2 d\mu_2 \qquad (2.34)
\end{aligned}$$

where $k_3 = k_1k_2$. This integration is the product of two separable terms so that (2.34) can be simplified as

$$Q(m|m^{(k)}) = k_3 \int E_1 d\mu_1 \int E_2 d\mu_2 \qquad (2.35)$$

where E_1 can be written as

$$E_1 = exp\{\frac{-1}{2\sigma_0^2}[\sum_{j=1}^{m}(y_j-\mu_1)^2 + \sum_{j=1}^{m^{(k)}}(y_j-\mu_1)^2] - \frac{1}{\sigma_{10}^2}(\mu_1-\mu_1^0)^2\} \qquad (2.36)$$

After some algebraic simplifications to complete the square terms in E_1, we have

$$E_1 = exp\{\frac{-1}{2\sigma_0^2}[\sum_{j=1}^{m} y_j^2 + \sum_{j=1}^{m^{(k)}} y_j^2] - \frac{1}{\sigma_{10}^2}(\mu_1^0)^2 + \frac{\beta^2}{2\alpha}\}exp\{-\frac{(\mu_1-\frac{\beta}{\alpha})^2}{2(\frac{1}{\alpha})}\} \qquad (2.37)$$

where

$$\begin{aligned}
\alpha &= \frac{m}{\sigma_0^2} + \frac{m^{(k)}}{\sigma_0^2} + \frac{2}{\sigma_{10}^2} \\
\beta &= -\frac{\sum_{j=1}^{m} y_j + \sum_{j=1}^{m^{(k)}} y_j}{\sigma_0^2} - \frac{2\mu_1^0}{\sigma_{10}^2} \qquad (2.38)
\end{aligned}$$

Integration of E_1 gives

$$\int E_1 d\mu_1 = exp\{\frac{-1}{2\sigma_0^2}(\sum_{j=1}^{m} y_j^2 + \sum_{j=1}^{m^{(k)}} y_j^2) - \frac{1}{\sigma_{10}^2}(\mu_1^0)^2 + \frac{\beta^2}{2\alpha}\}\sqrt{2\pi\alpha^{-1}} \qquad (2.39)$$

Following similar algebraic manipulation for E_2 leads to

$$\int E_2 d\mu_2 = exp\{\frac{-1}{2\sigma_0^2}(\sum_{j=m+1}^{n} y_j^2 + \sum_{j=m^{(k)}+1}^{n} y_j^2) - \frac{1}{\sigma_{20}^2}(\mu_2^0)^2 + \frac{\gamma^2}{2\zeta}\} \times \sqrt{2\pi\zeta^{-1}} \qquad (2.40)$$

where

$$\begin{aligned}
\zeta &= \frac{n-m}{\sigma_0^2} + \frac{n-m^{(k)}}{\sigma_0^2} + \frac{2}{\sigma_{20}^2} \\
\gamma &= -\frac{\sum_{j=m+1}^{n} y_j + \sum_{j=m^{(k)}+1}^{n} y_j}{\sigma_0^2} - \frac{2\mu_2^0}{\sigma_{20}^2} \qquad (2.41)
\end{aligned}$$

Finally, $Q(m|m^{(k)})$ is derived by multiplying (2.39) by (2.40).

From (2.39) and (2.40), one can see that Q-function depends on current available estimate of the parameter $m^{(k)}$ and the decision variable m. In the next step, M-step, Q-function is maximized with respect to unknown m. In other words, starting EM loop with an initial value for $m(0) = m_0$, E-step and M-step iterate until no increase is observed in Q-function.

2.4 Simplified Expectation Maximization (EM)

In this section, a simplified EM algorithm is proposed for detecting mean shift. In this framework, the integration is avoided by replacing the hidden variables with their expected values given the data and current estimate of the parameters in the likelihood function. In E-step, we have

$$Q(m|m^{(k)}) = E_{\mu_1,\mu_2|Y,m^{(k)}}\{log[p(Y,\mu_1,\mu_2|m)]\} \tag{2.42}$$

Using (2.20), (2.42) can be written as

$$Q(m|m^{(k)}) = E_{\mu_1,\mu_2|Y,m^{(k)}}\{log(P(Y|\mu_1,\mu_2,m)) + log(P(\mu_1)) + log(P(\mu_2))\} \tag{2.43}$$

In simplified version of EM, instead of conditional expectation, μ_1 and μ_2 are replaced by their expected values estimated from the data. Define Q-function as

$$Q(m|m^{(k)}) = log(P(Y|E(\mu_1),E(\mu_2),m)) + log(P(E(\mu_1)) + log(P(E(\mu_2))) \tag{2.44}$$

where

$$\begin{aligned} E(\mu_1) &= \bar{y}_{1:m^{(k)}} = \frac{\sum_{j=1}^{m^{(k)}} y_j}{m^{(k)}} \\ E(\mu_2) &= \bar{y}_{m^{(k)}+1:n} = \frac{\sum_{j=m^{(k)}+1}^{n} y_j}{n-m^{(k)}} \end{aligned} \tag{2.45}$$

Therefore, using the likelihood and priors as in (2.5) and (2.6), Q-function in (2.42) can be written as

$$\begin{aligned} Q(m|m^{(k)}) = & \frac{-1}{2\sigma_0^2}[\sum_{j=1}^{m}(y_j - E(\mu_1))^2 + \sum_{j=m+1}^{n}(y_j - E(\mu_2))^2] \\ & - \frac{1}{2\sigma_{10}^2}(E(\mu_1) - \mu_1^0)^2 - \frac{1}{2\sigma_{20}^2}(E(\mu_2) - \mu_2^0)^2 \end{aligned} \tag{2.46}$$

For simplicity, the priors are selected as $\sigma_{10} = \sigma_{20}$. Here, selection of wrong values for μ_1^0 and μ_2^0 does not have any effects on performance of the algorithm since the hidden variables, μ_1 and μ_2, are estimated from data given the current estimate $m^{(k)}$. Having derived Q-function, in the next step, maximization is conducted with respect to m and the parameter is estimated. In the next iteration, this new parameter is used to re-calculate the mean values μ_1 and μ_2 embedded in Q-function. Hence, these two steps iterate until convergence.

2.5 Simulation Results and Discussions

In literature, there are two commonly used performance measures of gross error detection. Here, we apply those measures in mean shift detection. Thus, through simulation studies, the performance of Bayesian and Expectation Maximization approaches is investigated. One of those measures, defined as an indication of method's ability in detection of biased instrument, is overall power, OP

$$OP = \frac{\text{Number of Gross Errors Correctly Identified}}{\text{Number of Gross Errors Simulated}} \tag{2.47}$$

The other performance measure is defined to represent the probability of false alarm as

$$AVTI = \frac{\text{Number of Gross Errors Wrongly Identified}}{\text{Number of Simulation Trials Made}} \tag{2.48}$$

These two performance measures are widely used in gross error detection literatures [1].

In Figure 2.1, the process measurement for the case of mean shift with constant variance is shown. The mean of variable is subject to change at an unknown time instant. In this work, we focus on change in the mean of data at a single time point but no change in the variance due to steady state nature of the process operation. We also assume that after occurrence of gross error, the bias magnitude remains constant.

The simulated data are generated from a normal distribution with mean μ and standard deviation σ_0 in MATLAB. The results are obtained from Monte Carlo simulation of 1000 runs. In Table 3.1, the results of OP and $AVTI$ are shown for three methods: Bayesian (BM), EM and simplified EM (SEM). The simulation parameters are selected as

$$\begin{aligned}&n = 500, m = 142, \mu = 1, \sigma_0^2 = \sigma_{10}^2 = \sigma_{20}^2 = 1,\\ &\delta = 3\sigma_0, \mu_1^0 = 1, \mu_2^0 = 2.\end{aligned}$$

The EM and SEM algorithms start from an initial guess $m^{(0)} = 100$ while the real m is 142. As illustrated in Table 3.1, the results of OP and $AVTI$ indicate that EM is more accurate in detection of gross errors but the performance difference between BM and EM is minor for indicated bias magnitudes. In other words, when the mean values of prior distributions are selected to be close to true mean, both methods have approximately similar performance and high OP is achieved in detection of biased instruments. The performance of EM and SEM is almost the same for different magnitudes of biases. In all EM or SEM simulations, after one or two iterations, the algorithms converge.

Next, the results are shown when the mean of priors are improperly selected as shown in Table 2.2, i.e. far from the true mean. For 1000 simulation runs with $\delta = 3\sigma_0$, the following performance is obtained: OP=0 by BM, 0.8 by EM and 0.9 by SEM, indicating that EM and SEM are capable of correctly detecting biased instruments in the case of wrong selection of priors. This nature of learning from the data is an advantage of EM or SEM which makes them efficient approaches in gross error detection.

In Table 2.3, as an example of $n = 60$, the effects of different bias shift points are investigated. There is no significant difference in the values of OP and $AVTI$ for BM and EM, because these methods do not rely on point estimation of mean from data. However, for the case of SEM that relies on point estimation of mean, when m or $n - m$ decreases, OP decreases.

Figure 2.2 shows the overall power, OP, with respect to the size of bias as a multiplier of standard error. With small bias that is more difficult to detect, the power of EM is larger than BM and SEM. However, with gross errors between $1.5\sigma_0$ to $2.5\sigma_0$, the power of BM is the largest. As bias magnitude further increases, the OP's for three methods approach each other.

Table 2.1: Performance Comparison for Different biases ($n = 500$, $m = 142$, $\mu_1^0 = \mu_2^0 = 1$)

	Bias	*OP*	*AVTI*
BM	$3\sigma_0$	0.85	0.145
	$4\sigma_0$	0.95	0.044
	$5\sigma_0$	0.98	0.015
EM	$3\sigma_0$	0.86	0.140
	$4\sigma_0$	0.95	0.044
	$5\sigma_0$	0.99	0.009
SEM	$3\sigma_0$	0.86	0.141
	$4\sigma_0$	0.95	0.042
	$5\sigma_0$	0.99	0.011

Table 2.2: Performance Comparison for Different Mean values for Prior Distributions($\mu_1^0 = 20, \mu_2^0 = 30, n = 500, \delta = 3\sigma$)

	BM	*EM*	*SEM*
OP	0	0.8	0.9
AVTI	1	0.16	0.1

In the case of improper selection of $\mu_1^0 = 20$ and $\mu_2^0 = 30$, OP's for BM and EM are illustrated in Figure 2.3 for different values of bias. For bias up to $4\sigma_0$, OP for BM is zero and as bias magnitude increases, the power increases dramatically. The power for EM, even for small size of gross errors is nonzero, demonstrating the efficiency of EM despite of wrong priors. In Figure 2.4, the measurements along with three detection criteria using BM, EM and SEM are illustrated for $n = 300, m = 135, \delta = 3\sigma_0, \mu_1^0 = 1$ and $\mu_2^0 = 2$. For all three methods, y axis is in logarithmic scale. The peak points of posterior probability as well as that of Q-function are located where the bias shift occurs. In other words, the global maximum is obtained at true change point. As noted earlier, the convergence of EM

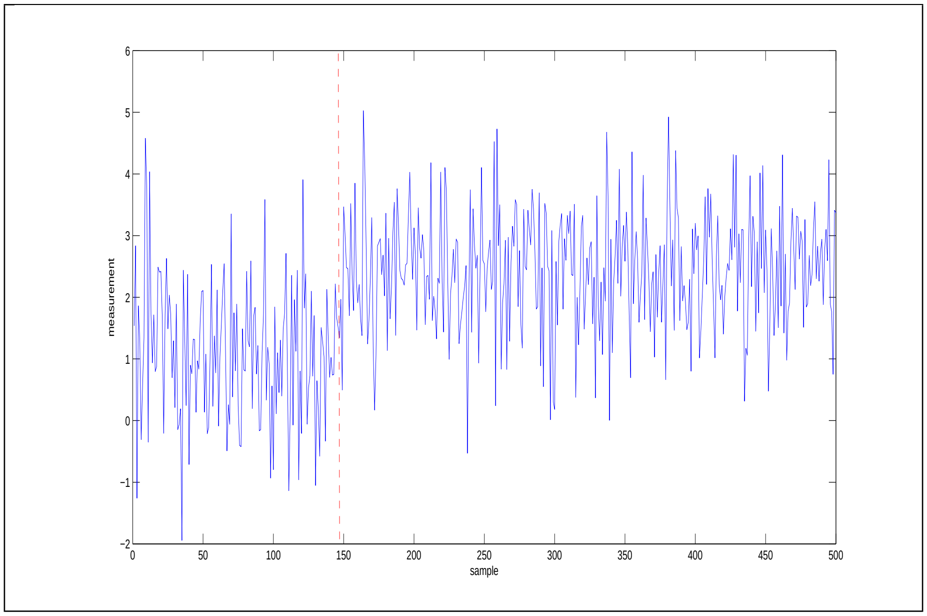

Figure 2.1: Measurement with Mean Shift and Constant Variance

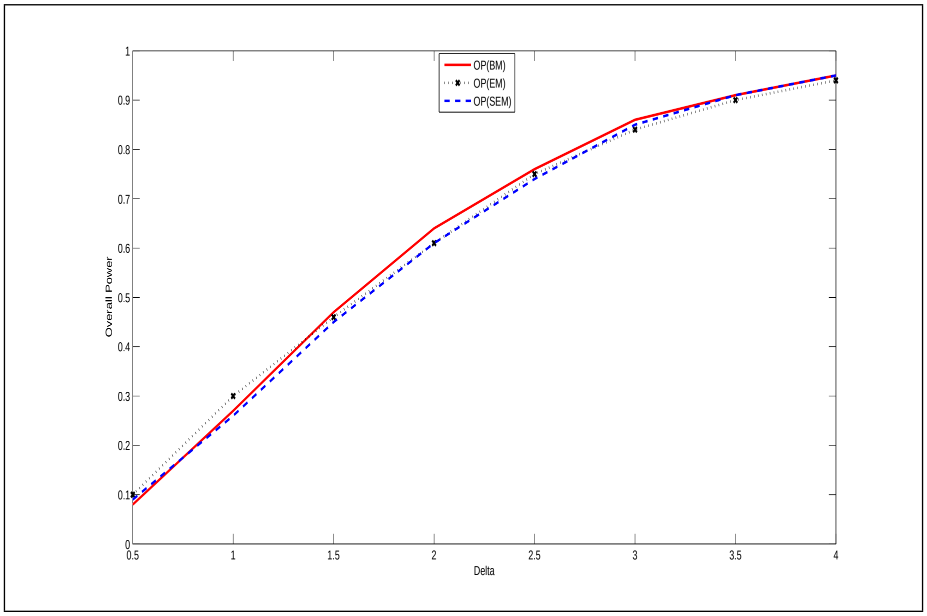

Figure 2.2: Power Curve for Three Methods: BM, EM and SEM

Table 2.3: Performance Comparison ($\delta = 3\sigma$, $n = 60$, $\mu_1^0 = 1$, $\mu_2^0 = 2$)

	m	OP	$AVTI$
BM	5	0.84	0.154
	17	0.85	0.144
	30	0.85	0.152
	45	0.85	0.141
EM	5	0.85	0.152
	17	0.84	0.160
	30	0.85	0.141
	45	0.83	0.153
SEM	5	0.83	0.161
	17	0.86	0.155
	30	0.85	0.140
	45	0.84	0.151

and SEM is fast and at most two iterations would suffice. Comparison of EM and SEM shows that with small m, EM has a larger OP. In the sense of complexity, EM follows a more general framework compared with SEM while SEM is simpler in both theory and application but with reduced power for small m.

In order to find the gross error magnitude, one can calculate the average of data for the two segments based on the estimated change point m and finally, obtain the difference between these calculated average values.

2.6 Conclusion

In this chapter, three methods for solving the mean shift detection problem are derived based on Bayesian and EM approaches. The focus was on univariate data with single change. In next chapter, this problem is solved taking into account the multivariate data with single and multiple changes. Moreover, problems arising from prior selection and initialization of EM are investigated.

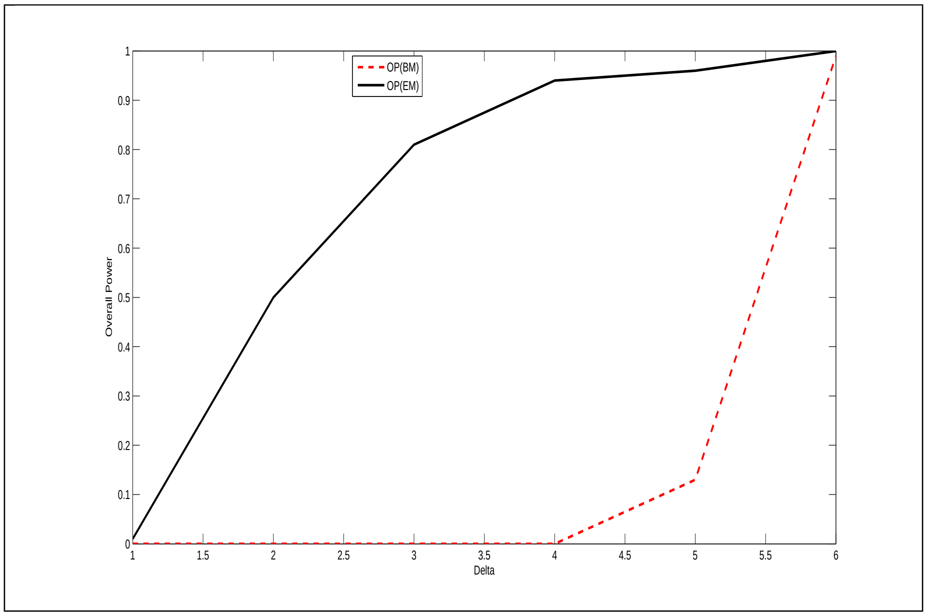

Figure 2.3: Power Curve for Two Methods: BM and EM ($n = 500$, No. of simulations=10000, $\mu_1^0 = 20$, $\mu_2^0 = 30$)

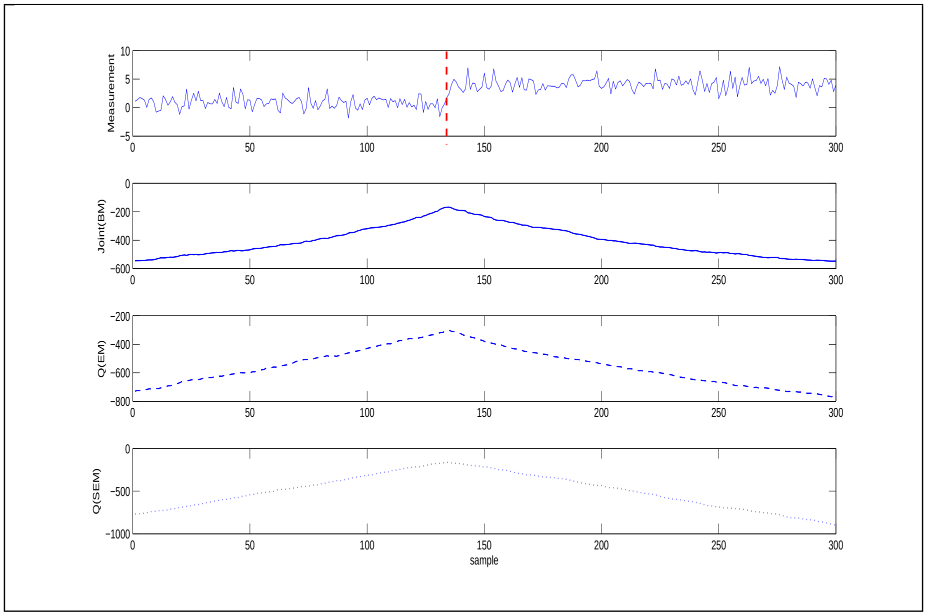

Figure 2.4: Joint Function (BM) and Q Function Curves (EM & SEM) ($n = 300$, $m = 135$, $\mu_1^0 = 1$, $\mu_2^0 = 2$, $\delta = 3\sigma_0$)

Chapter 3

Multivariate Mean Shift Detection

In this chapter, change point detection problem is formulated and solutions based on Bayesian and EM methods are derived. This change detection corresponds to mean shift detection in multivariate data.

In the following sections, a closed form solution to the Bayesian formulation of single and multiple change point detection problem is first considered for multivariate data and MAP is used for the estimation of the parameters. Moreover, considering the sensitivity of the Bayesian approach to prior selection, EM is adopted to solve both single or multiple change points detection problem. By comparison, it is shown that EM is more powerful when priors are highly uncertain while the Bayesian approach has its advantage of less computation demand.

3.1 Bayesian Change Point Detection

3.1.1 Problem Formulation for Single Change point

Here, a multivariate Bayesian formulation of change point detection is given where maximum a posteriori (MAP) is applied to infer the change point detection. Throughout this work, time instant for single change point is referred to as m and multiple time instants for multiple change points are represented by the vector $t = [t_1, ..., t_N]$ where N is the total number of change points. In addition, the covariance of data is assumed to be the same before and after the change points.

Consider n observations from p variables form a $p \times n$ matrix as

$$D = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} & y_{1(m+1)} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2m} & y_{2(m+1)} & \cdots & y_{2n} \\ \vdots & \vdots & \ldots & \vdots & \vdots & \ldots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pm} & y_{p(m+1)} & \cdots & y_{pn} \end{pmatrix} = (Y_1, Y_2, ..., Y_m, ..., Y_n)$$

$Y_1, ..., Y_m, ..., Y_n$ are measurements of p variables from time instant 1 to n. Assume that at the sampling instant m, a change occurs resulting in a shift in the mean vector. Thus, the whole data are split into two segments operating at two different means, μ_1 and μ_2, respectively with the same covariance matrix Σ. This general problem formulation framework is adopted throughout this work.

3.1.2 Existence of Change Point

If we are interested in verification of existence of change point, it is basically testing the hypotheses of "change" against "no change". In other words, two models are compared with each other:
No change model (H_0):

$$Y_i \sim \mathcal{N}_p(\mu_1, \Sigma), i = 1, 2, ..., n$$

Change model (H_1):

$$\begin{aligned} Y_i &\sim \mathcal{N}_p(\mu_1, \Sigma), i = 1, 2, ..., m \\ Y_i &\sim \mathcal{N}_p(\mu_2, \Sigma), i = m+1, m+2, ..., n \end{aligned}$$

In order to perform the analysis, Bayes factor can be used as in[8]. It is a defined as ratio of posterior to prior odds on H_1 against H_0 as

$$\begin{aligned} B(D) &= \frac{\frac{P(H_1|D)}{P(H_0|D)}}{\frac{P(H_1)}{P(H_0)}} \\ &= \frac{P(D|H_1)}{P(D|H_0)} \\ &= \frac{\int P(D|\mu_1, \mu_2, m, \Sigma)P(\mu_1, \mu_2, \Sigma)P(m)}{\int P(D|\mu_1, \Sigma)P(\mu_1, \Sigma)} \end{aligned} \tag{3.1}$$

where D is the data. In the case of unknown mean vectors and covariance, which is a general case, the closed form of Bayes factor is derived in [8]. The decision rule for selecting one hypothesis against the other is of the form

$$B(D) \succ \frac{l_{10}P(H_0)}{l_{01}P(H_1)} \tag{3.2}$$

l_{kj} is the loss caused from deciding on H_k while H_j is true. Also there is no loss when correct conclusions are made ($l_{kk} = l_{jj} = 0$). In determining the threshold of Bayes test, the prior probabilities of each hypothesis along with the loss associated with every wrong decision is considered. If one assumes equal probability for each hypothesis and also equivalent loss for wrong decision then the threshold of Bayes factor can be written as

$$B(D) \succ 1 \tag{3.3}$$

The closed form expression for Bayes factor can be derived as a function of hyperparameters of prior distributions. Having checked existence of change, in next sections, the problem for both single and multiple change points detection is studied further.

3.1.3 Single Change Point Detection

In this section, Bayesian solution is reviewed and a closed form solution for a single change point problem is derived. This solution assists in solving multiple change points detection problem. Using Bayesian analysis, the objective is to find $P(m|D)$ which is the posterior probability of change point given the data. According to Bayes rule, this probability is equal to $\frac{P(m,D)}{P(D)}$. In the following, this posterior probability is derived in detail. Here, it is assumed that observations are independent of each other and follow Gaussian distribution as

$$\begin{aligned} Y_i \sim &\mathcal{N}_p(\mu_1, \Sigma), i = 1, 2, ..., m \\ Y_i \sim &\mathcal{N}_p(\mu_2, \Sigma), i = m+1, m+2, ..., n \end{aligned} \tag{3.4}$$

where $\mu_1 \neq \mu_2$. The normal distribution function can be expressed as

$$\mathcal{N}_p(\mu, \Sigma) = (2\pi)^{-p/2}|\Sigma^{-1}|^{1/2} exp\{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\} \tag{3.5}$$

The shift time, m, can occur anywhere from 1 to $n-1$. The likelihood function of data, D, is therefore of the form

$$P(D = (Y_1, Y_2, ..., Y_n)|\mu_1, \mu_2, m, \Sigma) = \prod_{i=1}^{m} P(Y_i|\mu_1, \Sigma) \prod_{i=m+1}^{n} P(Y_i|\mu_2, \Sigma) \tag{3.6}$$

where μ_1, μ_2 and m are to be determined. The priors for μ_1, μ_2 and m are taken as

$$\begin{aligned} P(\mu_1|\mu_1^0, \Sigma_{01}) &= \mathcal{N}_p(\mu_1^0, \Sigma_{01}) \\ P(\mu_2|\mu_2^0, \Sigma_{02}) &= \mathcal{N}_p(\mu_1^0, \Sigma_{02}) \\ P(m) &= \text{Uniform distribution} \end{aligned} \tag{3.7}$$

where $\mu_1^0, \Sigma_{10}, \mu_2^0, \Sigma_{20}$ are the hyperparameters of prior distributions and assumption of uniform distribution of m follows from [7]. Denote $\beta_0 = (\mu_1^0, \Sigma_{10}, \mu_2^0, \Sigma_{20})$. Hence, the joint distribution between observations and parameters is represented as

$$\begin{aligned} P(D, \mu_1, \mu_2, &\Sigma, \beta_0, m) \\ &= P(D|\mu_1, \mu_2, \Sigma, \beta_0, m)P(\mu_1, \mu_2, \Sigma, \beta_0, m) \\ &= P(D|\mu_1, \mu_2, \Sigma, \beta_0, m)P(\mu_1|\mu_2, \Sigma, \beta_0, m)P(\mu_2, m, \Sigma, \beta_0) \\ &= P(D|\mu_1, \mu_2, \Sigma, \beta_0, m)P(\mu_1|\mu_2, \Sigma, \beta_0, m)P(\mu_2|m, \Sigma, \beta_0)P(m) \\ &= P(D|\mu_1, \mu_2, \Sigma, \beta_0, m)P(\mu_1|\beta_0)P(\mu_2|\beta_0)P(m) \end{aligned} \tag{3.8}$$

In derivation of (3.8), independence assumption of priors is incorporated. Moreover, the prior distributions of μ_1 and μ_2 are completely determined by their hyperparameters β_0. By substituting the priors as in (3.7) and likelihood as in (3.6), into (3.8), the joint distribution can be determined as

$$P(D, \mu_1, \mu_2, \Sigma, \beta_0, m) = k_1 \prod_{i=1}^{m} P(Y_i|\mu_1, \Sigma) \prod_{i=m+1}^{n} P(Y_i|\mu_2, \Sigma)\mathcal{N}_p(\mu_1^0, \Sigma_{01})\mathcal{N}_p(\mu_2^0, \Sigma_{02}) \tag{3.9}$$

where k_1 is a constant.

With the likelihood and priors shown in (3.6) and (3.7), the joint distribution of data and parameters has an explicit closed form expression as follows:

$$P(D, \mu_1, \mu_2, \Sigma, \beta_0, m) = F_1 \times F_2 \tag{3.10}$$

where

$$\begin{aligned} F_1 &= C_1 exp\{-\frac{1}{2}\left(\mu_1^T \Delta_n^{-1} \mu_1 - \mu_1^T \Delta_n^{-1} \Omega_n - \Omega_n^T \Delta_n^{-1} \mu_1 + \Omega_n^T \Delta_n^{-1} \Omega_n\right)\}\times \\ &\quad exp\{-\frac{1}{2}(\mu_1^{0^T} \Sigma_{01}^{-1} \mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i - B^T A^{-1} B)\} \\ &= C_1 exp\{-\frac{1}{2}\left(\mu_1 - \Omega_n\right)' \Delta_n^{-1}\left(\mu_1 - \Omega_n\right)\} \times exp\{-\frac{1}{2}(\mu_1^{0^T} \Sigma_{01}^{-1} \mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i - B^T A^{-1} B)\} \end{aligned} \tag{3.11}$$

and

$$F_2 = C_2 exp\{-\frac{1}{2}\left(\mu_2 - \Psi_n\right)' \Lambda_n^{-1}\left(\mu_2 - \Psi_n\right)\} \times exp\{-\frac{1}{2}(\mu_2^{0^T} \Sigma_{02}^{-1} \mu_2^0 + \sum_{i=m+1}^{n} y_i^T \Sigma^{-1} y_i - D^T C^{-1} D)\} \tag{3.12}$$

where C_1 and C_2 are constant and A, B, C and D are defined in Appendix A. The derivations of F_1 and F_2 are given in Appendix A.

In order to determine the change point, first we need to integrate out μ_1 and μ_2 from the joint distribution where F_1 is a function of μ_1 and F_2 is a function of μ_2. Integration of F_1 with respect to μ_1 leads to

$$\begin{aligned} \int_{-\infty}^{+\infty} F_1\, d\mu_1 &= C_1 exp\{-\frac{1}{2}(\mu_1^{0^T} \Sigma_{01}^{-1} \mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i - B^T A^{-1} B)\}\times \\ &\quad \int_{-\infty}^{+\infty} exp\{-\frac{1}{2}\left(\mu_1 - \Omega_n\right)' \Delta_n^{-1}\left(\mu_1 - \Omega_n\right)\}\, d\mu_1 \\ &= C_1 exp\{-\frac{1}{2}(\mu_1^{0^T} \Sigma_{01}^{-1} \mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i - B^T A^{-1} B)\} \times (2\pi)^{p/2} |\Delta_n^{-1}|^{-1/2} \end{aligned} \tag{3.13}$$

Note that the integration is with respect to μ_1 which is a p-dimensional vector. Since the integrand is a multivariate normal density function with mean Ω_n and covariance Δ_n, we have used the following relation in the derivation of (3.13):

$$\int_{-\infty}^{+\infty} exp\{-\frac{1}{2}\left(\mu_1 - \Omega_n\right)' \Delta_n^{-1}\left(\mu_1 - \Omega_n\right)\}\, d\mu_1 = (2\pi)^{p/2} |\Delta_n^{-1}|^{-1/2} \tag{3.14}$$

Similarly by integrating F_2 with respect to μ_2, we have

$$\int_{-\infty}^{+\infty} F_2\, d\mu_2 = C_2 exp\{-\frac{1}{2}(\mu_2^{0^T} \Sigma_{02}^{-1} \mu_2^0 + \sum_{i=m+1}^{n} y_i^T \Sigma^{-1} y_i - D^T C^{-1} D)\} \times (2\pi)^{p/2} |\Lambda_n^{-1}|^{-1/2} \tag{3.15}$$

Combining the marginal distribution of (3.13) and (3.15) yields

$$
\begin{aligned}
P(m|D) &\propto \int\int P(D, \mu_1, \mu_2, \Sigma, \beta_0, m) d\mu_1 d\mu_2 \\
&= C_1 exp\{-\frac{1}{2}(\mu_1^{0^T}\Sigma_{01}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i - B^T A^{-1} B)\} \times (2\pi)^{p/2}|\Delta_n^{-1}|^{-1/2} \times \\
&C_2 exp\{-\frac{1}{2}(\mu_2^{0^T}\Sigma_{02}^{-1}\mu_2^0 + \sum_{i=m+1}^{n} y_i^T \Sigma^{-1} y_i - D^T C^{-1} D)\} \times (2\pi)^{p/2}|\Lambda_n^{-1}|^{-1/2} \quad (3.16)
\end{aligned}
$$

The single change point can now be determined by maximizing (3.16) with respect to m where $m = 1, 2, ..., n-1$.

Having obtained a solution for single change point detection, one can use the results obtained to solve the detection problem with multiple change points.

3.1.4 Problem Formulation for Multiple Change Points

In reality, the data can be subject to change at multiple time points. Consider a set of data with length n and N possible change points. The shift points occur at $t = [t_1, ..., t_N]$. In other words, the data can be split into $N+1$ segments: each segment has its own mean vector. It is assumed that the covariance does not change before and after each change point. Also, $t_0 = 1$ and $t_{N+1} = n$. A similar problem was considered in [10] but it was solved numerically based on MCMC method and Gibbs sampling. In this following section, an explicit analytical solution is derived which will facilitate the optimization solution.

3.1.5 Multiple Change Points Detection

Using Bayesian methods requires a prior distribution for vector t. The change point can be modelled by

$$t_{i+1} = t_i + \epsilon_i \quad with \quad \epsilon_i \sim \mathcal{P}(\lambda) \tag{3.17}$$

where ϵ_i is independently identically distributed (iid) with Poisson distribution [10]. λ is a priori mean value of time intervals $t_{i+1} - t_i$. Set $\lambda = \frac{n}{N+1}$ and thus

$$p(t|\lambda, N) = p(\Delta t|\lambda, N) = \prod_{i=0}^{N} \mathcal{P}((t_{i+1} - t_i)|\lambda) = \prod_{i=0}^{N} (e^{-\lambda}\frac{\lambda^{(t_{i+1}-t_i)}}{(t_{i+1}-t_i)!}) \tag{3.18}$$

If the data segments are represented as $D = (Y_0, Y_1, ..., Y_N)^T$ so that $Y_i = (y(t_i+1), y(t_i+2), ..., y(t_{i+1}))^T$, then the likelihood function can be expressed as

$$P(D|\mu_i, t, \Sigma, N) = \prod_{i=0}^{N} P(Y_i|\mu_i, \Sigma) \tag{3.19}$$

The prior of t is given by (3.18) and that of μ_i follows Gaussian distribution, i.e. $P(\mu_i) = \mathcal{N}_p(\mu_i^0, \Sigma_{i0})$. Thus, the posterior probability can be written as

$$\begin{aligned} P(t, \mu_i | D, \lambda, \Sigma, N) &\propto P(D|\mu_i, t, \Sigma, N) P(t|\lambda, N) P(\mu_i|N) \\ &\propto \prod_{i=0}^{N} P(Y_i|\mu_i, \Sigma) \prod_{i=0}^{N} [e^{-\lambda} \frac{\lambda^{(t_{i+1}-t_i)}}{(t_{i+1}-t_i)!} P(\mu_i)] \end{aligned} \tag{3.20}$$

In order to further derive the desired posterior with respect to the change point $P(t|Y, \lambda, \Sigma, N)$, one can integrate out (3.20) with respect to μ_i and obtain the marginal posterior distribution. In (3.20), the first and third terms depend on μ_i. Thus, following the same procedure as in the derivation of (3.16), integrating out μ_i in (3.20) yields

$$\begin{aligned} P(t|Y, \lambda, \Sigma, N) \propto \prod_{i=0}^{N} (e^{-\lambda} \frac{\lambda^{(t_{i+1}-t_i)}}{(t_{i+1}-t_i)!}) \times C_k exp\{-\frac{1}{2} \sum_{i=0}^{N} (\mu_i^{0T} \Sigma_{i0}^{-1} \mu_i^0 \\ + \sum_{j=t_i+1}^{t_{i+1}} [y_j^T \Sigma^{-1} y_j] - B_i^T A_i^{-1} B_i)\} \times |\Delta_i^{-1}|^{-1/2} \end{aligned} \tag{3.21}$$

where C_k is a constant, $|.|$ represents determinant of matrix and

$$\begin{aligned} A_i &= (t_{i+1} - t_i)\Sigma^{-1} + \Sigma_{i0}^{-1} \\ B_i &= (t_{i+1} - t_i)\Sigma^{-1}\bar{y} + \Sigma_{i0}^{-1}\mu_i^0 \\ \bar{y} &= \frac{1}{t_{i+1} - t_i} \sum_{i=t_i}^{t_{i+1}} y_i \\ \Delta_i &= A_i^{-1} = ((t_{i+1} - t_i)\Sigma^{-1} + \Sigma_{i0}^{-1})^{-1} \end{aligned} \tag{3.22}$$

The next step is to maximize (3.21) with respect to the vector $t = [t_1, ..., t_N]$ and determine the change points. i.e.,

$$\hat{t} = \arg\max_t P(t|Y, \lambda, \Sigma, N) \tag{3.23}$$

Having derived the closed form expression for marginal posterior probability, the change points can be determined through optimization. This algorithm performs well provided that the priors are selected appropriately. However, if selection of priors has considerable uncertainty, an intuitive solution would be to improve the priors using available data. Using EM algorithm, one can achieve this objective even if the priors are not chosen properly. In next section, EM formulation of change point problem is presented.

3.2 Expectation Maximization(EM)

In chapter one, EM solution to change point detection is derived. But that solution was limited to univariate data and single change point detection. In the following, we generalize the solution to a broader range of problem.

3.2.1 Multivariate EM for Single Change Point Detection

In this section, an EM algorithm is applied to solve single change point detection problem as formulated earlier. To solve the problem, μ_1 and μ_2 are treated as the missing or hidden variables, $D = (Y_1, Y_2, ..., Y_n)$ are the observed data and m is the parameter of the interest to be determined. Thus, E-step can be expressed as

$$Q(m|m^{(k)}) = E_{\mu_1,\mu_2|Y,m^{(k)}}\{P(D, \mu_1, \mu_2|m)\} \tag{3.24}$$

In M-step, the maximization is performed with respect to the parameter, m, as

$$m^{(k+1)} = \arg\max_m Q(m|m^{(k)}) \tag{3.25}$$

In order to derive the Q-function in E-step, we write

$$\begin{aligned} P(D, \mu_1, \mu_2|m) &= P(D|\mu_1, \mu_2, m)P(\mu_1, \mu_2|m) \\ &= P(D|\mu_1, \mu_2, m)P(\mu_1|\mu_2, m)P(\mu_2|m) \end{aligned} \tag{3.26}$$

Since m, the shift time instant, is independent of the means, μ_1 and μ_2, and the two mean vectors are also independent of each other, (3.26) can be written as

$$P(D, \mu_1, \mu_2|m) = P(D|\mu_1, \mu_2, m)P(\mu_1)P(\mu_2) \tag{3.27}$$

By substituting (3.27) in (3.24), we have

$$\begin{aligned} Q(m|m^{(k)}) &= E_{\mu_1,\mu_2|D,m^{(k)}}\{P(D|\mu_1, \mu_2, m)P(\mu_1)P(\mu_2)\} \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P(D|\mu_1, \mu_2, m)P(\mu_1)P(\mu_2)P(\mu_1, \mu_2|D, m^{(k)})d\mu_1 d\mu_2 \end{aligned} \tag{3.28}$$

Since μ_1 and μ_2 are independent, Q-function can be expressed as

$$Q(m|m^{(k)}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P(D|\mu_1, \mu_2, m)P(\mu_1)P(\mu_2)P(\mu_1|D, m^{(k)})P(\mu_2|D, m^{(k)})d\mu_1 d\mu_2 \tag{3.29}$$

where $P(\mu_1|D, m^{(k)})$ can be determined following the Bayesian rule such that

$$\begin{aligned} P(\mu_1|D, m^{(k)}) &= \frac{P(D|\mu_1, m^{(k)})P(\mu_1|m^{(k)})}{p(D|m^{(k)})} \\ &= \frac{P(Y_1, Y_2, ..., Y_{m^{(k)}}|\mu_1, m^{(k)})P(Y_{m^{(k)}+1}, Y_{m^{(k)}+2}, ..., Y_n|m^{(k)})P(\mu_1)}{P(D|m^{(k)})} \\ &= k_1 P(Y_1, Y_2, ..., Y_{m^{(k)}}|\mu_1, m^{(k)})P(\mu_1) \end{aligned} \tag{3.30}$$

where $k_1 = \frac{P(Y_{m^{(k)}+1}, Y_{m^{(k)}+2}, ..., Y_n|m^{(k)})}{P(D|m^{(k)})}$.

Similarly, $P(\mu_2|D, m^{(k)})$ can also be derived as

$$P(\mu_2|D, m^{(k)}) = k_2 P((Y_{m^{(k)}+1}, Y_{m^{(k)}+2}, ..., Y_n)|\mu_2, m^{(k)})P(\mu_2) \tag{3.31}$$

where $k_2 = \frac{P(Y_1,Y_2,...,Y_m^{(k)}|m^{(k)})}{P(D|m^{(k)})}$.

As a result, (3.29) can be rewritten as

$$\begin{aligned}
&Q(m|m^{(k)}) = k_1 k_2 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P(Y|\mu_1,\mu_2,m)P(\mu_1)P(\mu_2)\times \\
&P(Y_{1:m^{(k)}}|\mu_1,m^{(k)})P(\mu_1)P(Y_{m^{(k)}+1:n}|\mu_2,m^{(k)})P(\mu_2)d\mu_1 d\mu_2 \\
&= k_3 \int_{-\infty}^{+\infty} P(Y_{1:m}|\mu_1,m)P(Y_{1:m^{(k)}}|\mu_1,m^{(k)})(P(\mu_1))^2 d\mu_1\times \\
&\int_{-\infty}^{+\infty} P(Y_{m+1:n}|\mu_2,m)P(Y_{m^{(k)}+1:n}|\mu_2,m^{(k)})(P(\mu_2))^2 d\mu_2
\end{aligned} \tag{3.32}$$

where $k_3 = k_1 k_2$. Consequently, (3.32) can be simplified as

$$Q(m|m^{(k)}) = k_3 \int_{-\infty}^{+\infty} E_1 d\mu_1 \int_{-\infty}^{+\infty} E_2 d\mu_2 \tag{3.33}$$

where E_1 and E_2 are

$$\begin{aligned}
E_1 &= P(Y_{1:m}|\mu_1,m)P(Y_{1:m^{(k)}}|\mu_1,m^{(k)})(P(\mu_1))^2 \\
E_2 &= P(Y_{m+1:n}|\mu_2,m)P(Y_{m^{(k)}+1:n}|\mu_2,m^{(k)})(P(\mu_2))^2
\end{aligned} \tag{3.34}$$

After some algebraic simplification, we have

$$\begin{aligned}
Q(m|m^{(k)}) = {}& C_1 \times exp\{-\frac{1}{2}(2{\mu_1^0}^T{\Sigma_{01}}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T\Sigma^{-1}y_i + \sum_{i=1}^{m^{(k)}} y_i^T\Sigma^{-1}y_i - {B_1}^T{A_1}^{-1}B_1)\}\times \\
& (2\pi)^{p/2}|{\Delta_{nn}}^{-1}|^{-1/2}\times \\
& C_2 \times exp\{-\frac{1}{2}(2{\mu_2^0}^T{\Sigma_{02}}^{-1}\mu_2^0 + \sum_{i=m+1}^{n} y_i^T\Sigma^{-1}y_i + \sum_{i=m^{(k)}+1}^{n} y_i^T\Sigma^{-1}y_i - {D_1}^T{G_1}^{-1}D_1)\}\times \\
& (2\pi)^{p/2}|{\Lambda_{nn}}^{-1}|^{-1/2}
\end{aligned} \tag{3.35}$$

where C_1 and C_2 are constant and

$$\begin{aligned}
&A_1 = m\Sigma^{-1} + m^{(k)}\Sigma^{-1} + 2{\Sigma_{01}}^{-1} \\
&B_1 = m\Sigma^{-1}\bar{y}_1 + m^{(k)}\Sigma^{-1}\bar{y}_{1k} + 2{\Sigma_{01}}^{-1}\mu_1^0 \\
&G_1 = (n-m)\Sigma^{-1} + (n-m^{(k)})\Sigma^{-1} + 2{\Sigma_{02}}^{-1} \\
&D_1 = (n-m)\Sigma^{-1}\bar{y}_2 + (n-m^{(k)})\Sigma^{-1}\bar{y}_{2k} + 2{\Sigma_{02}}^{-1}\mu_2^0 \\
&\Delta_{nn} = {A_1}^{-1}, \Lambda_{nn} = {G_1}^{-1} \\
&\bar{y}_1 = \frac{1}{m}\sum_{i=1}^{m} y_i,\ \ \bar{y}_{1k} = \frac{1}{m^{(k)}}\sum_{i=1}^{m^{(k)}} y_i,\ \ \bar{y}_2 = \frac{1}{n-m}\sum_{i=m+1}^{n} y_i,\ \ \bar{y}_{2k} = \frac{1}{n-m^{(k)}}\sum_{i=m^{(k)}+1}^{n} y_i
\end{aligned} \tag{3.36}$$

Having derived the Q-function, the next step, the M-step, is to solve the following optimization problem:

$$m^{(k+1)} = \arg\max_m Q(m|m^{(k)}) \tag{3.37}$$

where Q-function depends on both current estimate of the parameter $m^{(k)}$ and decision variable m. In M-step of EM algorithm, this function is maximized with respect to m. In other words, starting this loop with an initial value for $m(0) = m_0$, E-step and M-step iterate until no further change is observed in Q-function.

3.2.2 EM for Multiple Change Points Detection of Multivariate Data

Here, EM solution for single change point detection is extended in order to solve the multiple change points detection problem as formulated in previous section. In this EM formulation, the hidden variables are μ_i for $i = 0, 1, ..., N$. The vector $t = [t_1, ..., t_N]$ is the parameter to be estimated. Thus, the E-step in EM can be represented by

$$Q(t|t^{(k)}) = E_{\mu_0,\mu_1,....,\mu_N|D,t^{(k)}}\{P(Y, \mu_0, \mu_1, ..., \mu_N|t)\} \tag{3.38}$$

where

$$P(D, \mu_0, \mu_1, ..., \mu_N|t) = \prod_{i=0}^{N} P(Y_i|\mu_i, \Sigma) \prod_{i=0}^{N} P(\mu_i) \tag{3.39}$$

Thus, E-step can be formulated as

$$\begin{aligned} Q(t|t^{(k)}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \prod_{i=0}^{N} P(Y_i|\mu_i, \Sigma) \prod_{i=0}^{N} P(\mu_i) \times \\ P(\mu_0|D, t^k)P(\mu_1|D, t^k)....P(\mu_N|D, t^k)d\mu_0 d\mu_1...d\mu_N \end{aligned} \tag{3.40}$$

In the derivation of (3.40), the independence assumption of priors are taken into account, $P(\mu_i) = \mathcal{N}_p(\mu_i^0, \Sigma_{i0})$, for $\mu_i, i = 0, ..., N$. Following the same derivation as (3.30) and (3.31) for each $P(\mu_i|D, t^k)$, we have

$$\begin{aligned} P(\mu_i|D, t^k) &= \frac{P(D|\mu_i, t^k)P(\mu_i|t^k)}{P(D|t^k)} \\ &= \frac{P(Y_0, Y_1, ..., Y_N|\mu_i, t^k)P(\mu_i)}{P(D|t^k)} \\ &= \frac{P(Y_i|\mu_i, t^k)P(Y_0, Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_N|t^k)P(\mu_i)}{P(D|t^k)} \\ &= k_i P(Y_i|\mu_i, t^k)P(\mu_i) \end{aligned} \tag{3.41}$$

where k_i can be written as

$$k_i = \frac{P(Y_0, Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_N|t^k)}{P(Y_0, Y_1, ..., Y_{i-1}, Y_i, Y_{i+1}, ..., Y_i|t^k)} \tag{3.42}$$

Substituting (3.41) into (3.40) yields

$$\begin{aligned} Q(t|t^{(k)}) = C_1 \int_{-\infty}^{+\infty} & \mathcal{N}_p(Y_0|\mu_0, \Sigma, t)\mathcal{N}_p(Y_0|\mu_0, \Sigma, t^k)(\mathcal{N}_p(\mu_0|\mu_0^0, \Sigma_{00}))^2 d\mu_0 \times \\ & \int_{-\infty}^{+\infty} \mathcal{N}_p(Y_1|\mu_1, \Sigma, t)\mathcal{N}_p(Y_1|\mu_1, \Sigma, t^k)(\mathcal{N}_p(\mu_1|\mu_1^0, \Sigma_{10}))^2 d\mu_1 \times ... \\ & .... \int_{-\infty}^{+\infty} \mathcal{N}_p(Y_N|\mu_0, \Sigma, t)\mathcal{N}_p(Y_1|\mu_1, \Sigma, t^k)(\mathcal{N}_p(\mu_N|\mu_N^0, \Sigma_{N0}))^2 d\mu_N \end{aligned} \tag{3.43}$$

where C_1 is a constant. Using the same algebraic simplification as performed in (3.35) and (3.36) for single change detection leads to

$$Q(t|t^{(k)}) = C_2 \prod_{i=0}^{N} exp\{-\frac{1}{2}[\sum_{i=0}^{N}\{2{\mu_i^0}^T \Sigma_{i0}^{-1} \mu_i^0 + \sum_{j\in[t_i,t_{i+1}]} [Y_j^T \Sigma^{-1} Y_j] - {B_i}^T {A_i}^{-1} B_i\}]\}(2\pi)^{p/2}|{\Delta_i}^{-1}|^{-1/2} \tag{3.44}$$

where C_2 is a constant, $|.|$ represents determinant of matrix and

$$\begin{aligned} A_i &= (t_{i+1} - t_i)\Sigma^{-1} + (t_{i+1}^k - t_i^k)\Sigma^{-1} + 2{\Sigma_{i0}}^{-1} \\ B_i &= (t_{i+1} - t_i)\Sigma^{-1}\bar{y} + (t_{i+1}^k - t_i^k)\Sigma^{-1}\bar{y}_k + 2{\Sigma_{i0}}^{-1}\mu_i^0 \\ \bar{y} &= \frac{1}{t_{i+1} - t_i} \sum_{j\in[t_i,t_{i+1}]} Y_j \\ \bar{y}_k &= \frac{1}{t_{i+1}^k - t_i^k} \sum_{j\in[t_i^k,t_{i+1}^k]} Y_j \\ \Delta_i &= A_i^{-1} \end{aligned} \tag{3.45}$$

Using this Q-function, the M-step can be written as

$$t^{(k+1)} = \arg\max_t Q(t|t^{(k)}) \tag{3.46}$$

and consequently every change point can be determined through this optimization problem.

3.3 Discussion

In the following sections, we discuss the effects of prior in Bayesian formulation and initial value selection in EM.

3.3.1 Prior Selection

The advantage of the Bayesian framework lies in the fact that one can combine the prior with the observations through the likelihood. However, this advantage can also turn to a disadvantage. Selection of prior is often subjective. In other words, converting prior knowledge and belief into a mathematical prior distribution is not an easy task . If priors are improperly selected, false or misleading results could be expected and the posterior expression may be affected by the priors significantly [21].

In this work, the prior distributions considered for mean vectors before and after the change point have been assumed to follow a Gaussian distribution. For the change point, a uniform prior distribution has been commonly selected for the single change point problem [7] and a Poisson distribution for multiple change points problem [10]. The selection of distribution for mean values has an impact on the posterior distribution. For instance, in single change point problem, according to (3.16) one can see that selection of μ_1^0 and μ_2^0 can influence both B, D and consequently the final estimates. Assuming that one is not aware of permissible range of mean value before and after the change, selection of improper values for such hyperparameters of prior distribution can clearly lead to false detection if the observed data are not sufficiently large. In other words, Bayesian method can be sensitive to selection of these values as the prior. Using the EM algorithm, this problem can be alleviated since EM works on conditional probability according to the most updated prior based on the data and current estimation of parameters. If (3.16) and (3.35) are compared, one can see that in (3.35), updating of parameters such as B, D and Q-function is in the direction that leads to increase in the likelihood function and consequently converges to an optimal value. However, EM is an optimization based algorithm and from optimization perspective, EM algorithm has to deal with initial value which is discussed in detail in the following section.

3.3.2 Initialization of EM Algorithm

Like other optimization techniques, EM can also be sensitive to the selection of initial value. EM increases the likelihood at each iteration till convergence to a stationary point which could be a local optimum. Consequently, initial value selection can play an important role. Moreover, the rate of convergence and the number of iterations can also be dependent on the initial values. In the context of multivariate data, it can be even worse [35]. A good reference on properties of EM can be found in [22].

There are various strategies developed to tackle this problem. Some are based on multiple random initial values to produce multiple solutions and then select the one that leads to the largest likelihood. Some are based on clustering methods [36][37]. Initialization of EM using Principal Component Analysis (PCA) is another approach used in [38]. In [35], a detailed review and a comparative study of existing methods for selection of initial values are given.

In this work, random sets of initial values are selected and the one that maximizes the likelihood is obtained. This method for selection of the initial value has been proven to be effective.

3.4 Simulation

In this section, two simple numerical examples are given first to demonstrate the proposed algorithms, followed by a CSTR simulation example.

3.4.1 Performance Evaluation

In chapter two, two performance measures are defined to evaluate the power of algorithm. One is overall power (OP) and the other one is probability of false alarm (AVTI). These two performance measures are applied here to change point detection problem.

3.4.2 Single Change Point Simulation

In order to simulate the problem of multivariate data with single change point, two correlated variables ($p = 2$) are generated. The simulation parameters are as follows:

$$p = 2,\, n = 200\ ,\, \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\ m = 112,\, \mu_1^0 = [3,4]^T\ ,\, \mu_2^0 = [3,5]^T$$
$$\Sigma_{01} = \Sigma_{02} = I_{2\times 2}\ ,\, \delta = [1,1]^T$$

A time instant $m = 112$, the bias in added to measurements. The actual mean before the shift is $\mu = [1,2]^T$. The magnitude of actual shift is $\delta = [1,1]^T$ which is equal to one standard deviation of the noise added to the data. The measurements $y = [y_1, y_2]^T$ are illustrated in Figure 3.1. The results of MAP solution to the Bayesian method is shown in Figure 3.2. The mode of the marginal posterior is exactly the same as the true m which is where the change point occurs. Figure 3.3 shows the result based on EM algorithm, where $Q-$function for the last iteration is illustrated. The number of iteration to convergence is two indicating fast convergence of EM. This function has reached its peak point at true change point. In addition, in both Figures 3.2 and 3.3, y-axes are shown in logarithmic scale. The results of EM for 3 different magnitudes of δ are shown in Figure 3.4. Apparently, for larger bias, which are easy to detect, the peak is sharper.

To evaluate the repeatability of Bayesian and EM methods, a Monte Carlo simulation of 100 runs is performed. For different bias sizes, the results of overall power, OP and $AVTI$ are shown in Table **??**. In this table, σ_i represents the variance of variable i. As we can see from Table 1, the power of EM especially for small magnitudes of bias is larger than Bayesian approach. Obviously, as magnitude of bias increases, the power increases and hence the rate of false detection decreases accordingly.

As mentioned earlier, improper selection of priors can affect the Bayesian method. To illustrate this, improper values are selected for hyperparameters of prior distribution, i.e. μ_1^0 and μ_2^0. In such a case, the Bayesian method gives wrong results. In other words, the maximum of marginalized posterior occurs at a wrong location. As an example, the results of Bayesian and EM are shown in Figure 3.5 for $\mu_1^0 = [20; 30]^T$ and $\mu_1^0 = [25; 35]^T$. Note that the true shift point is $m = 112$. Obviously, Bayesian method cannot detect the change point correctly while EM is able to identify it. The number of iteration for EM algorithm is three in this case indicating fast convergence of EM.

As a comparison, both Bayesian and EM have advantages and disadvantages. EM shows sensitivity to initial value selection while Bayesian inference is sensitive to priors. However, EM algorithm can be alleviated from its disadvantage by randomization of the initial values. On the other hand, due to iterative nature, EM has heavier computation than Bayesian.

In the next simulated example, multiple change point detection is investigated.

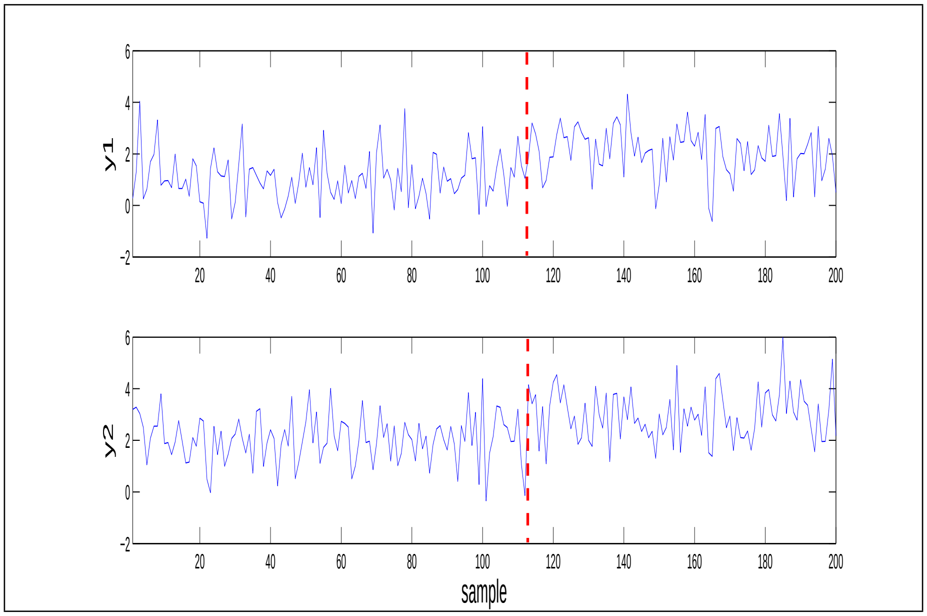

Figure 3.1: Two Correlated Measurements, $y = [y_1, y_2]^T$

3.4.3 Multiple Change Points Simulation

To evaluate the performance of Bayesian approach and the proposed EM for multiple shifts of correlated variables, random variables with $p = 3$ are generated and the number of change points is set as $N = 3$. The change points are denoted as $t = [t_1, t_2, t_3]$. Simulation parameters are selected as

$$p = 3,\ n = 50,\ N = 3,\ \lambda = \tfrac{50}{4} = 12.5,\ \Sigma = \begin{pmatrix} 3 & 0.5 & 0.1 \\ 0.5 & 1 & 0.75 \\ 0.1 & 0.75 & 1 \end{pmatrix},$$

$$t = [t_1, t_2, t_3] = [12, 27, 43],\ \delta_1 = [1.2, 0, 0]^T\ ,\ \delta_2 = [0, 1.3, 0]^T,\ \delta_3 = [0, 0, 1.4]^T\ \Sigma_{0i} = I_{3\times 3}\ ,$$
$$i = 0, 1, 2, 3\ \mu_0^0 = [2, 2, 2]^T,\ \mu_1^0 = [2, 2, 2]^T,\ \mu_2^0 = [3, 3, 3]^T,\ \mu_3^0 = [2, 2, 2]^T$$

The bias is added to every variable as follows: at time $t = t_1$, the bias with magnitude δ_1 is added to measurement vector, i.e. $y = [y_1, y_2, y_3]^T$; at time $t = t_2$, the bias with magnitude δ_2 is added to y; at time $t = t_3$, the bias of magnitude δ_3 is introduced to y. The measurements are shown in Figure 3.6. The results of simulation based on the

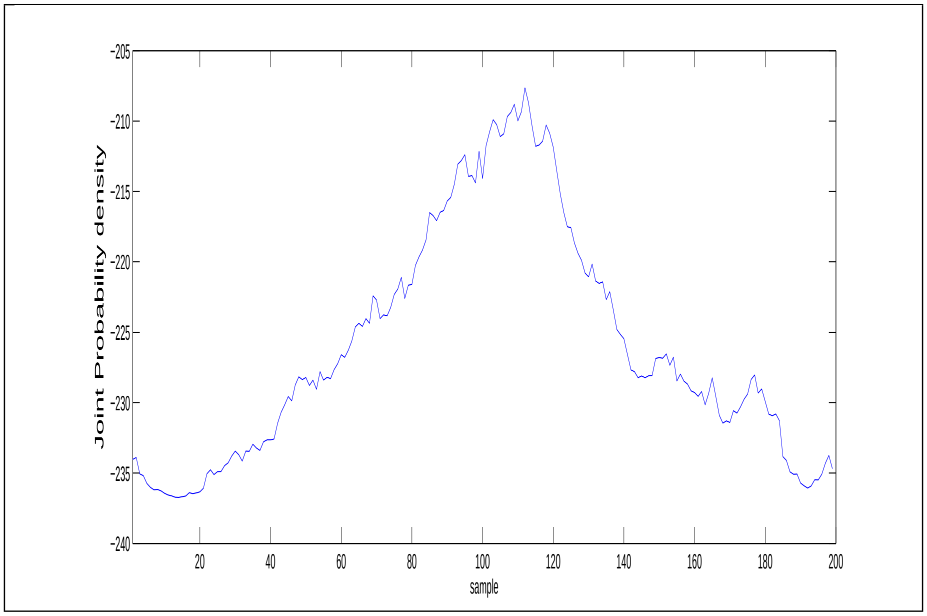

Figure 3.2: Logarithmic Scale of Bayesian Joint Distribution Diagram

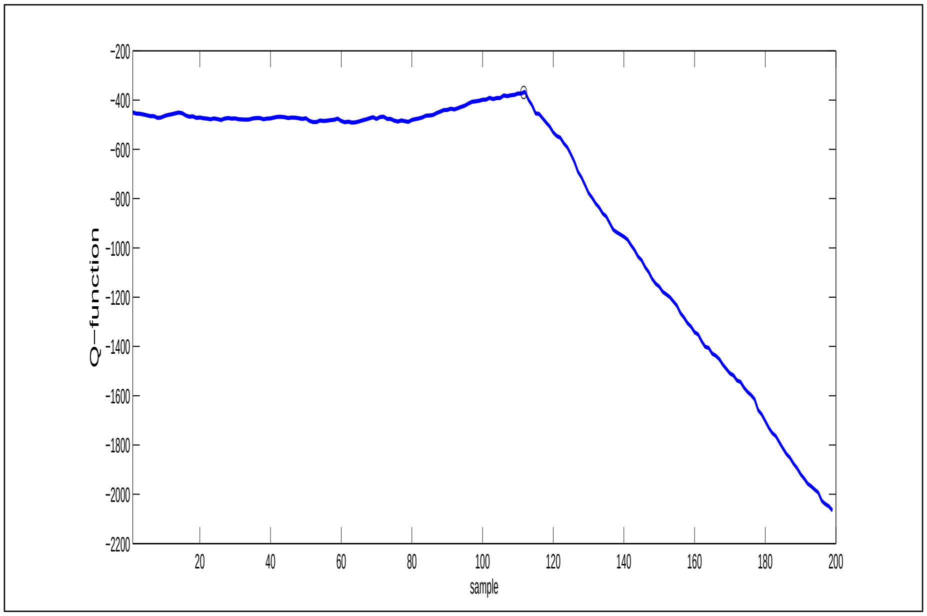

Figure 3.3: Logarithmic Scale of $Q-$function at the Last Iteration of EM

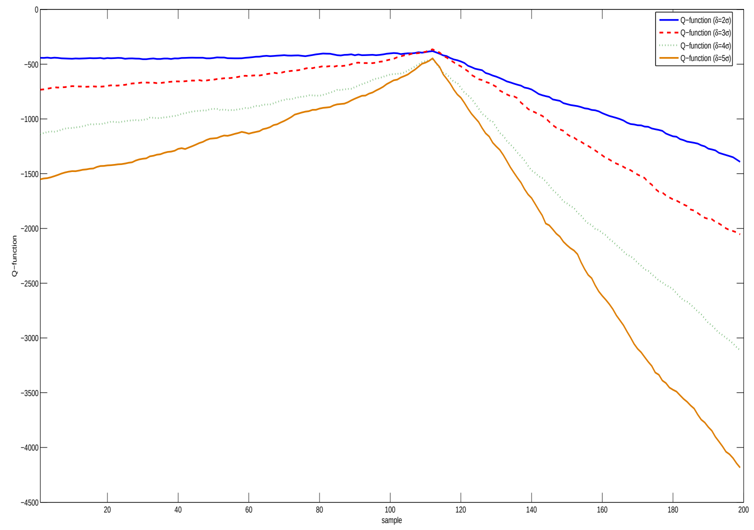

Figure 3.4: The $Q-$function at the Last iteration (3rd Iteration) for Different Magnitudes of Bias

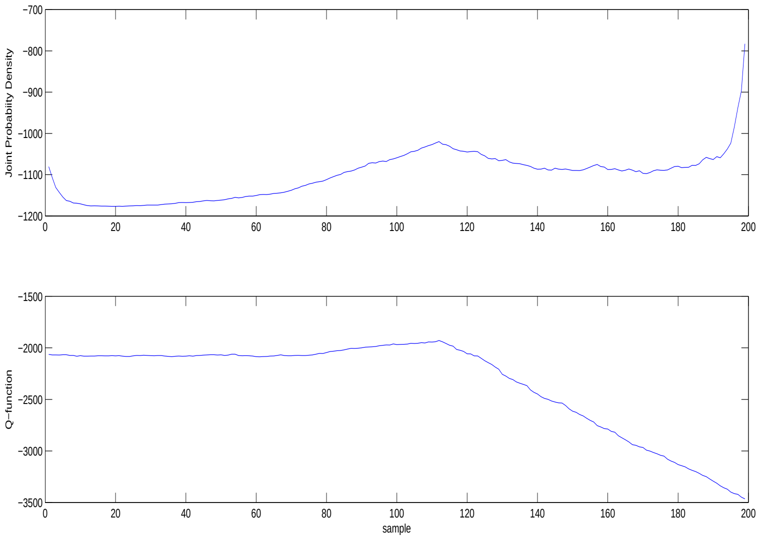

Figure 3.5: The Bayesian Marginalized Joint Distribution (Upper Diagram) and Q-function of EM (Lower Diagram)for $\mu_1^0 = [20; 30]^T$ and $\mu_1^0 = [25; 35]^T$

Table 3.1: Performance Comparison of Bayesian Method (BM) and EM for Different biases ($n = 200$, $m = 112$, $\mu_1^0 = [3; 4], \mu_2^0 = [3; 5]$)

	$Bias$	OP	$AVTI$
BM	$0.5\sigma_i$	0.2	0.8
	$1\sigma_i$	0.36	0.64
	$1.5\sigma_i$	0.56	0.44
	$2\sigma_i$	0.79	0.21
	$2.5\sigma_i$	0.82	0.18
	$3\sigma_i$	0.92	0.08
	$3.5\sigma_i$	1	0
EM	$0.5\sigma_i$	0.32	0.68
	$1\sigma_i$	0.37	0.63
	$1.5\sigma_i$	0.59	0.41
	$2\sigma_i$	0.77	0.23
	$2.5\sigma_i$	0.89	0.11
	$3\sigma_i$	1	0
	$3.5\sigma_i$	1	0

Bayesian method using (3.21) are illustrated in Figure 3.7. Apparently, the MAP gives the maxima at true change points for all three change points. A similar problem was considered using MCMC method in [10] which requires a more complex computation compared with MAP since our Bayesian algorithm has used an analytical solution as derived. As a result, this optimization problem can be simply solved using an integer optimization that in the simplest form, can be performed using function evaluation.

Applying similar simulation parameters as the first example, the Q-function in EM is shown in Figure 3.8. The initial values in EM are selected using the procedure mentioned in Section 3.3.2 The true change points are $t = [t_1, t_2, t_3] = [12, 27, 43]$ and the number of samples is $n = 50$. In Table 3.2, for various initial values, the maximum of Q-function at the last iteration (logarithm of Q-function) as well as the final estimates of shift points are given. As we can see, among the randomly selected initial values, $t = [20, 25, 47]$ is the initial value corresponding to largest value which has also converged to true points. The number of iteration for all initial values excluding row (1) is two. For initial values $t = [10, 25, 40]$, the number of iteration is three which shows the effect of selection of initial value on the number of iteration.

Monte Carlo simulation with 100 runs is also performed for this example. The results of overall power OP and $AVTI$ for three different magnitudes of bias are depicted in Table 3.3 using Bayesian method (BM) and EM. The results are consistent with those in previous example. For small sizes of bias, the power of EM is larger than the Bayesian method.

The test of improper values of hyperparameters is also performed. The hyperparameters are selected as: $\mu_0^0 = [10; 10; 10]^T$, $\mu_1^0 = [12; 12; 12]^T$, $\mu_2^0 = [13; 13; 14]^T$ and

$\mu_1^0 = [25; 25; 35]^T$. EM is capable of estimating the true shift points while Bayesian method cannot correctly identify them.

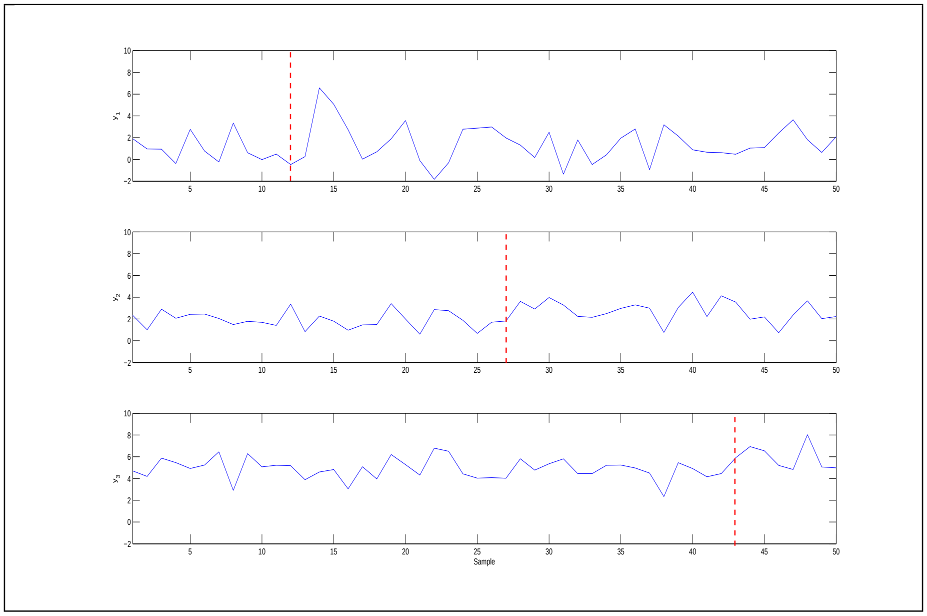

Figure 3.6: Three Measurement $y = [y_1, y_2, y_3]$ with Different Change Points

Table 3.2: Initialization of EM in Multiple Change Points Problem. True Change points are $t = [12, 27, 43]$

Row	t_1	t_2	t_3	Q_{max}	Final Solution
1	10	25	40	-166.61	$t = [11, 27, 43]$
2	10	20	35	-155.01	$t = [19, 27, 43]$
3	11	15	30	-146.25	$t = [12, 14, 27]$
4	13	28	45	-136.86	$t = [15, 24, 43]$
5	15	30	46	-134.31	$t = [12, 27, 43]$
6	20	25	47	-133.104	$t = [12, 27, 43]$
7	30	35	48	-160.92	$t = [29, 32, 44]$
8	15	20	38	-159.69	$t = [11, 26, 42]$

The next case study is selected to investigate the problem of change point detection in a chemical engineering example.

3.4.4 Continuous Stirred Tank Reactor (CSTR)

To further test the change point determination in correlated multivariate data, a chemical process is selected. The irreversible exothermic reaction $A \rightarrow B$ occurs inside a constant-volume reactor cooled by a single coolant stream. The system dynamics as shown in [39]

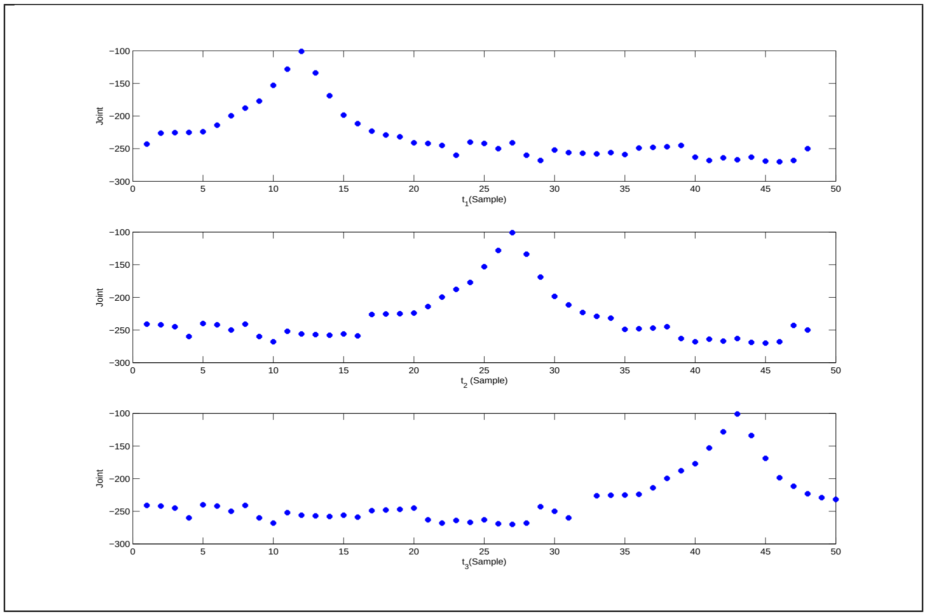

Figure 3.7: Bayesian Marginalised Probability Density (Logarithmic Scale) for Multiple Change Points

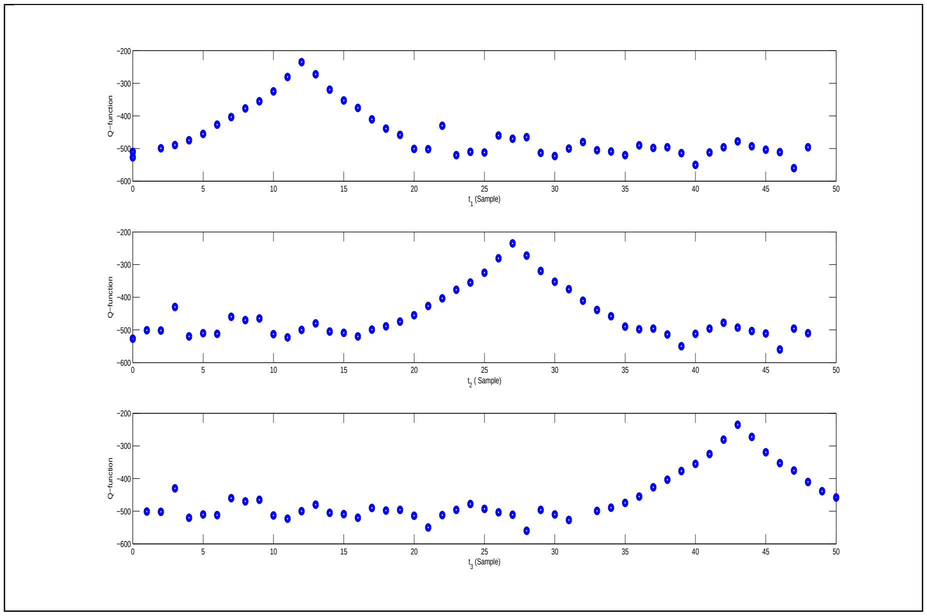

Figure 3.8: Q-function (logarithmic Scale) for Multiple Change Points

Table 3.3: Performance Comparison of Bayesian Method (BM) and EM for Different biases ($n = 50$, $t = [12, 27, 43]$, $\mu_0^0 = [2, 2, 2]^T$, $\mu_1^0 = [2, 2, 2]^T$, $\mu_2^0 = [3, 3, 3]^T$, $\mu_3^0 = [2, 2, 2]^T$)

	$Bias$	OP	$AVTI$
BM	$0.5\sigma_i$	0.33	0.67
	$1\sigma_i$	0.42	0.58
	$1.5\sigma_i$	0.56	0.44
EM	$0.5\sigma_i$	0.39	0.61
	$1\sigma_i$	0.53	0.47
	$1.5\sigma_i$	0.58	0.42

can be written as

$$\frac{C_A(t)}{dt} = \frac{q(t)}{V}(C_{A0}(t) - C_A(t)) - k_0 C_A(t) exp(-\frac{E}{RT(t)}) \tag{3.47}$$

$$\begin{aligned} \frac{T(t)}{dt} &= \frac{q(t)}{V}(T_0(t) - T(t)) + \frac{\Delta H k_0 C_A(t)}{\rho C_p} exp(-\frac{E}{RT(t)}) + \\ & \frac{\rho_c C_{pc}}{\rho C_p V} q_c(t)\{1 - exp(\frac{-hA}{q_c(t)\rho C_p})\}(T_{c0} - T(t)) \end{aligned} \tag{3.48}$$

The parameter definition and nominal values for CSTR are illustrated in Table 3.4. The states of the system are C_A (concentration of component A) and T (the reactor temperature). The outputs are the same as states but with measurement noise. The measurement noise added to each output is of Gaussian distribution with mean zero and standard deviation 10% of the states. The input is coolant flow rate, q_c, which determines the operating points varying within the interval [97 109]. By changing the input, the operating point changes and this affects the outputs accordingly. In this study, initially a constant input $q_c = 97$ is fed to the system so the system operates under the steady state condition.

In order to introduce a change to the measurement, at a certain time instant, the system input changes from one operating point to another driving the outputs to a new condition.

The histograms of data before (with white colour) and after (with grey colour) change point are shown in Figure 3.9. The mean values of outputs before and after the change are close to each other. The time trend plots of measurements are also shown in Figure 3.10. The collected data are down sampled so that the dynamics of process or transient response can be ignored and only steady state conditions are considered. The input is changed at time $t = 159$ with the magnitude of change being 2 L/min. When this occurs, the outputs start to change after a short delay. EM and Bayesian methods are applied to determine the time points at which the outputs have changed.

The parameters of EM and BM are: $\mu_1^0 = [0.5; 410]$, $\mu_2^0 = [0.5; 420]$, $\Sigma_{01} = \Sigma_{02} = I_{2\times2}$. In the case of EM, the number of iterations to convergence is 2 corresponding to the initial value $m = 80$. The results of Q-function with respect to time, for the two iterations of EM, as well as Bayesian joint probability are illustrated in Figure 3.11. In both cases, y-axis is in logarithmic scale.

Table 3.4: Model Parameters for the CSTR

Parameters	Nominal Values
production concentration of component A, C_A	output 1
temperature of the reactor, T	output 2
feed concentration of component A, C_{A0}	$1mol/L$
feed temperature, T_0	350.0 K
specific heats,C_p, C_{pc}	1 cal/(gk)
liquid density, ρ, ρ_c	$1 \times 10^3 g/L$
heat of reaction, $-\Delta H$	$-2 \times 10^5 cal/mol$
activation energy term,E/R	$1 \times 10^4 K$
reaction rate constant, k_0	$7.2 \times 10^{10} min^{-1}$
heat transfer term, hA	$7 \times 10^5 cal/(minK)$
reactor volume, V	$100L$
inlet coolant temperature, T_{c0}	$350K$
process flow rate,q	$100L/min$
coolant flow rate,q_c	input

As shown in Figure 3.11, in the second iteration, the increase in Q-function is apparent. From this, one can see that both methods show successful detection of change point at $t = 162$.

The effects of improper selection of priors are also investigated. If hyperparameters are $\mu_1^0 = [0.8; 400]$, $\mu_2^0 = [0.8; 410]$, then Bayesian method has false detection while EM is capable of detecting the change point. Using EM, depending on initial values, the Q-function is different. For some initial values, EM may exhibit some sorts of non-convergent behavior.

3.5 Experimental Evaluation: Hybrid Tank System

In order to evaluate the performance of change point detection using the proposed solution, the method is tested on real data obtained from experimental studies of a hybrid tank system. The schematic of hybrid tank is shown in Figure 3.12. This system was considered in the literature [40] for hybrid modelling, fault detection and reconfiguration. It consists of three connected tanks, six on/off valves and two pumps. The valves can be manipulated to change the flow rate to or out of these tanks. By opening or closing these valves, the system dynamics changes accordingly.

There are two cascade loops for the left and right tanks which control the levels of these tanks by manipulating the set point of flow controllers. Two proportional controllers are designed to maintain the level inside the left and right tank approximately at 75% percent. At first, the valves $V1$, $V2$, $V3$, $V4$ are closed and $V5$ to $V9$ are open.

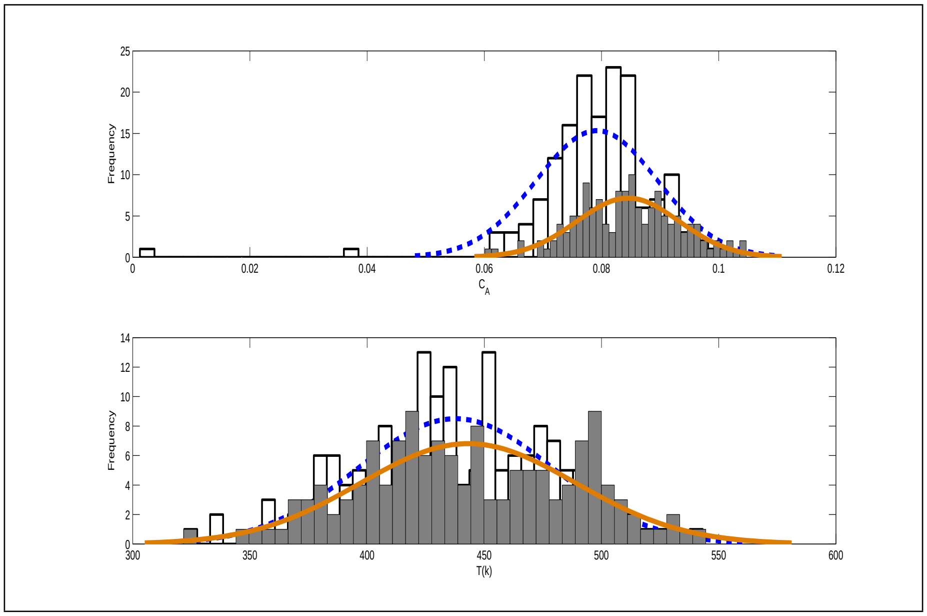

Figure 3.9: Histogram of Product Concentration (C_A) and Reactor Temperature (T) Before (white) and After Change Point(Grey)

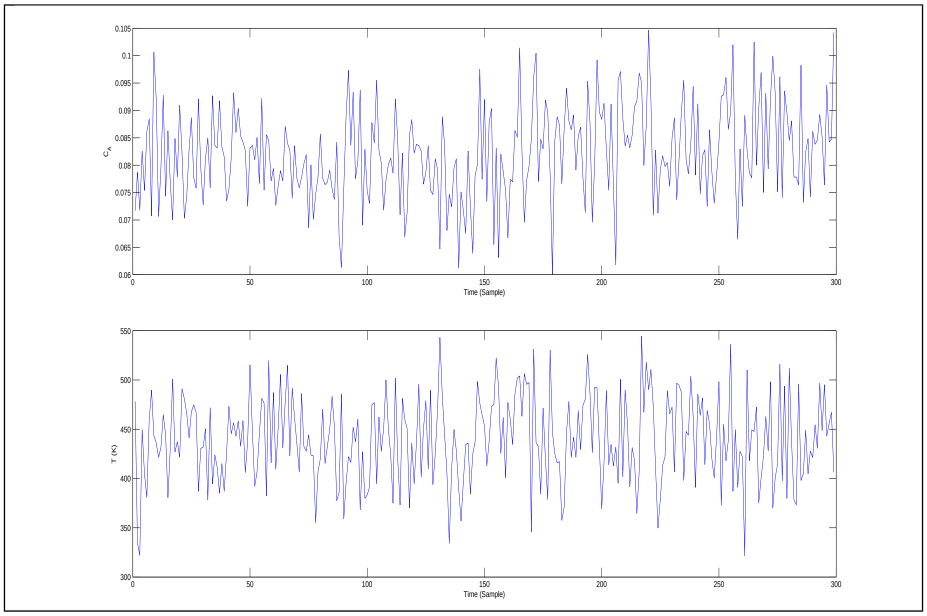

Figure 3.10: Product Concentration (C_A) and Reactor Temperature (T)

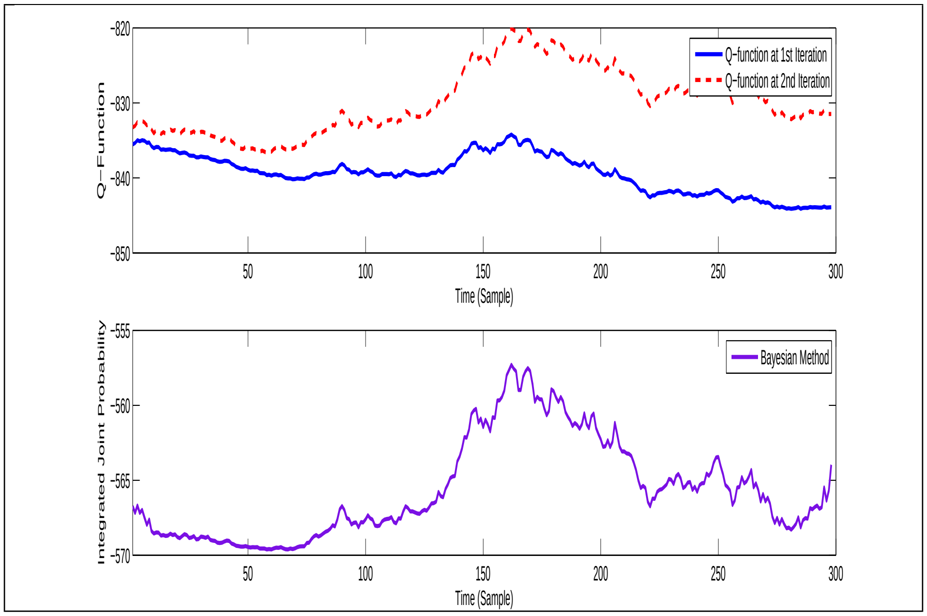

Figure 3.11: Q-function of EM (Upper Plot) and Integrated Joint Probability of Bayesian (Lower Plot) for CSTR System

Three changes are considered in this experiment. As the level in both left and right tanks reach 75%, the valves $V1$ and $V3$ are open imposing a change in levels of the tanks.

In order to introduce the second change, the valves $V2$ and $V4$ are open resulting in a second change in the tank levels. Having reached the second steady state, the valves $V1$ and $V3$ are closed leading to the third change. The left and right levels as well as the valves positions are shown in Figure 3.13. Note that logic value of 1 indicates close position and 0 indicates open position. First, by opening the valves $V1$ and $V3$, the level of left tank starts to decrease but the right tank does not experience a significant change. However, opening $V2$ and $V4$ in Figure 3.13 causes both levels to change. The magnitude of level change is more significant for the right tank than the left tank. Finally, by closing $V1$ and $V3$, both tanks experience a change in levels.

The hyperparameters selected for prior distributions are $\mu_1^0 = [70; 70]$, $\mu_2^0 = [71; 72]$, $\mu_3^0 = [71; 72]$, $\mu_4^0 = [71; 72]$, $\Sigma_{01} = \Sigma_{02} = \Sigma_{03} = \Sigma_{04} = I_{2\times2}$. Q-function values with respect to three variables t_1, t_2 and t_3 are illustrated in Figure 3.14. In order to evaluate the values of Q-function, various combinations of time instants are evaluated leading to high-dimensional matrix of Q. This dimension increases depending on the number of samples.

The results verify successful detection of the time instants at which the system dynamics is subject to change. Using such parameters in Bayesian method also leads to successful detection of change points.

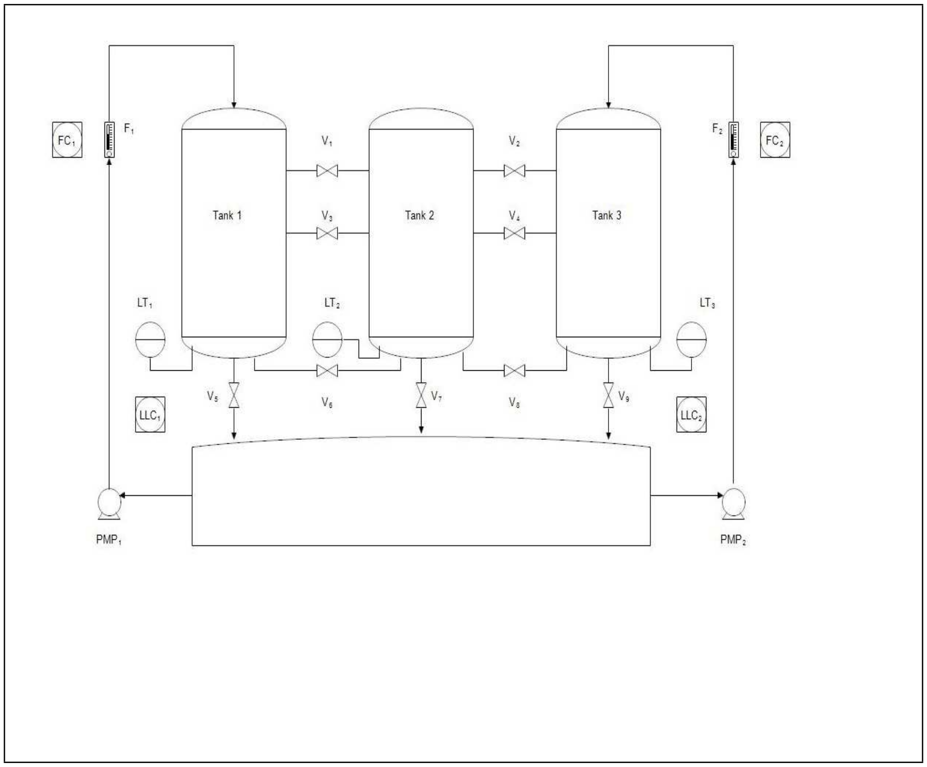

Figure 3.12: Schematic of Hybrid Tank

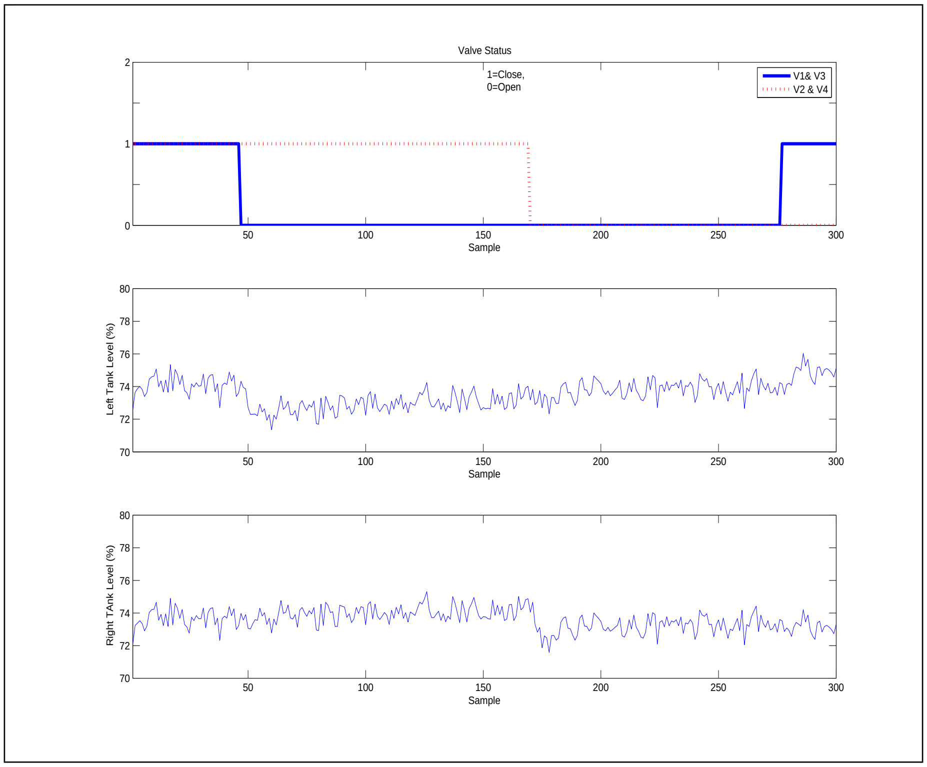

Figure 3.13: Valve Status(Upper Plot), left Tank Level(Middle Plot) and Right Tank Level(Lower Plot)

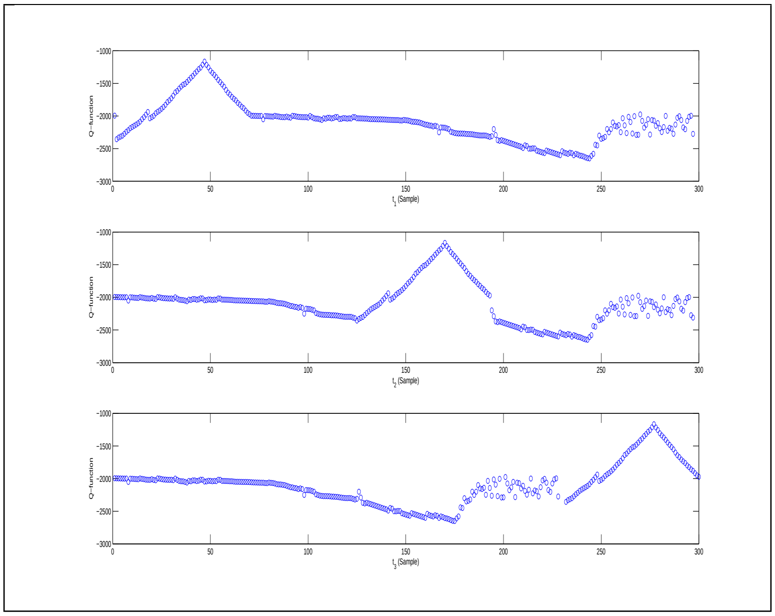

Figure 3.14: $Q-$function of EM for Hybrid System

3.6 Conclusion

In this chapter, Expectation Maximization (EM) is applied to the change point detection problem. The analytical solution of EM is derived for multivariate data for both single and multiple change points. In addition, a closed form of posterior probability is also derived for the Bayesian approach and MAP is used to infer the parameters. The performance of Bayesian and EM is compared for various scenarios of priors. It is shown that Bayesian inference is sensitive to the selection of priors but EM may show sensitivity to initial values for the iteration. The performance of proposed algorithms is evaluated using several examples including a CSTR process, one experimental case study and two simulated examples. The results show satisfactory performance of both Bayesian and EM methods. In the case of small changes and improper priors, however, EM demonstrates better performance. On the other hand, using EM approach, the challenges in selection of proper initial values and hence tackling the problem of non-convergent behaviour or local optimum can be regarded as its drawback. Besides, due to several numbers of iterations required in EM, it requires more computation compared with Bayesian approach.

In next chapter, the change point detection problem is more extended to unknown covariance and changing covariance.

Chapter 4

Multivariate Change Detection in the Presence of Unknown and Changing Covariance

4.1 Introduction

In [41], the change point problem was solved for mean shift detection with known and constant covariance. In real data, however, the covariance of data is often unknown and varying. One way to deal with this problem is to estimate the covariance from the data. This approach, in the presence of several changes in the mean of data may not work efficiently or may not work at all without knowing the change instances. An alternative way to handle this is through incorporation of prior knowledge in terms of prior distributions for this unknown parameter. In this chapter, a Bayesian data analysis in the context of parameter estimation is reviewed first, followed by solving the change point detection problem in the presence of unknown covariance through EM algorithm. Then, the problem is extended to a more general framework where both mean and covariance of data, as unknown parameters, can change simultaneously.

As mentioned before, EM has a flexible framework through which several problems can be tackled. For instance, EM handles improper priors more effectively than the Bayesian method. This is one of the main advantages of using EM as noted earlier while it may be computationally heavier. In previous chapter, this advantage was demonstrated through several examples. In this chapter, the focus is on derivation of EM solution, as one of the most efficient methods in solving change point detection problem.

4.2 Preliminary

Assume that in a sequence of independent and identically distributed (i.i.d) data, both mean and covariance are unknown and the objective is to estimate the mean and covariance of the data. A commonly used prior distribution for covariance matrix is Inverse Wishart (IW) distribution as

$$P(\Sigma|\nu_0, \Psi_0) = \frac{|\Psi_0|^{\frac{\nu_0}{2}} |\Sigma|^{-\frac{\nu_0+p+1}{2}} exp(\frac{-tr(\Psi_0\Sigma^{-1})}{2})}{2^{\frac{\nu_0 p}{2}} Z(\nu_0, p)} \tag{4.1}$$

where

$$Z(n, p) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma(\frac{n+1-i}{2}) \tag{4.2}$$

and ν_0 is degrees of freedom. Note that one must have $\nu > p - 1$. Ψ_0 is $p \times p$ positive definite matrix. Also, assume a Gaussian prior distribution for the mean [21] as

$$P(\mu|\mu_0, \frac{\Sigma}{k_0}) = (2\pi)^{-p/2}|\Sigma|^{-1/2} exp\{-\frac{k_0}{2}(\mu - \mu_0)^T \Sigma^{-1}(\mu - \mu_0)\} \tag{4.3}$$

where k_0 and μ_0 are the hyperparameters. From (4.3), we can see that the prior distribution of mean depends on Σ. In this definition, a hierarchial prior structure is used to take into account the correlated priors which provides a more general framework. In the context of Gaussian distribution assumption for the data, the likelihood for multivariate data can be expressed as

$$P(y|\mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2} \prod_{i=1}^{n} exp\{-\frac{1}{2}(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)\} \tag{4.4}$$

$$= (2\pi)^{-np/2}|\Sigma|^{-n/2} exp\{-\frac{1}{2}\sum_{i=1}^{n}[(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)]\} \tag{4.5}$$

$$= (2\pi)^{-np/2}|\Sigma|^{-n/2} exp\{-\frac{1}{2}[(n-1)s^2 + n(\bar{y} - \mu)^T \Sigma^{-1}\bar{y} - \mu)]\} \tag{4.6}$$

where $s = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^T \Sigma^{-1}(y_i - \bar{y})$. The product of priors and the likelihood results in joint posterior as

$$\begin{aligned}
&P(\mu, \Sigma|y) \propto P(y|\mu, \Sigma) \times P(\mu|\mu_0, \frac{\Sigma}{k_0}) \times P(\Sigma|\nu_0, \Psi_0) \\
&\propto (2\pi)^{-np/2}|\Sigma|^{-n/2} exp\{-\frac{1}{2}\sum_{i=1}^{n}[(y_i - \mu)^T \Sigma^{-1}(y_i - \mu)]\}\times \\
&(2\pi)^{-p/2}|\Sigma|^{-1/2} exp\{-\frac{k_0}{2}(\mu - \mu_0)^T \Sigma^{-1}(\mu - \mu_0)\} \frac{|\Psi_0|^{\frac{\nu_0}{2}} |\Sigma|^{-\frac{\nu_0+p+1}{2}} exp(\frac{-tr(\Psi_0\Sigma^{-1})}{2})}{2^{\frac{\nu_0 p}{2}} Z(\nu_0, p)} \\
&\propto |\Psi_0|^{\frac{\nu_0}{2}}|\Sigma|^{-1/2}|\Sigma|^{-\frac{n+\nu_0+p+1}{2}} exp[-\frac{1}{2}\{tr(\Psi_0\Sigma^{-1}) + (n-1)s^2 + n(\bar{y} - \mu)^T \Sigma^{-1}\bar{y} - \mu) + \\
&k_0(\mu - \mu_0)^T \Sigma^{-1}(\mu - \mu_0)\}]
\end{aligned} \tag{4.7}$$

After some algebraic simplification such as completing the square, the joint posterior is derived as

$$\begin{aligned}
&P(\mu, \Sigma|y) \propto |\Psi_0|^{\frac{\nu_0}{2}}|\Sigma|^{-1/2}|\Sigma|^{-\frac{n+\nu_0+p+1}{2}} exp[-\frac{1}{2}\{tr(\Psi_0\Sigma^{-1}) + (n-1)s^2 + \\
&\frac{k_0 n}{k_0 + n}(\bar{y} - \mu)^T \Sigma^{-1}(\bar{y} - \mu)\}] \times exp[-\frac{1}{2}\{(k_0 + n)(\mu - \mu_n)^T \Sigma^{-1}(\mu - \mu_n)\}]
\end{aligned} \tag{4.8}$$

The following matrix properties have been used in derivation of (4.8):

$$x^T\Sigma^{-1}x = tr(x^T\Sigma^{-1}x) \tag{4.9}$$

$$tr(A) + tr(B) = tr(A + B) \quad \text{for two square matrices A, B} \tag{4.10}$$

$$tr(AB) = tr(BA) \quad \text{for two square matrices A, B} \tag{4.11}$$

$$tr(xx^T) = tr(x^Tx) \tag{4.12}$$

$$tr(x^T\sigma^{-1}x) = tr(\sigma^{-1}xx^T) = tr(xx^T\sigma^{-1}) \tag{4.13}$$

The joint posterior of (4.8) is Normal-Inverse Wishart (NIW) distribution expressed in the form of $N(\mu|\mu_n, (k_0 + n)^{-1}\Sigma) \times IW(\Sigma^{-1}|\nu_n, \Psi_n)$ as

$$\begin{aligned} P(\mu, \Sigma|y) &\propto |\Psi_n|^{\frac{\nu_n}{2}}|\Sigma|^{-\frac{\nu_n+p+1}{2}}exp[-\frac{1}{2}tr(\Psi_n\Sigma^{-1})] \\ &\times |\Sigma|^{-1/2}exp[\frac{k_n}{2}(\mu - \mu_n)^T\Sigma^{-1}(\mu - \mu_n)] \end{aligned} \tag{4.14}$$

where $\nu_n = \nu_0 + n$, $k_n = k_0 + n$, $\mu_n = \frac{1}{k_0+n}(k_0\mu_0 + n\bar{y})$ and $\Psi_n = \Psi_0 + \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T + \frac{k_0 n}{k_0+n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T$. Thus, given the data and Σ, the marginal probability of μ can be written as

$$P(\mu|\Sigma, Y) = N(\mu|\mu_n, (k_n)^{-1}\Sigma) \propto exp[\frac{k_n}{2}(\mu - \mu_n)\Sigma^{-1}(\mu - \mu_n)] \tag{4.15}$$

and the marginal probability of Σ is derived as

$$P(\Sigma|Y) = IW(\Sigma^{-1}|\nu_n, \Psi_n) \tag{4.16}$$

4.3 Expectation Maximization (EM) Algorithm

EM algorithm is based on maximum likelihood estimation principle. This method was first introduced by [26] in 1977. Numerous applications in various areas can be found in literature based on this method such as in machine learning, computer vision, medical imaging, mixture models, speech recognition, etc. It is effective especially when it is not easy to find the maximum of $P(parameter|data)$ directly [21] [22]. This algorithm consists of iteration between two steps: expectation-step or E-step and maximization-step or M-step. In other words, EM can be formulated as (1) finding the conditional expectation with respect to missing variables given the data and the current estimate of the parameter and (2) maximizing the expectation derived in the previous step to estimate the parameters [21]. Convergence of EM algorithm is guaranteed because at each iteration, the likelihood function is non-decreasing [22]. In this framework, E-step can be formulated as

$$Q(\theta|\theta^{(k)}) = E_{Z|D,\theta^{(k)}}\{P(D, Z|\theta)\} \tag{4.17}$$

where Z is the missing data or hidden variable, D is the observed data and θ is the parameter to be estimated. $\theta^{(k)}$ is the current estimate of parameter. In essence, in E-step, missing data are marginalized given the observation and the current estimate of unknown parameters. In other words, in E-step, a conditional expectation is derived where, depending on the types of missing variables, continues or discrete variables, an integration or summation is computed. In the M-step, the new parameter, $\theta^{(k+1)}$, is chosen so that it maximizes $Q(\theta|\theta^{(k)})$; that is,

$$Q(\theta^{(k+1)}|\theta^{(k)}) \geq Q(\theta|\theta^{(k)}), \forall\theta \tag{4.18}$$

M-step can be expressed as

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta|\theta^{(k)}) \tag{4.19}$$

Starting EM with initial values for the parameters, E-step and M-step are repeated until a suitable stopping rule criterion is satisfied [22].

4.4 EM in Change Detection with Unknown Mean and Covariance

In the following, the problem is formulated for occurrence of single change in the mean of multivariate data.

4.4.1 Problem Formulation for Single Change Point Detection

Consider that n observations from p variables form a $p \times n$ matrix as

$$D = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} & y_{1(m+1)} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2m} & y_{2(m+1)} & \cdots & y_{2n} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ y_{p1} & y_{p2} & \cdots & y_{pm} & y_{p(m+1)} & \cdots & y_{pn} \end{pmatrix} = (Y_1, Y_2, ..., Y_m, ..., Y_n)$$

$Y_1, ..., Y_m, ..., Y_n$ are measurements of p variables from time instant 1 to n. Assume that at the sampling instant m, a change occurs resulting in a shift in the mean vector. As a result, the whole data are split into two segments operating at two different means, μ_1 and μ_2, respectively with the same but unknown covariance matrix Σ. Assume that the mean of data before change is μ_1 and after change is μ_2. As discussed in Section 4.2, the prior distributions for unknown parameters can be selected as

$$P(\mu_1|\mu_1^0, \frac{\Sigma}{k_{01}}) = (2\pi)^{-p/2}|\Sigma|^{-1/2} exp\{-\frac{k_{01}}{2}(\mu_1 - \mu_1^0)^T\Sigma^{-1}(\mu_1 - \mu_1^0)\}$$

$$P(\Sigma|\nu_0, \Psi_0) = \frac{|\Psi_0|^{\frac{\nu_0}{2}}|\Sigma|^{-\frac{\nu_0+p+1}{2}} exp(\frac{-tr(\Psi_0\Sigma^{-1})}{2})}{2^{\frac{\nu_0 p}{2}} Z(\nu_0, p)} \tag{4.20}$$

Under the EM framework, μ_1, μ_2, Σ are treated as hidden variables or missing data. The unknown parameter is the time instant at which the change occurs. Hence, E-step can be formulated as

$$Q(m|m^{(k)}) = E_{\mu_1,\mu_2,\Sigma|Y,m^{(k)}}\{P(Y, \mu_1, \mu_2, \Sigma|m)\} \tag{4.21}$$

$$= \int\int\int P(Y, \mu_1, \mu_2, \Sigma|m)P(\mu_1, \mu_2, \Sigma|Y, m^{(k)})d\mu_1 d\mu_2 d\Sigma \tag{4.22}$$

The first term in integrator of (4.22) can be written as

$$\begin{aligned}P(Y,&\mu_1,\mu_2,\Sigma|m) = P(Y|\mu_1,\mu_2,\Sigma,m)P(\mu_1|\Sigma,m)P(\mu_2|\Sigma,m)P(\Sigma|m)\\ &= P(Y_{1:m}|\mu_1,\Sigma,m)P(Y_{m+1:n}|\mu_2,\Sigma,m)P(\mu_1|\Sigma)P(\mu_2|\Sigma)P(\Sigma)\\ &= \prod_{i=1}^{m}\mathcal{N}(Y_i|\mu_1,\Sigma)\prod_{i=m+1}^{n}\mathcal{N}(Y_i|\mu_2,\Sigma)\mathcal{N}(\mu_1|\mu_1^0,(k_{01})^{-1}\Sigma)\mathcal{N}(\mu_2|\mu_2^0,(k_{02})^{-1}\Sigma)IW(\nu_0,\Psi_0)\end{aligned} \tag{4.23}$$

Note that it is reasonable to assume that m is dependent of μ_1,μ_2 and Σ and this fact has been used in the derivation of (4.23). The second term in (4.22) can be expressed as

$$P(\mu_1,\mu_2,\Sigma|Y,m^{(k)}) = P(\mu_1|\Sigma,Y,m^{(k)})P(\mu_2|\Sigma,Y,m^{(k)})P(\Sigma|Y,m^{(k)}) \tag{4.24}$$

In derivation of (4.24), we have used the fact that given the change point $m^{(k)}$, μ_1 and μ_2 are conditionally independent. According to Bayesian analysis as in the previous section, $P(\mu_1|\Sigma,Y,m^k)$ and $P(\mu_2|\Sigma,Y,m^k)$ can be derived similarly as in the derivation of (4.15). Thus, one can write

$$\begin{aligned}P(\mu_1|\Sigma,Y,m^{(k)}) &\sim \mathcal{N}(\mu_1|\mu_{1mk},(k_{1mk})^{-1}\Sigma)\\ P(\mu_2|\Sigma,Y,m^{(k)}) &\sim \mathcal{N}(\mu_2|\mu_{2mk},(k_{2mk})^{-1}\Sigma)\end{aligned} \tag{4.25}$$

and since Σ is the the same for all the n observations, from (4.16) we have

$$P(\Sigma|Y,m^{(k)}) = IW(\nu_n,\Psi_n) \tag{4.26}$$

where the parameters are defined as

$$\nu_n = \nu_0 + n \tag{4.27}$$

$$k_{1mk} = k_{01} + m^{(k)} \tag{4.28}$$

$$k_{2mk} = k_{02} + n - m^{(k)} \tag{4.29}$$

$$\mu_{1mk} = \frac{1}{k_{01}+m^{(k)}}(k_{01}\mu_1^0 + m^{(k)}\bar{y}_1) \tag{4.30}$$

$$\mu_{2mk} = \frac{1}{k_{02}+n-m^{(k)}}(k_{02}\mu_2^0 + (n-m^{(k)})\bar{y}_2) \tag{4.31}$$

$$\begin{aligned}\Psi_n = \Psi_0 &+ \sum_{i=1}^{m^{(k)}}(y_i-\bar{y}_1)(y_i-\bar{y}_1)^T + \sum_{i=m^{(k)}+1}^{n}(y_i-\bar{y}_2)(y_i-\bar{y}_2)^T +\\ &\frac{k_{01}m^{(k)}}{k_{01}+m^{(k)}}(\bar{y}_1-\mu_1^0)(\bar{y}_1-\mu_1^0)^T + \frac{k_{02}(n-m^{(k)})}{k_{02}+n-m^{(k)}}(\bar{y}_2-\mu_2^0)(\bar{y}_2-\mu_2^0)^T\end{aligned} \tag{4.32}$$

Thus, (4.24) yields

$$P(\mu_1,\mu_2,\Sigma|Y,m^{(k)}) = \mathcal{N}(\mu_1|\mu_{1mk},(k_{1mk})^{-1}\Sigma)\mathcal{N}(\mu_2|\mu_{2mk},(k_{2mk})^{-1}\Sigma)IW(\nu_n,\Psi_n) \tag{4.33}$$

By substituting (4.25) and (4.26) into (4.22), we have

$$Q(m|m^{(k)}) = \int_{\mu_1}\int_{\mu_2}\int_{\Sigma}\prod_{i=1}^{m}\mathcal{N}(Y_i|\mu_1,\Sigma)\prod_{i=m+1}^{n}\mathcal{N}(Y_i|\mu_2,\Sigma)\mathcal{N}(\mu_1|\mu_1^0,(k_{01})^{-1}\Sigma)$$
$$\mathcal{N}(\mu_2|\mu_2^0,(k_{02})^{-1}\Sigma)IW(\nu_0,\Psi_0)\mathcal{N}(\mu_1|\mu_{1mk},(k_{1mk})^{-1}\Sigma)\times$$
$$\mathcal{N}(\mu_2|\mu_{2mk},(k_{2mk})^{-1}\Sigma)IW(\nu_n,\Psi_n)d\mu_1 d\mu_2 d\Sigma \tag{4.34}$$

This integration can be simplified as

$$Q(m|m^{(k)}) = \int_{\Sigma}\int_{\mu_1}[\prod_{i=1}^{m}\mathcal{N}(Y_i|\mu_1,\Sigma)\mathcal{N}(\mu_1|\mu_1^0,(k_{01})^{-1}\Sigma)\mathcal{N}(\mu_1|\mu_{1mk},(k_{1mk})^{-1}\Sigma)]d\mu_1$$
$$\times\int_{\mu_2}[\prod_{i=m+1}^{n}\mathcal{N}(Y_i|\mu_2,\Sigma)\mathcal{N}(\mu_2|\mu_2^0,(k_{02})^{-1}\Sigma)\mathcal{N}(\mu_2|\mu_{2mk},(k_{2mk})^{-1}\Sigma)]d\mu_2$$
$$\times[IW(\nu_0,\Psi_0)IW(\nu_n,\Psi_n)]d\Sigma \tag{4.35}$$

The first inner integration in (4.35) is with respect to μ_1. Denote the first inner integration in (4.35) as Z_1. We have

$$Z_1 = \int_{\mu_1}[\prod_{i=1}^{m}\mathcal{N}(Y_i|\mu_1,\Sigma)\mathcal{N}(\mu_1|\mu_1^0,(k_{01})^{-1}\Sigma)\mathcal{N}(\mu_1|\mu_{1mk},(k_{1mk})^{-1}\Sigma)]d\mu_1$$
$$= (2\pi)^{-mp/2}|\Sigma|^{-m/2}(2\pi)^{-p/2}|(k_{01})^{-1}\Sigma|^{-1/2}(2\pi)^{-p/2}|(k_{1mk})^{-1}\Sigma|^{-1/2}\times$$
$$\int_{\mu_1}exp\{-\frac{1}{2}\{\sum_{i=1}^{m}[(y_i-\mu_1)^T\Sigma^{-1}(y_i-\mu_1)]+k_{01}(\mu_1-\mu_1^0)^T\Sigma^{-1}(\mu_1-\mu_1^0)\}+$$
$$k_{1mk}(\mu_1-\mu_{1mk})^T\Sigma^{-1}(\mu_1-\mu_{1mk})\}\}d\mu_1 \tag{4.36}$$

After some algebraic simplification such as completing the square, one can write

$$Z_1 = \int_{\mu_1}...d\mu_1 = (2\pi)^{-(m+1)p/2}(k_{1mk})^{p/2}(k_{01})^{p/2}|\Sigma|^{-m/2-1}|\Delta_m|^{1/2}$$
$$\times\, exp\{-\frac{1}{2}[\sum_{i=1}^{m}(y_i^T\Sigma^{-1}y_i)+k_{01}{\mu_1^0}^T\Sigma^{-1}\mu_1^0+k_{1mk}\mu_{1mk}^T\Sigma^{-1}\mu_{1mk}-B^TA^{-1}B]\} \tag{4.37}$$

where

$$A = (k_{01}+m+k_{1mk})\Sigma^{-1} \tag{4.38}$$
$$\Delta_m = A^{-1} \tag{4.39}$$
$$B = \Sigma^{-1}(\sum_{i=1}^{m}y_i - k_{01}\mu_1^0 - k_{1mk}\mu_{1mk}) \tag{4.40}$$

Denote the second inner integration in (4.35) as Z_2. Following the same approach with respect to μ_2 as derivation of Z_1 yields

$$Z_2 = \int_{\mu_2}...d\mu_2 = (2\pi)^{-(n-m+1)p/2}(k_{2mk})^{p/2}(k_{02})^{p/2}|\Sigma|^{-(n-m)/2-1}|\Omega_m|^{1/2}$$
$$exp\{-\frac{1}{2}[\sum_{i=m+1}^{n}(y_i^T\Sigma^{-1}y_i)+k_{02}{\mu_2^0}^T\Sigma^{-1}\mu_2^0+k_{2mk}\mu_{2mk}^T\Sigma^{-1}\mu_{2mk}-D^TC^{-1}D]\} \tag{4.41}$$

where

$$C = (k_{02} + n - m + k_{2mk})\Sigma^{-1} \tag{4.42}$$

$$\Omega_m = C^{-1} \tag{4.43}$$

$$D = \Sigma^{-1}(\sum_{i=m+1}^{n} y_i - k_{02}\mu_2^0 - k_{2mk}\mu_{2mk}) \tag{4.44}$$

Multiplying (4.37) by (4.41) results in

$$\begin{aligned}
Z_1 \times Z_2 = &(2\pi)^{-(n+2)p/2}(k_{1mk})^{p/2}(k_{2mk})^{p/2}(k_{01})^{p/2}(k_{02})^{p/2}|\Sigma|^{-n/2-2}|\Delta_m|^{1/2}|\Omega_m|^{1/2}\times \\
&exp\{-\frac{1}{2}[\sum_{i=1}^{m}(y_i^T\Sigma^{-1}y_i) + k_{01}\mu_1^{0^T}\Sigma^{-1}\mu_1^0 + k_{1mk}\mu_{1mk}^T\Sigma^{-1}\mu_{1mk} - B^TA^{-1}B]\}\times \\
&exp\{-\frac{1}{2}[\sum_{i=m+1}^{n}(y_i^T\Sigma^{-1}y_i) + k_{02}\mu_2^{0^T}\Sigma^{-1}\mu_2^0 + k_{2mk}\mu_{2mk}^T\Sigma^{-1}\mu_{2mk} - D^TC^{-1}D]\} \\
&= (2\pi)^{-(n+2)p/2}(k_{1mk})^{p/2}(k_{2mk})^{p/2}(k_{01})^{p/2}(k_{02})^{p/2}(k_{01} + m + k_{1mk})^{-p/2}\times \\
&(k_{02} + n - m + k_{2mk})^{-p/2}|\Sigma|^{-n/2-1} \times exp\{-\frac{1}{2}tr[Q_0\Sigma^{-1}]\}
\end{aligned} \tag{4.45}$$

where

$$\begin{aligned}
Q_0 = &\sum_{i=1}^{m}(y_iy_i^T) + k_{01}\mu_1^0\mu_1^{0^T} + k_{1mk}\mu_{1mk}\mu_{1mk}{}^T + \sum_{i=m+1}^{n}(y_iy_i^T) + k_{02}\mu_2^0\mu_2^{0^T} + k_{2mk}\mu_{2mk}\mu_{2mk}{}^T + \\
&(\sum_{i=1}^{m} y_i - k_{01}\mu_1^0 - k_{1mk}\mu_{1mk})(\sum_{i=1}^{m} y_i - k_{01}\mu_1^0 - k_{1mk}\mu_{1mk})^T + \\
&(\sum_{i=m+1}^{n} y_i - k_{02}\mu_2^0 - k_{2mk}\mu_{2mk})(\sum_{i=m+1}^{n} y_i - k_{02}\mu_2^0 - k_{2mk}\mu_{2mk})^T
\end{aligned} \tag{4.46}$$

In derivation of (4.45), the properties (4.9) to (4.13) have been used to simplify the expression. To integrate out Σ, (4.45) is substituted into (4.35). Consequently, we have

$$\begin{aligned}
Q(m|m^{(k)}) = &(2\pi)^{-(n+2)p/2}(k_{1mk})^{p/2}(k_{2mk})^{p/2}(k_{01})^{p/2}(k_{02})^{p/2}(k_{01} + m + k_{1mk})^{-p/2}\times \\
&(k_{02} + n - m + k_{2mk})^{-p/2}\int_{\Sigma}[|\Sigma|^{-n/2-1}exp(-\frac{1}{2}tr[Q_0\Sigma^{-1}])IW(\nu_0, \Psi_0)IW(\nu_n, \Psi_n)]d\Sigma \\
&= h(k)\int_{\Sigma}[|\Sigma|^{-n/2-1}exp(-\frac{1}{2}tr[Q_0\Sigma^{-1}])IW(\nu_0, \Psi_0)IW(\nu_n, \Psi_n)]d\Sigma
\end{aligned} \tag{4.47}$$

where

$$\begin{aligned}
h(k) = &(2\pi)^{-(n+2)p/2}(k_{1mk})^{p/2}(k_{2mk})^{p/2}(k_{01})^{p/2}(k_{02})^{p/2}(k_{01} + m + k_{1mk})^{-p/2}\times \\
&(k_{02} + n - m + k_{2mk})^{-p/2}
\end{aligned} \tag{4.48}$$

Thus, (4.47) can be expressed as

$$
\begin{aligned}
Q(m|m^{(k)}) &= h(k)\frac{|\Psi_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 p}{2}}Z(\nu_0,p)}\frac{|\Psi_n|^{\frac{\nu_n}{2}}}{2^{\frac{\nu_n p}{2}}Z(\nu_n,p)}\int_\Sigma [|\Sigma|^{-n/2-1}exp(-\frac{1}{2}tr(Q_0\Sigma^{-1}))|\Sigma|^{-\frac{\nu_0+p+1}{2}}\times \\
&\quad exp(-\frac{1}{2}tr(\Psi_0\Sigma^{-1}))|\Sigma|^{-\frac{\nu_n+p+1}{2}}exp(-\frac{1}{2}tr(\Psi_n\Sigma^{-1}))]d\Sigma \\
&= h(k)\frac{|\Psi_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 p}{2}}Z(\nu_0,p)}\frac{|\Psi_n|^{\frac{\nu_n}{2}}}{2^{\frac{\nu_n p}{2}}Z(\nu_n,p)}\int_\Sigma |\Sigma|^{-\frac{\nu_{nn}+p+1}{2}}exp(-\frac{1}{2}tr[\Psi_{nn}\Sigma^{-1}])d\Sigma \qquad (4.49)
\end{aligned}
$$

where $\Psi_{nn} = Q_0 + \Psi_0 + \Psi_n$ and $\nu_{nn} = n + \nu_0 + \nu_n + p + 3$. Finally, the closed form solution of (4.49) can be derived as

$$
\begin{aligned}
Q(m|m^{(k)}) &= h(k)\frac{|\Psi_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 p}{2}}Z(\nu_0,p)}\frac{|\Psi_n|^{\frac{\nu_n}{2}}}{2^{\frac{\nu_n p}{2}}Z(\nu_n,p)}\frac{2^{\frac{\nu_{nn}p}{2}}Z(\nu_{nn},p)}{|\Psi_{nn}|^{\frac{\nu_{nn}}{2}}} \\
&= h(k)2^{(\nu_{nn}-\nu_0-\nu_n)p/2}\frac{|\Psi_0|^{\frac{\nu_0}{2}}|\Psi_n|^{\frac{\nu_n}{2}}}{|\Psi_{nn}|^{\frac{\nu_{nn}}{2}}}\frac{Z(\nu_{nn},p)}{Z(\nu_0,p)Z(\nu_n,p)} \qquad (4.50)
\end{aligned}
$$

Having derived the closed form expression for Q-function with respect to hyperparameters and current estimate of the unknown parameter, m^k, the next step, M-step, can be written as

$$
m^{(k+1)} = \arg\max_m Q(m|m^{(k)}) \qquad (4.51)
$$

In (4.51) the maximization is performed with respect to m, which can be solved relatively easily by, for example, numerical evaluation of the Q-function over different integer value of m. The iteration between E-step and M-step continues until no change in Q-function is observed.

In the next section, the EM derivation for the case of single change is extended to multiple change detection problem.

4.4.2 Problem Formulation for Multiple Change Points Detection

In real industrial process data, the measurements may be subject to mean shifts at various time instants. In such cases, the single mean shift model with unknown covariance, derived in previous section, does not hold and one should develop a model based on multiple change points. The same formulation for multiple change points as [41] can be followed but the covariance is unknown here, which introduces a significant challenge. Assume that, the interval of length n is of interest and there are N mean shifts. The shift points are characterised by vector $t = [t_1, ..., t_N]$ and hence, the data are split into $N + 1$ segments; each segment has its own mean vector but the covariance is unknown and also $t_0 = 1$, $t_{N+1} = n$.

Under EM framework, the hidden variables are μ_i and Σ for $i = 0, 1, ..., N$. The vector $t = [t_1, ..., t_N]$ characterizes the parameters of the interest to be estimated. Therefore, E-step can be expressed as

$$
Q(t|t^{(k)}) = E_{\mu_0,\mu_1,...,\mu_N,\Sigma|Y,t^{(k)}}\{P(Y, \mu_0, \mu_1, ..., \mu_N, \Sigma|t)\} \qquad (4.52)
$$

Similar to previous section, $P(Y, \mu_0, \mu_1, ..., \mu_N, \Sigma|t)$ can be represented as

$$\begin{aligned} P(Y, \mu_0, \mu_1, ..., \mu_N, \Sigma|t) &= P(Y|\mu_0, \mu_1, ..., \mu_N, \Sigma, t)P(\mu_0|\Sigma, t)P(\mu_1|\Sigma, t)...P(\mu_N|\Sigma, m)P(\Sigma|t) \\ &= P(Y_{t_0:t_1}|\mu_0, \Sigma, t)P(Y_{t1:t_2}|\mu_1, \Sigma, t)...P(Y_{t_N:t_{N+1}}|\mu_N, \Sigma, t)\times \\ &P(\mu_0|\Sigma, t)P(\mu_1|\Sigma, t)...P(\mu_N|\Sigma, t)P(\Sigma|t) \end{aligned} \tag{4.53}$$

Similarly, the prior for mean vector and also the covariance can be expressed as

$$\begin{aligned} P(\mu_i|\Sigma) &= (2\pi)^{-p/2}|\Sigma|^{-1/2}exp\{-\frac{k_{0i}}{2}(\mu_i - \mu_i^0)^T\Sigma^{-1}(\mu_i - \mu_i^0)\} \quad i = 0, 1, ..., N \\ P(\Sigma|\nu_0, \Psi_0) &= \frac{|\Psi_0|^{\frac{\nu_0}{2}}|\Sigma|^{-\frac{\nu_0+p+1}{2}}exp(\frac{-tr(\Psi_0\Sigma^{-1})}{2})}{2^{\frac{\nu_0 p}{2}}Z(\nu_0, p)} \end{aligned} \tag{4.54}$$

On the other hand, based on the chain rule in probability, $P(\mu_0, \mu_1, ..., \mu_N, \Sigma|Y, t^{(k)})$ can be derived as

$$P(\mu_0, \mu_1, ..., \mu_N, \Sigma|Y, t^{(k)}) = P(\mu_0|\Sigma, Y, t^{(k)})P(\mu_1|\Sigma, Y, t^{(k)})...P(\mu_N|\Sigma, Y, t^{(k)})P(\Sigma|Y, t^{(k)}) \tag{4.55}$$

Again in the derivation of (4.55), we have used the fact that given the change point $t^{(k)}, \mu_i, i = 1, \ldots, N$ are conditionally independent. As noted before, given the data and Σ, the posterior for μ_i, $i = 0, 1, ..., N$ is of the following form:

$$P(\mu_0|\Sigma, Y, t^{(k)}) \sim \mathcal{N}(\mu_{0tk}, (k_{0tk})^{-1}\Sigma) \tag{4.56}$$

$$P(\mu_1|\Sigma, Y, t^{(k)}) \sim \mathcal{N}(\mu_{1tk}, (k_{1tk})^{-1}\Sigma) \tag{4.57}$$

.

.

.

$$P(\mu_N|\Sigma, Y, t^{(k)}) \sim \mathcal{N}(\mu_{Ntk}, (k_{Ntk})^{-1}\Sigma) \tag{4.58}$$

where

$$k_{itk} = k_{0i} + (t_{i+1}^{(k)} - t_i^{(k)}) \quad i = 0, 1, ..., N \tag{4.59}$$

$$\mu_{itk} = \frac{1}{k_{0i} + t_{i+1}^{(k)} - t_i^{(k)}}(k_{0i}\mu_i^0 + (t_{i+1}^{(k)} - t_i^{(k)})\bar{y}_{t_i^{(k)}:t_{i+1}^{(k)}}) \quad i = 0, 1, ..., N \tag{4.60}$$

$$\bar{y}_{t_i^{(k)}:t_{i+1}^{(k)}} = \frac{\sum_{i=t_i^{(k)}+1}^{t_{i+1}^{(k)}} y_i}{t_{i+1}^{(k)} - t_i^{(k)}} \quad i = 0, 1, ..., N \tag{4.61}$$

In this derivation, $\mu = [\mu_0, \mu_1, ..., \mu_N]$ is the mean with respect to various segments identified by t_0 to t_N. Since Σ is the same for all the n observations, we have

$$P(\Sigma|Y, t^{(k)}) = IW(\nu_n, \Psi_n) \tag{4.62}$$

where

$$\Psi_n = \Psi_0 + \sum_{i=t_0}^{i=t_1^{(k)}} (y_i - \bar{y}_0)(y_i - \bar{y}_0)^T + \sum_{i=t_1^{(k)}+1}^{i=t_2^{(k)}} (y_i - \bar{y}_1)(y_i - \bar{y}_1)^T + ...$$
$$+ \sum_{i=t_N^{(k)}}^{i=t_{N+1}^{(k)}} (y_i - \bar{y}_1)(y_i - \bar{y}_1)^T + \frac{k_{00}(t_1^{(k)} - t_0)}{k_{00} + (t_1^{(k)} - t_0)} (\bar{y}_0 - \mu_0^0)(\bar{y}_0 - \mu_0^0)^T +$$
$$\frac{k_{01}(t_2^{(k)} - t_1^{(k)})}{k_{01} + (t_2^{(k)} - t_1^{(k)})} (\bar{y}_1 - \mu_1^0)(\bar{y}_1 - \mu_1^0)^T + ... + \frac{k_{0N}(t_{N+1} - t_N^{(k)})}{k_{0N} + (t_{N+1}^{(k)} - t_N^{(k)})} (\bar{y}_N - \mu_N^0)(\bar{y}_N - \mu_N^0)^T \tag{4.63}$$

$$\nu_n = \nu_0 + n \tag{4.64}$$

By substituting (4.56)-(4.58) and (4.62) into (4.55) and also (4.53) into (4.52), we have the Q-function as

$$\int_\Sigma [\int_{\mu_0} P(Y_{t_0:t_1}|\mu_0, \Sigma, t) P(\mu_0|\Sigma, Y, t^{(k)}) P(\mu_0|\Sigma, t) d\mu_0$$
$$\int_{\mu_1} P(Y_{t1:t_2}|\mu_1, \Sigma, t) P(\mu_1|\Sigma, Y, t^{(k)}) P(\mu_1|\Sigma, t) d\mu_1 ...$$
$$\int_{\mu_N} P(Y_{t_N:t_{N+1}}|\mu_N, \Sigma, t) P(\mu_1|\Sigma, Y, t^{(k)}) P(\mu_1|\Sigma, t) d\mu_N \ \ ] P(\Sigma|t) P(\Sigma|Y, t^{(k)}) d\Sigma \tag{4.65}$$

Every single integration with respect to μ_i can be calculated using similar approach as in the derivation of (4.37). We can write

$$\int_{\mu_i} ...d\mu_i = (2\pi)^{-(t_{i+1}-t_i)+1)p/2} (k_{itk})^{p/2} (k_{0i})^{p/2} |\Sigma|^{-(t_{i+1}-t_i)+1)/2-1} |\Delta_i|^{1/2}$$
$$\times exp\{-\frac{1}{2}[\sum_{j=t_i+1}^{t_{i+1}} (y_j^T \Sigma^{-1} y_j) + k_{0i} {\mu_i^0}^T \Sigma^{-1} \mu_i^0 + k_{imk} \mu_{itk}^T \Sigma^{-1} \mu_{itk} - B_i^T A_i^{-1} B_i]\} \quad i = 0, 1, ..., N \tag{4.66}$$

where

$$A_i = (k_{0i} + t_{i+1} - t_i + k_{itk}) \Sigma^{-1} \quad i = 0, 1, ..., N \tag{4.67}$$
$$\Delta_i = A_i^{-1} \quad i = 0, 1, ..., N \tag{4.68}$$
$$B_i = \Sigma^{-1} (\sum_{j=t_i+1}^{t_{i+1}} y_j - k_{0i} \mu_i^0 - k_{itk} \mu_{itk}) \quad i = 0, 1, ..., N \tag{4.69}$$

After some simplifications, the integration in (4.65) can be written as

$$Q(t|t^{(k)}) = g(k) \int_\Sigma |\Sigma|^{-(n+(N+1)/2)} \times exp(-\frac{1}{2} \sum_{i=0}^{N} \{[\sum_{j=t_i+1}^{t_{i+1}} (y_j^T \Sigma^{-1} y_j) +$$
$$k_{0i} {\mu_i^0}^T \Sigma^{-1} \mu_i^0 + k_{imk} \mu_{itk}^T \Sigma^{-1} \mu_{itk} - B_i^T A_i^{-1} B_i]\}) P(\Sigma|t) P(\Sigma|Y, t^{(k)}) d\Sigma \tag{4.70}$$

where

$$g(k) = (2\pi)^{-n(N+1)p/2} \prod_{i=0}^{N} [(k_{itk})^{p/2} (k_{0i})^{p/2} (k_{0i} + t_{i+1} - t_i + k_{itk})^{-p/2} \tag{4.71}$$

Using trace properties in matrices, (4.70) can be written as

$$\begin{aligned} &\int_{\Sigma} ...d\Sigma = g_1(k) \int_{\Sigma} |\Sigma|^{-(n+(N+1)/2+\frac{\nu_n+p+1}{2}+\frac{\nu_0+p+1}{2})} exp(-\frac{1}{2} \sum_{i=0}^{N} \{[\sum_{j=t_i+1}^{t_{i+1}} (y_j^T \Sigma^{-1} y_j) + \\ &k_{0i} {\mu_i^0}^T \Sigma^{-1} \mu_i^0 + k_{imk} \mu_{itk}^T \Sigma^{-1} \mu_{itk} - B_i^T A_i^{-1} B_i]\}) IW(\nu_0, \Psi_0) IW(\nu_n, \Psi_n) d\Sigma \\ &= g_1(k) \int_{\Sigma} |\Sigma|^{-(n+(N+1)/2+\frac{\nu_n+p+1}{2}+\frac{\nu_0+p+1}{2})} exp(-\frac{1}{2} tr[Q_0 \Sigma^{-1} + \Psi_0 \Sigma^{-1} + \Psi_n \Sigma^{-1}]) d\Sigma \\ &= g_1(k) \int_{\Sigma} |\Sigma|^{-\frac{\nu_{nn}+p+1}{2}} exp(-\frac{1}{2} tr[\Psi_{nn} \Sigma^{-1}]) d\Sigma \end{aligned} \tag{4.72}$$

where Q_0 can be expressed as

$$\begin{aligned} Q_0 = &\sum_{i=0}^{N} [\sum_{j=t_i+1}^{t_{i+1}} (y_j y_j^T) + k_{0i} \mu_i^0 {\mu_i^0}^T + k_{itk} \mu_{itk} {\mu_{itk}}^T + \\ &(\sum_{j=t_i+1}^{t_{i+1}} y_j - k_{0i} \mu_i^0 - k_{itk} \mu_{itk})(\sum_{j=t_i+1}^{t_{i+1}} y_j - k_{0i} \mu_i^0 - k_{itk} \mu_{itk})^T] \end{aligned} \tag{4.73}$$

and $\Psi_{nn} = Q_0 + \Psi_0 + \Psi_n$ and $\nu_{nn} = N + 2 + 2n + \nu_0 + \nu_n + p$

$$g_1(k) = g(k) \frac{|\Psi_0|^{\frac{\nu_0}{2}}}{2^{\frac{\nu_0 p}{2}} Z(\nu_0, p)} \frac{|\Psi_n|^{\frac{\nu_n}{2}}}{2^{\frac{\nu_n p}{2}} Z(\nu_n, p)} \tag{4.74}$$

Thus, (4.70) can be derived in a closed form as

$$Q(t|t^{(k)}) = g_1(k) \frac{2^{\frac{\nu_{nn} p}{2}} Z(\nu_{nn}, p)}{|\Psi_{nn}|^{\frac{\nu_{nn}}{2}}} \tag{4.75}$$

The next step in EM is M-step which is obtained as

$$t^{(k+1)} = \arg\max_t Q(t|t^{(k)}) \tag{4.76}$$

which can be easily solved.

So far, the problem is solved under scenario of unknown constant covariance. In next section, we derive the solution under the EM framework to change point detection problem where both mean and covariance of data change simultaneously. Similarly, the problem is solved for single change detection first and then the solution is extended to multiple change points.

4.5 EM in Simultaneous Mean and Covariance Change Detection

4.5.1 Single Mean and Covariance Change Detection

In this section, the mean change detection problem is generalised to the case that both mean and covariance of data, which are unknown, are subject to simultaneous change at a random point. In other words, we assume that the mean and covariance of data are first μ_1 and Σ_1 and then they change to μ_2 and Σ_2. The assumption for prior distributions, like previous sections, can be selected as

$$P(\Sigma_1|\nu_{01},\Psi_{01}) = \frac{|\Psi_{01}|^{\frac{\nu_{01}}{2}}|\Sigma_1|^{-\frac{\nu_{01}+p+1}{2}}exp(\frac{-tr(\Psi_{01}\Sigma_1^{-1})}{2})}{2^{\frac{\nu_{01}p}{2}}Z(\nu_{01},p)} \tag{4.77}$$

$$P(\mu_1|\mu_1^0,\frac{\Sigma_1}{k_{01}}) = (2\pi)^{-p/2}|\Sigma_1|^{-1/2}exp\{-\frac{k_{01}}{2}(\mu_1-\mu_1^0)^T\Sigma_1^{-1}(\mu_1-\mu_1^0)\} \tag{4.78}$$

$$P(\Sigma_2|\nu_{02},\Psi_{02}) = \frac{|\Psi_{02}|^{\frac{\nu_{02}}{2}}|\Sigma_2|^{-\frac{\nu_{02}+p+1}{2}}exp(\frac{-tr(\Psi_{02}\Sigma_2^{-1})}{2})}{2^{\frac{\nu_{02}p}{2}}Z(\nu_{02},p)} \tag{4.79}$$

$$P(\mu_2|\mu_2^0,\frac{\Sigma_2}{k_{02}}) = (2\pi)^{-p/2}|\Sigma_2|^{-1/2}exp\{-\frac{k_{02}}{2}(\mu_2-\mu_2^0)^T\Sigma_2^{-1}(\mu_2-\mu_2^0)\} \tag{4.80}$$

As we can see, since the number of unknown parameters has increased, the number of priors and hence the number of hyperparameters has also increased accordingly. Similarly, the hidden variables become $\mu_1,\mu_2,\Sigma_1,\Sigma_2$. Therefore, E-step formulation of problem has the following form:

$$\begin{aligned} Q(m|m^{(k)}) &= E_{\mu_1,\mu_2,\Sigma_1,\Sigma_2|Y,m^{(k)}}\{P(Y,\mu_1,\mu_2,\Sigma_1,\Sigma_2|m)\} \\ &= \int P(Y,\mu_1,\mu_2,\Sigma_1,\Sigma_2|m)P(\mu_1,\mu_2,\Sigma_1,\Sigma_2|Y,m^{(k)})d\mu_1 d\mu_2 d\Sigma_1 d\Sigma_2 \end{aligned} \tag{4.81}$$

The probabilities in integration of (4.81) can be expressed as

$$\begin{aligned} P(Y,\mu_1,\mu_2,\Sigma_1,\Sigma_2|m) &= P(Y_{1:m}|\mu_1,\Sigma_1,m)P(Y_{m+1:n}|\mu_2,\Sigma_2,m)P(\mu_1|\Sigma_1)P(\mu_2|\Sigma_2)P(\Sigma_1)P(\Sigma_2) \\ &= \prod_{i=1}^{m}\mathcal{N}(Y_i|\mu_1,\Sigma_1)\prod_{i=m+1}^{n}\mathcal{N}(Y_i|\mu_2,\Sigma_2)\mathcal{N}(\mu_1|\mu_1^0,(k_{01})^{-1}\Sigma_1)\times \\ &\quad \mathcal{N}(\mu_2|\mu_2^0,(k_{02})^{-1}\Sigma_2)IW(\nu_{01},\Psi_{01})IW(\nu_{02},\Psi_{02}) \end{aligned} \tag{4.82}$$

Here, it is assumed that Σ_1, Σ_2 and m are prior independent. It is also assumed that μ_1 and μ_2 are prior independent. In addition, using chain rule, one can write

$$\begin{aligned} &P(\mu_1,\mu_2,\Sigma_1,\Sigma_2|Y,m^{(k)}) \\ &\quad = P(\mu_1|\Sigma_1,Y,m^{(k)})P(\mu_2|\Sigma_2,Y,m^{(k)})P(\Sigma_1|Y,m^{(k)})P(\Sigma_1|Y,m^{(k)}) \\ &\quad = \mathcal{N}(\mu_1|\mu_{1mk},(k_{1mk})^{-1}\Sigma_1)\mathcal{N}(\mu_2|\mu_{2mk},(k_{2mk})^{-1}\Sigma_2)IW(\nu_{1m},\Psi_{1m})IW(\nu_{2m},\Psi_{2m}) \end{aligned} \tag{4.83}$$

where

$$\nu_{1m} = \nu_{01} + m^{(k)} \tag{4.84}$$

$$\nu_{2m} = \nu_{02} + n - m^{(k)} \tag{4.85}$$

$$k_{1mk} = k_{01} + m^{(k)} \tag{4.86}$$

$$k_{2mk} = k_{02} + n - m^{(k)} \tag{4.87}$$

$$\mu_{1mk} = \frac{1}{k_{01} + m^{(k)}}(k_{01}\mu_1^0 + m^{(k)}\bar{y}_1) \tag{4.88}$$

$$\mu_{2mk} = \frac{1}{k_{02} + n - m^{(k)}}(k_{02}\mu_2^0 + (n - m^{(k)})\bar{y}_2) \tag{4.89}$$

$$\Psi_{1m} = \Psi_{01} + \sum_{i=1}^{m^{(k)}}(y_i - \bar{y}_1)(y_i - \bar{y}_1)^T + \frac{k_{01}m^{(k)}}{k_{01} + m^{(k)}}(\bar{y}_1 - \mu_1^0)(\bar{y}_1 - \mu_1^0)^T \tag{4.90}$$

$$\Psi_{2m} = \Psi_{02} + \sum_{i=m^{(k)}+1}^{n}(y_i - \bar{y}_2)(y_i - \bar{y}_2)^T + \frac{k_{02}(n - m^{(k)})}{k_{02} + n - m^{(k)}}(\bar{y}_2 - \mu_2^0)(\bar{y}_2 - \mu_2^0)^T \tag{4.91}$$

As noted earlier, in derivation of (4.83), the fact that given the change point $m^{(k)}$, μ_1 and μ_2 are conditionally independent is taken into account. By incorporating (4.82) and (4.83) in (4.81), Q-function can be derived as

$$\begin{aligned}Q(m|m^{(k)}) = &\int_{\Sigma_1}\int_{\mu_1}\prod_{i=1}^{m}\mathcal{N}(Y_i|\mu_1, \Sigma_1)\mathcal{N}(\mu_1|\mu_1^0, (k_{01})^{-1}\Sigma_1)\mathcal{N}(\mu_1|\mu_{1mk}, (k_{1mk})^{-1}\Sigma_1)\\ &IW(\nu_{01}, \Psi_{01})IW(\nu_{1m}, \Psi_{1m})d\mu_1 d\Sigma_1 \times \int_{\Sigma_2}\int_{\mu_2}\prod_{i=m+1}^{n}\mathcal{N}(Y_i|\mu_2, \Sigma_2)\mathcal{N}(\mu_2|\mu_2^0, (k_{02})^{-1}\Sigma_2)\\ &\times N(\mu_2|\mu_{2mk}, (k_{2mk})^{-1}\Sigma_2)IW(\nu_{02}, \Psi_{02})IW(\nu_{2m}, \Psi_{2m})d\mu_2 d\Sigma_2\end{aligned} \tag{4.92}$$

Integration with respect to μ_1 leads to

$$\begin{aligned}\int_{\mu_1}...d\mu_1 = &(2\pi)^{-(m+1)p/2}(k_{1mk})^{p/2}(k_{01})^{p/2}|\Sigma_1|^{-m/2-1}|\Delta_1|^{1/2}\\ &\times exp\{-\frac{1}{2}[\sum_{i=1}^{m}(y_i^T\Sigma_1^{-1}y_i) + k_{01}{\mu_1^0}^T\Sigma_1^{-1}\mu_1^0 + k_{1mk}\mu_{1mk}^T\Sigma_1^{-1}\mu_{1mk} - B^TA^{-1}B]\}\end{aligned} \tag{4.93}$$

where

$$A = (k_{01} + m + k_{1mk})\Sigma_1^{-1} \tag{4.94}$$

$$\Delta_1 = A^{-1} = (k_{01} + m + k_{1mk})^{-1}\Sigma_1 \tag{4.95}$$

$$B = \Sigma_1^{-1}(\sum_{i=1}^{m} y_i - k_{01}\mu_1^0 - k_{1mk}\mu_{1mk}) \tag{4.96}$$

By integrating out Σ_1, it follows that

$$\int_{\Sigma_1} ...d\Sigma_1 = w(k) \times \int_{\Sigma_1} |\Sigma_1|^{-\frac{\nu_{nn}+p+1}{2}} exp(-\frac{1}{2}tr[\Psi_{nn}\Sigma_1^{-1}])d\Sigma_1 \tag{4.97}$$

where

$$\begin{aligned} w(k) &= (2\pi)^{-(m+1)p/2}(k_{1mk})^{p/2}(k_{01})^{p/2}(k_{01}+m+k_{1mk})^{-p/2}|\Psi_{01}|^{\frac{\nu_{01}}{2}}|\Psi_{1m}|^{\frac{\nu_{1m}}{2}} \times \\ &\frac{1}{2^{\frac{\nu_{01}p}{2}}Z(\nu_{01},p)2^{\frac{\nu_{1m}p}{2}}Z(\nu_{1m},p)} \end{aligned} \tag{4.98}$$

$$\Psi_{1nn} = Q_{01} + \Psi_{01} + \Psi_{1m} \tag{4.99}$$

$$\nu_{1nn} = \nu_{01} + \nu_{1m} + p + m + 2 \tag{4.100}$$

$$\begin{aligned} Q_{01} &= \sum_{i=1}^{m} y_i y_i^T + k_{01}\mu_1^0 {\mu_1^0}^T + k_{1mk}\mu_{1mk}{\mu_{1mk}}^T + \\ &(\sum_{i=1}^{m} y_i - k_{01}\mu_1^0 - k_{1mk}\mu_{1mk})(\sum_{i=1}^{m} y_i - k_{01}\mu_1^0 - k_{1mk}\mu_{1mk})^T \end{aligned} \tag{4.101}$$

Since the integrand in (4.97) is Inverse-Wishart distribution as $IW(\nu_{1nn}, \Psi_{1nn})$, the first integration in (4.92) can be calculated as

$$\int_{\Sigma_1} ...d\Sigma_1 = w(k)\frac{2^{\frac{\nu_{1nn}p}{2}}Z(\nu_{1nn},p)}{|\Psi_{1nn}|^{\frac{\nu_{1nn}}{2}}} \tag{4.102}$$

Using similar approach, the second integration in (4.92) yields

$$\int_{\Sigma_2} ...d\Sigma_2 = J(k)\frac{2^{\frac{\nu_{2nn}p}{2}}Z(\nu_{2nn},p)}{|\Psi_{2nn}|^{\frac{\nu_{2nn}}{2}}} \tag{4.103}$$

where

$$\begin{aligned} J(k) &= (2\pi)^{-(n-m+1)p/2}(k_{2mk})^{p/2}(k_{02})^{p/2}(k_{02}+n-m+k_{2mk})^{-p/2}|\Psi_{02}|^{\frac{\nu_{02}}{2}} \\ &\times |\Psi_{2m}|^{\frac{\nu_{2m}}{2}} \frac{1}{2^{\frac{\nu_{02}p}{2}}Z(\nu_{02},p)2^{\frac{\nu_{2m}p}{2}}Z(\nu_{2m},p)} \end{aligned} \tag{4.104}$$

$$\Psi_{2nn} = Q_{02} + \Psi_{02} + \Psi_{2m} \tag{4.105}$$

$$\nu_{2nn} = \nu_{02} + \nu_{2m} + p + n - m + 2 \tag{4.106}$$

$$\begin{aligned} Q_{02} &= \sum_{i=m+1}^{n} y_i y_i^T + k_{02}\mu_2^0 {\mu_2^0}^T + k_{2mk}\mu_{2mk}{\mu_{2mk}}^T + \\ &(\sum_{i=m+1}^{n} y_i - k_{02}\mu_2^0 - k_{2mk}\mu_{2mk})(\sum_{i=m+1}^{n} y_i - k_{02}\mu_2^0 - k_{2mk}\mu_{2mk})^T \end{aligned} \tag{4.107}$$

Thus, Q-function is derived by multiplying (4.102) and (4.103):

$$Q(m|m^{(k)}) = w(k)\frac{2^{\frac{\nu_{1nn}p}{2}}Z(\nu_{1nn},p)}{|\Psi_{1nn}|^{\frac{\nu_{1nn}}{2}}} \times J(k)\frac{2^{\frac{\nu_{2nn}p}{2}}Z(\nu_{2nn},p)}{|\Psi_{2nn}|^{\frac{\nu_{2nn}}{2}}} \tag{4.108}$$

Eventually, the M-step of EM is formulated as

$$m^{(k+1)} = \arg\max_{m} Q(m|m^{(k)}) \tag{4.109}$$

E-step and M-step iterate until convergence.

4.5.2 Simultaneous Multiple Mean and Covariance Changes Detection

Having derived the Q-function in the case of single change in the mean and covariance in Section 4.5.1, we can extend the result to multiple changes detection. For every segment of the data, two prior distributions are defined for mean and covariance respectively. These priors, in every segment, are assumed to be independent of each other. Therefore, E-step can be written as

$$\begin{aligned}
&Q(t|t^{(k)}) = E_{\mu_0,\mu_1,...,\mu_N,\Sigma_0,\Sigma_1,...,\Sigma_N|Y,t^{(k)}}\{P(Y,\mu_0,\mu_1,...,\mu_N,\Sigma_0,\Sigma_1,...,\Sigma_N|t)\} \\
&= \int P(Y,\mu_0,\mu_1,...,\mu_N,\Sigma_1,\Sigma_2,...,\Sigma_N|t)P(\mu_0,\mu_1,...,\mu_N,\Sigma_0,\Sigma_1,...,\Sigma_N|Y,t^{(k)})... \\
&d\mu_0 d\mu_1...d\mu_N d\Sigma_0 d\Sigma_1..d\Sigma_N \qquad (4.110)
\end{aligned}$$

By separating the probabilities and integrals, and following the same approach as in Section 4.5.1, Q-function is derived as

$$Q(t|t^{(k)}) = \prod_{i=0}^{N} w_i(k) \frac{2^{\frac{\nu_{inn}p}{2}} Z(\nu_{inn},p)}{|\Psi_{inn}|^{\frac{\nu_{inn}}{2}}} \qquad (4.111)$$

where

$$\begin{aligned}
&w_i(k) = (2\pi)^{-(t_{i+1}-t_i+1)p/2}(k_{imk})^{p/2}(k_{0i})^{p/2}(k_{0i}+m+k_{imk})^{-p/2}|\Psi_{0i}|^{\frac{\nu_{0i}}{2}}|\Psi_{im}|^{\frac{\nu_{im}}{2}} \times \\
&\quad \frac{1}{2^{\frac{\nu_{0i}p}{2}} Z(\nu_{0i},p) 2^{\frac{\nu_{im}p}{2}} Z(\nu_{im},p)} \qquad (4.112) \\
&\Psi_{inn} = Q_{0i} + \Psi_{0i} + \Psi_{im} \qquad (4.113) \\
&Q_{0i} = \sum_{j=t_i+1}^{t_{i+1}} [y_j y_j^T] + k_{0i}\mu_i^0 {\mu_i^0}^T + k_{imk}\mu_{imk}{\mu_{imk}}^T + \\
&\quad (\sum_{j=t_i+1}^{t_{i+1}} y_j - k_{0i}\mu_i^0 - k_{imk}\mu_{imk})(\sum_{ij=t_i+1}^{t_{i+1}} y_j - k_{0i}\mu_i^0 - k_{imk}\mu_{imk})^T \qquad (4.114) \\
&\qquad (4.115)
\end{aligned}$$

and likewise, having derived the Q-function, the next step is M-step which is

$$t^{(k+1)} = \arg\max_t Q(t|t^{(k)}) \qquad (4.116)$$

In the following, the performance of the proposed algorithms is evaluated through several examples followed by a pilot-scale experimental study.

4.6 Simulation

4.6.1 Single Mean Change Detection with Unknown Covariance

Here, single change point detection in the presence of unknown covariance is simulated for multivariate data, $p = 3$, i.e. $y = [y_1, y_2, y_3]$. $n = 100$ samples are generated. At time instant $m = 80$, the mean of data changes. The hyperparameters of prior distributions are selected as:

$$\mu_{initial} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \delta_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \ n = 100, m = 80, m_{initial} = 60 \ \nu_0 = 5, \Psi_0 = \begin{pmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{pmatrix},$$

$$k_{01} = 10, k_{02} = 5, \ \mu_1^0 = [2, 2, 2]^T, \mu_2^0 = [2, 2, 2]^T$$

The uncertainty in prior distributions of mean values is governed by the hyperparameters k_{01} or k_{02}. The smaller these free parameters, the less prior knowledge. The reason is, increasing these parameters results in a higher covariance and hence more uncertainties, which is equivalent to less prior information. The variables, $y = [y_1, y_2, y_3]$, are illustrated in Figure 4.1. As we can see, the change is small and not easy to detect by visualization. Using EM formulation, in 6 iterations, the algorithm converges. The results of Q-function at each iteration are shown in Figure 4.2.

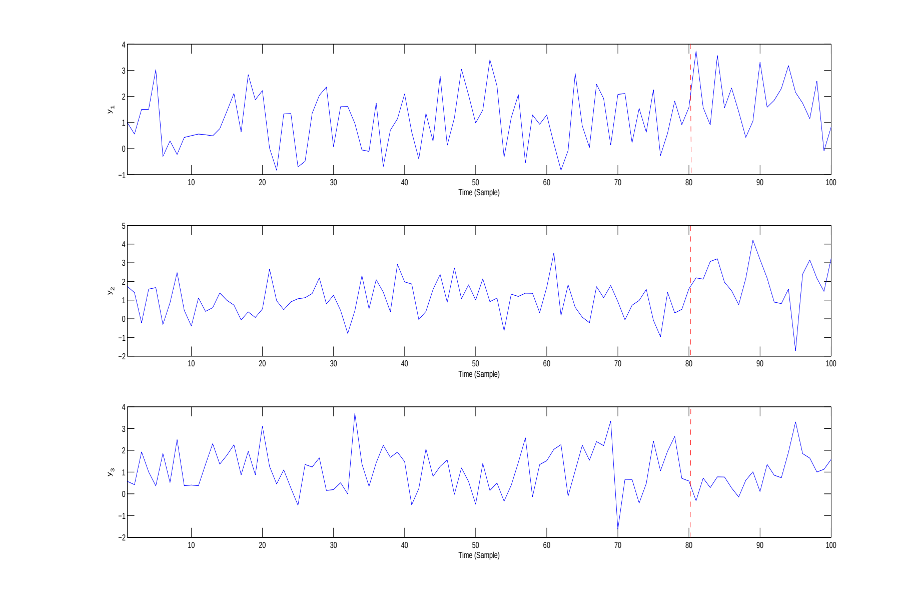

Figure 4.1: Three Variables $y = [y_1, y_2, y_3]$

From Figure 4.2 , it is apparent that at every iteration the Q-function has increased till convergence to the fixed point. At the last iteration, the maximum is achieved at true change point, $m = 80$. As we can see, the algorithm has fast convergence in terms of the number of iterations.

In order to quantify the performance of change point detection, two performance measures, OP and AVTI, are defined as

$$OP = \frac{\text{Number of Change Points Correctly Identified}}{\text{Number of Change Points Occurred}}$$

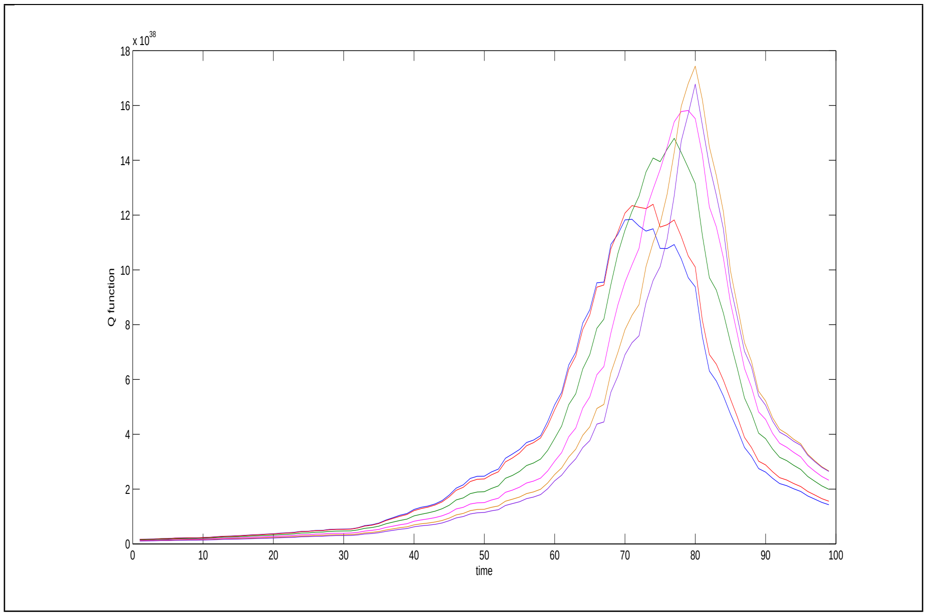

Figure 4.2: Q-Function at 6 Iterations Until Convergence, i.e. m=80

The other performance measure is defined to represent the probability of false alarm as

$$AVTI = \frac{\text{Number of Change Points Wrongly Identified}}{\text{Number of Simulation Runs Made}}$$

These two performance measures are widely used for gross error detection in the literature which have also been applied to change point detection [1].

Here, the OP and $AVTI$ are calculated for 100 Monte Carlo runs of the algorithm. The results are shown in Table 4.1. Not surprisingly, as the change magnitude increases, the algorithm performs better. From this table, we can observe that the proposed algorithm is capable of detecting small changes as well. The performance of the algorithm in the case of improper hyperparameters is also evaluated. EM can be less sensitive to wrong hyperparameters [41]; however, it is sensitive to initial values of the iterations. [41] recommended the randomization approach for initializing EM that has been adopted in this example. In other words, random sets of initial values are selected and the one that maximizes the likelihood is obtained. In the following, a more realistic simulated example using Continuous Stirred Tank Reactor (CSTR) is given for both mean and covariance change detection.

Table 4.1: Performance Results of EM for Different Change Magnitudes

	$Bias$	OP	$AVTI$
EM	$0.5\sigma_i$	0.45	0.55
	$0.75\sigma_i$	0.75	0.25
	$1\sigma_i$	1	0

4.6.2 Continuous Stirred Tank Reactor (CSTR)

The CSTR case study, as discussed in detail in [39], is used again to evaluate the EM algorithm performance in change detection of mean and covariance. The irreversible exothermic reaction $A \rightarrow B$ occurs inside a constant-volume reactor cooled by a single coolant stream. The system dynamics can be written as

$$\dot{C_A}(t) = \frac{q(t)}{V}(C_{A0}(t) - C_A(t)) - k_0 C_A(t) exp(-\frac{E}{RT(t)}) \tag{4.117}$$

$$\begin{aligned} \dot{T}(t) = \frac{q(t)}{V}(T_0(t) - T(t)) + \frac{\Delta H k_0 C_A(t)}{\rho C_p} exp(-\frac{E}{RT(t)}) + \\ \frac{\rho_c C_{pc}}{\rho C_p V} q_c(t)\{1 - exp(\frac{-hA}{q_c(t)\rho C_p})\}(T_{c0} - T(t)) \end{aligned} \tag{4.118}$$

The parameter definition and nominal values for CSTR are given in [41]. The states of the system are C_A (concentration of component A) and T (the reactor temperature). The outputs are the same as states but with measurement noise. Similar to [41], the change in outputs is due to change in the input. In other words, in order to introduce a change to the measurement, at a certain time instant, the system input changes from one operating point to another driving the outputs to a new operating condition. Moreover, when input changes to a new value, the variance of simulated noise also changes. In this study, initially a constant input $q_c = 97$ is fed to the system so the system operates under the steady state condition. The input is changed at time $t = 159$ with the magnitude of change being 2 L/min. As this occurs, the outputs start to change after a short delay. The measurement noise is also forced to increase at this time. The standard deviation of output noise is first 10 % of the states and then it increases to 15 % after the change. The histogram of outputs along with the time trends of data are shown in Figure 4.3 and 4.4 respectively.

As we can see, the mean and covariance of data both have changed. When applying EM algorithm, the hyperparameters are selected randomly as:

$$n = 248, p = 2, m_{initial} = 100 \ \nu_{01} = 7, \nu_{02} = 6, \Psi_{01} = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 5 \end{pmatrix}, \ \Psi_{02} = 0.5 \begin{pmatrix} 1 & 0.1 \\ 0.1 & 5 \end{pmatrix},$$

$$\Psi_{03} = 0.7 \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}, \ k_{01} = 50, k_{02} = 0.9, \ \mu_1^0 = [0.1, 410]^T, \ \mu_2^0 = [0.1; 410]^T$$

The randomization approach for the selection of the initial values to start the EM algorithm is adopted here again. Besides, in this example, the number of iterations to convergence is two indicating fast convergence. The Q-function at the last iteration is illustrated in Figure 4.5 with the maximum achieved at $m = 162$. Note that the y-axis is in logarithmic scale.

As we can see, EM also performs satisfactorily in the case of both mean and covariance change. In the following section, an experimental case study is also provided to further evaluate the proposed algorithm.

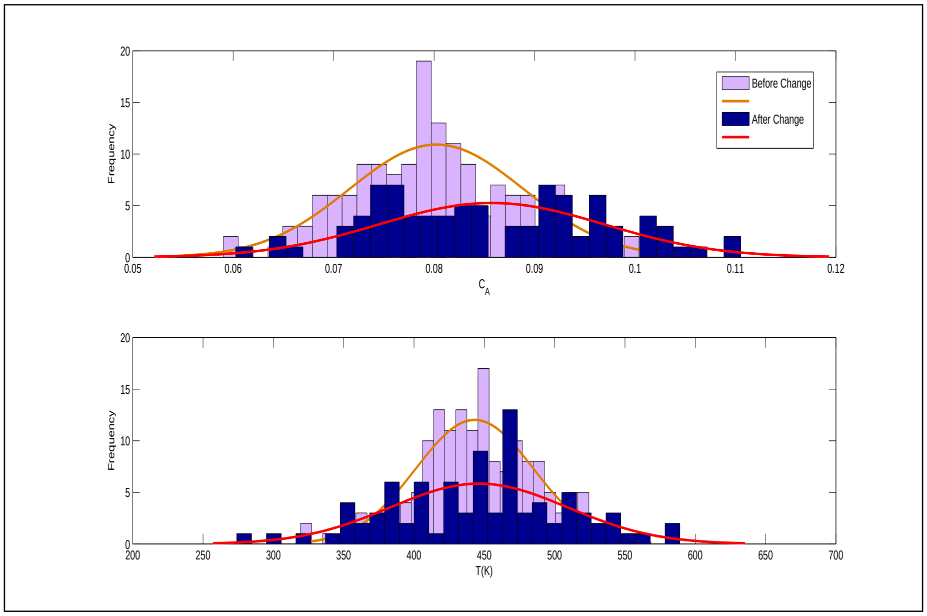

Figure 4.3: Histogram of Concentration and Temperature Before and After Change

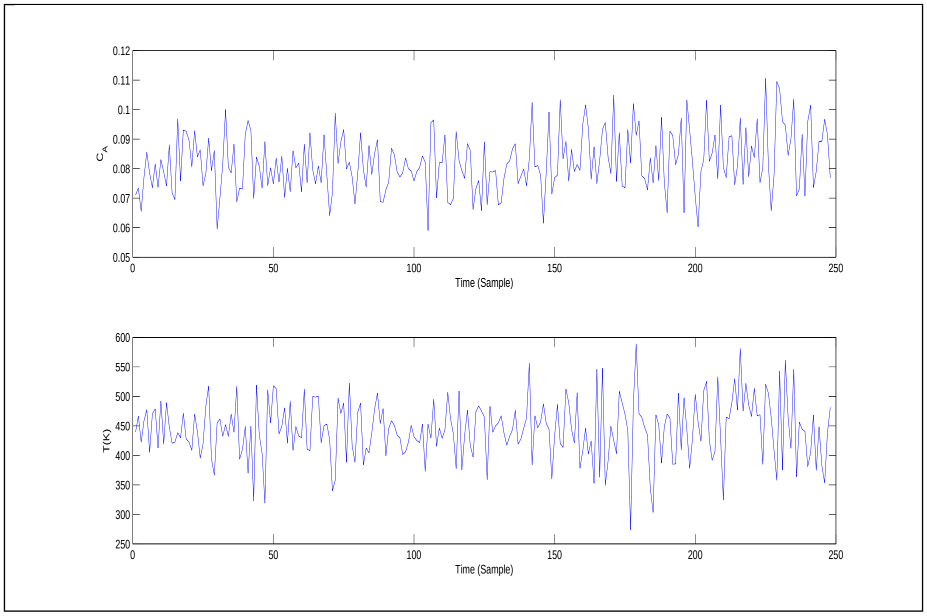

Figure 4.4: Time Trend Plot of Concentration (Upper) and Temperature (Lower)

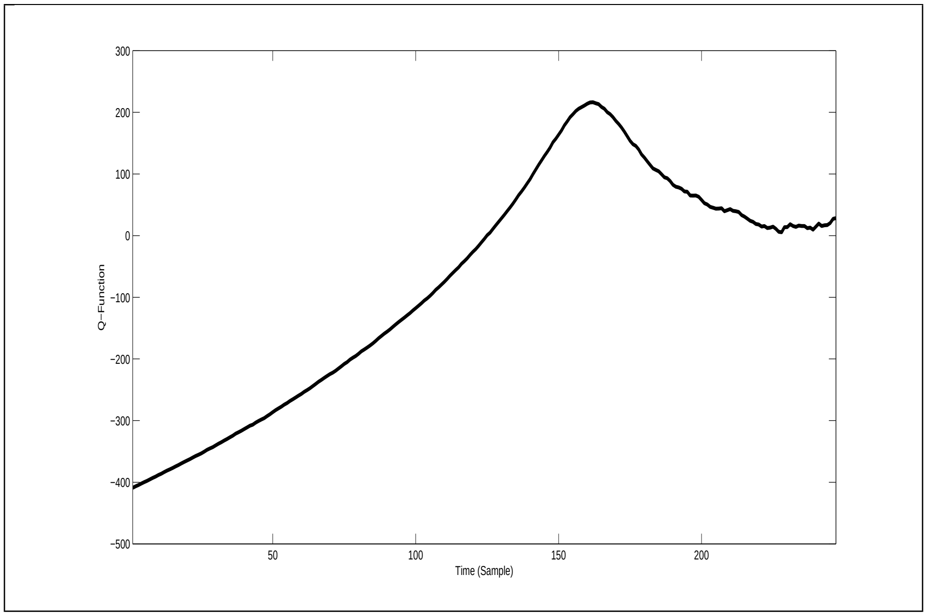

Figure 4.5: $Q-$function in Last Iteration of EM

4.7 Experimental Evaluation: Hybrid Tank System

The hybrid tank system used in [41], is selected to collect data but this time, the experiment is performed in a different way. This system consists of three connected tanks, six on/off valves and two pumps. The schematic of hybrid tank is shown in Figure 4.6. The valves can be manipulated to change the flow rate to or out of these tanks. By opening or closing these valves, the system dynamics changes accordingly.

There are two cascade loops for the left and right tanks which control the levels of these tanks by manipulating the set point of flow controllers. Two proportional controllers are designed to maintain the level inside the left and right tank approximately at 75% percent. At first, the valves $V1$, $V2$, $V3$, $V4$ are closed and $V5$ to $V9$ are open.

Since the objective is to incorporate the covariance change along with the mean change, as the valve status changes, the measurement noise changes so as to incorporate the change in covariance of the data. In order to implement the changing covariance, a switch is provided which changes the noise variance at those instants when a change in valves occurs. The valves status changes in the same way as they did in [41]. There are totally three changes. As the level in both left and right tanks reach 75%, the valves $V1$ and $V3$ are open imposing a change in levels of the tanks. In order to introduce the second change, the valves $V2$ and $V4$ are open resulting in the second change in the tank levels. Having reached the second steady state condition, the valves $V1$ and $V3$ are closed leading to the third change. The valves positions and the left and right levels are shown in Figures 4.7 and 4.8 respectively. As we can see, in this experiment, we have multiple change points in both the mean and covariance of data. It is also worth mentioning that all the communications between Matlab and the hybrid tank system are established through OLE for Process Control (OPC) which is a common platform in industry for data communication.

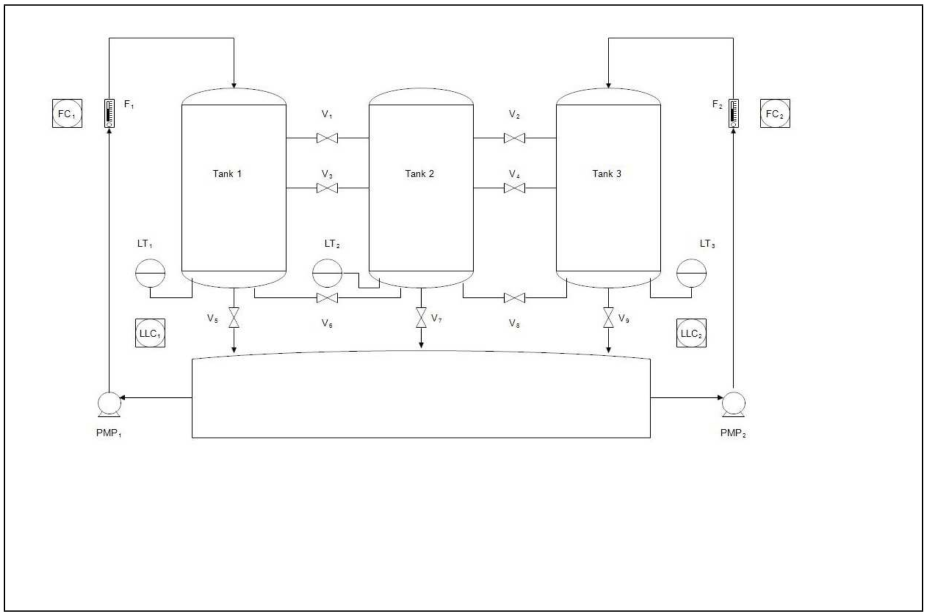

Figure 4.6: Schematic of Hybrid Tank

To detect the change points using the EM algorithm, the hyperparameters are selected randomly as:

$$\mu_{initial} = \begin{pmatrix}1\\1\\1\end{pmatrix}, \delta_1 = \begin{pmatrix}1\\1\\0\end{pmatrix}, \; n = 218, p = 2, t_{initial} = [20, 90, 150]$$

$$\nu_{01} = 6, \nu_{02} = 7, \nu_{03} = 7, \nu_{04} = 7, \Psi_{00} = \begin{pmatrix}1 & 0.1\\0.1 & 1\end{pmatrix}, \; \Psi_{01} = 0.5\begin{pmatrix}1 & 0.1\\0.1 & 1\end{pmatrix},$$

$$\Psi_{02} = 0.6\begin{pmatrix}1 & 0.1\\0.1 & 1\end{pmatrix}, \; \Psi_{03} = 0.7\begin{pmatrix}1 & 0.1\\0.1 & 1\end{pmatrix}, \; k_{01} = 10, k_{02} = 10, k_{03} = 10, k_{04} = 5,$$

$$\mu_1^0 = [60, 60]^T, \; \mu_2^0 = [60, 60]^T, \; \mu_3^0 = [60, 60]^T, \; \mu_4^0 = [60, 60]^T$$

In Figure 4.9, Q-function is illustrated with respect to time samples. Note that evaluation of Q-function with respect to all possible values of t, provided that one imposes constraints on consecutive samples, can reduce the computation at every iteration depending on the number of samples and also the number of change points. In evaluation of Q-function, when t_1 changes from 1 to $n-1$, t_2 changes from t_1+1 to $n-2$ and t_3 changes from t_2+1 to $n-3$. In this application, Q-function is derived as a $215 \times 215 \times 215$ matrix. Figure 4.9 illustrates the trajectory of Q-function. The y axis is in logarithmic scale. As we can see, EM is successful in detection of change points in the presence of unknown and changing covariance.

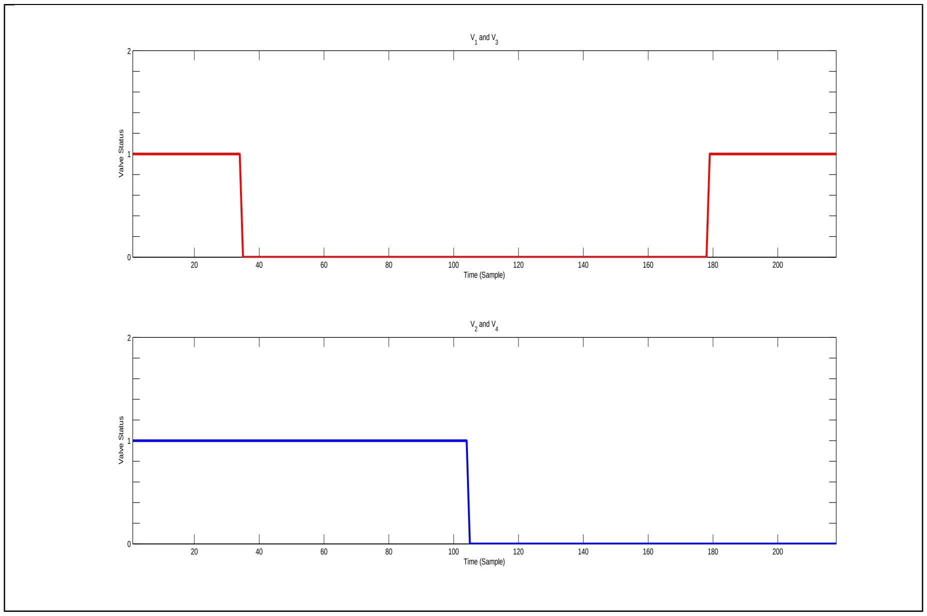

Figure 4.7: Valves Status in Hybrid Tank, 1=close and 0=open

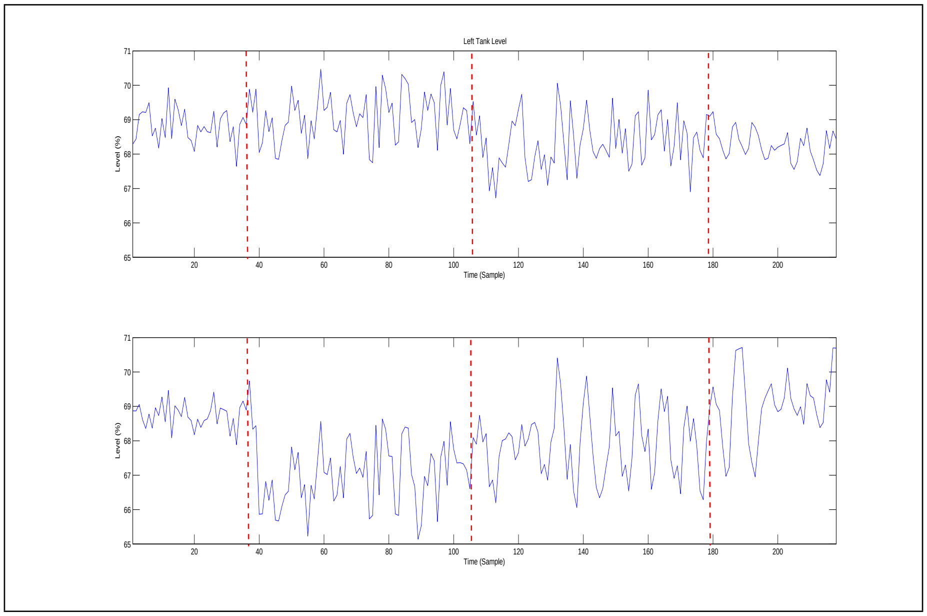

Figure 4.8: Valve Status in Hybrid Tank, 1 = *close* and 0 = *open*

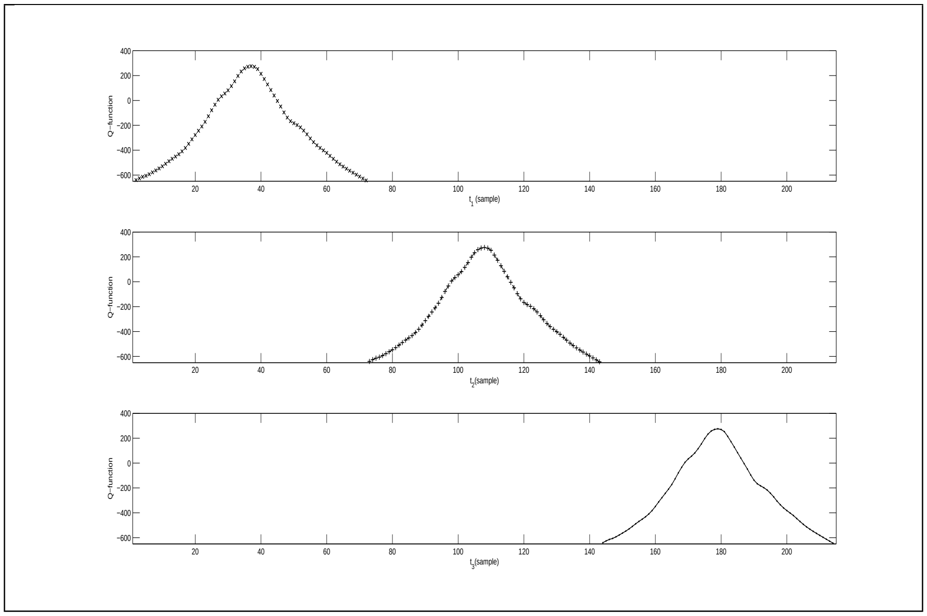

Figure 4.9: Trajectory of Q-function

4.8 Conclusion

In this chapter, a closed form solution based on EM algorithm is developed for change point detection problem in the presence of unknown and simultaneous changing mean and covariance. The proposed algorithm has satisfactory performance in identifying both single and multiple change points. Moreover, in the case of small changes and improper selection of priors, EM also performs effectively. Through several case studies, it was shown that the algorithm can achieve efficient and satisfactory performance.

Chapter 5

Mean Estimation in the Presence of Process Constraints Using Change Point Models

5.1 Mean Estimation Using EM Algorithm

As shown in previous chapters, in EM approach, depending on the selected hidden variables and desired parameters, the structure is adequately flexible to vary. For instance, assuming that the data consist of a mixture of distributions with various densities, a common framework for parameter estimation for this kind of problems is through EM algorithm. Depending on the model, the solution can vary. Thus, the model structure selection is a necessary step in parameter estimation.

In previous chapters, we estimated the change point assuming some prior distributions for certain unknown variables. One can also apply the change point formulation for mixture density parameter estimation problem. In other words, change point detection problem can be reformulated in order to estimate the mixture parameters. In the following section, change point model is employed to estimate the parameters before and after the change points.

5.1.1 Mean Estimation in Multiple Change Points Models

In this section, the objective is to estimate the parameters of mixture densities such as mean in different segments of data given the observation and prior of time instants at which changes occurs. Assume that N changing points exist in the multivariate data and all observations are denoted $Y = (Y_0, Y_1, ..., Y_N)$ where each Y_i represents the all data in segment i. There are N+1 segments. The means of $N+1$ segments of data can be expressed as $\mu = (\mu_0, \mu_1, ..., \mu_N)$.

Denote the hidden variables as $t = [t_1, t_2, ..., t_N]$, i.e. the vector of change points and the parameters to be estimated as $\mu = (\mu_0, \mu_1, ..., \mu_N)$. Again, it is assumed that $t_0 = 1$ and $t_{N+1} = n$ where n is the number of total observations. The prior information of change points is the same as in (3.17) and (3.18). The E-step can be written as

$$Q(\mu|\mu^{(k)}) = E_{t|Y,\mu^{(k)}}\{logP(Y,t|\mu)\} \tag{5.1}$$

where the term $P(Y,t|\mu)$ is the complete likelihood function. It can be simplified as

$$\begin{aligned} P(Y,t|\mu) &= \frac{P(Y,t,\mu)}{P(\mu)} \\ &= \frac{P(d|t,\mu)P(t,\mu)}{P(\mu)} = P(Y|t,\mu)P(t) \end{aligned} \tag{5.2}$$

$P(t,\mu) = P(t)P(\mu)$ since t and μ are independent. The expression in (5.1) can be rewritten as

$$\begin{aligned} Q(\mu|\mu^{(k)}) &= E_{t|Y,\mu^{(k)}}\{logP(Y,t|\mu)\} \\ &= E_{t|Y,\mu^{(k)}}\{log[P(Y|t,\mu)P(t)]\} \\ &= E_{t|Y,\mu^{(k)}}\{logP(Y|t,\mu) + logP(t)\} \end{aligned} \tag{5.3}$$

Since t is a discrete vector, the expectation has the following form:

$$Q(\mu|\mu^{(k)}) = \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} \{logP(Y|t,\mu) + logP(t)\}P(t|Y,\mu^{(k)}) \tag{5.4}$$

Also, the conditional probability $P(t|Y,\mu^{(k)})$ can be expressed as

$$\begin{aligned} P(t|Y,\mu^{(k)}) &= \frac{P(t,Y,\mu^{(k)})}{P(Y,\mu^{(k)})} = \frac{P(Y|t,\mu^{(k)})P(t,\mu^{(k)})}{P(Y|\mu^{(k)})P(\mu^{(k)})} \\ &= \frac{P(Y|t,\mu^{(k)})P(t)}{P(Y|\mu^{(k)})} \end{aligned} \tag{5.5}$$

Since $\mu^{(k)}$ is given in each iteration, then for a given mean vector, $P(Y|\mu^{(k)})$ is a constant value. As a result, the equation in (5.5) can be simplified as

$$P(t|Y,\mu^{(k)}) = c_1 P(Y|t,\mu^{(k)})P(t) \tag{5.6}$$

where $c_1 = \frac{1}{P(Y|\mu^{(k)})}$. Thus, (5.4) yields

$$\begin{aligned} Q(\mu|\mu^{(k)}) &= c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} \{logP(Y|t,\mu) + logP(t)\}P(Y|t,\mu^{(k)})P(t) \\ &= c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} \{P(Y|t,\mu^{(k)})P(t)logP(Y|t,\mu) + P(Y|t,\mu^{(k)})P(t)logP(t)\} \\ &= c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} \{P(Y|t,\mu^{(k)})P(t)logP(Y|t,\mu) + c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t,\mu^{(k)})P(t)logP(t)\} \end{aligned} \tag{5.7}$$

In the M-step, the maximization is performed with respect to μ. To achieve this, the derivative of $Q(\mu|\mu^{(k)})$ is calculated. Obviously, the second term in (5.7) does not depend on μ and hence maximization can be solved as

$$\begin{aligned} &\frac{d}{d\mu}Q(\mu|\mu^{(k)}) = 0 \\ &\frac{d}{d\mu}[c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} \{P(Y|t,\mu^{(k)})P(t)logP(Y|t,\mu)\}] = 0 \end{aligned} \tag{5.8}$$

Since $\mu = (\mu_0, \mu_1, ..., \mu_N)$, we have

$$\frac{d}{d\mu} = (\frac{d}{d\mu_0}, \frac{d}{d\mu_1}, ..., \frac{d}{d\mu_N}) \tag{5.9}$$

The derivative in (5.9) essentially consists of $N+1$ derivatives. Consider that the derivative is taken with respect to μ_i.

$$\frac{d}{d\mu_i}[c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} \{P(Y|t,\mu^{(k)})P(t)logP(Y|t,\mu)\}] = 0 \tag{5.10}$$

that is

$$c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t,\mu^{(k)})P(t)\frac{d}{d\mu_i}logP(Y|t,\mu) = 0 \tag{5.11}$$

As a result,

$$c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t,\mu^{(k)})P(t)\frac{d}{d\mu_i}logP(Y = (Y_0, Y_1, ..., Y_N)|t, \mu = (\mu_0, \mu_1, ...\mu_i, ..., \mu_N)) = 0 \tag{5.12}$$

Due to independent observations, we have

$$P(Y = (Y_0, Y_1, ..., Y_N)|t, \mu = (\mu_0, \mu_1, ...\mu_i, ..., \mu_N)) = \prod_{i=0}^{N} \mathcal{N}_p(Y_i|\mu_i, \Sigma) \tag{5.13}$$

where

$$\mathcal{N}_p(Y_i|\mu_i, , \Sigma) = (2\pi)^{-p/2}|\Sigma^{-1}|^{1/2}exp\{-\frac{1}{2}(Y_i - \mu_i)^T\Sigma^{-1}(Y_i - \mu_i)\} \tag{5.14}$$

As noted earlier, every i represents a segment of data in the interval $t_{i+1} - t_i$. The data in each segment are independent; hence they can be expressed as

$$\mathcal{N}_p(Y_i|\mu_i, , \Sigma) = (2\pi)^{-p/2}|\Sigma^{-1}|^{1/2}exp\{-\frac{1}{2}\sum_{j=t_i+1}^{t_{i+1}}(Y_{ij} - \mu_i)^T\Sigma^{-1}(Y_{ij} - \mu_i)\} \tag{5.15}$$

where Y_{ij} are the data in segment i and each j represents a single sample belonging to interval $t_{i+1} - t_i$. The time instances for these data are essentially $t_i + 1$ to t_{i+1}. Therefore, (5.13) yields

$$P(Y = (Y_0, Y_1, ..., Y_N)|t, \mu = (\mu_0, \mu_1, ...\mu_i, ..., \mu_N)) = \prod_{i=0}^{N} (2\pi)^{-p/2} |\Sigma^{-1}|^{1/2}$$
$$exp\{-\frac{1}{2} \sum_{j=t_i+1}^{t_{i+1}} (Y_{ij} - \mu_i)^T \Sigma^{-1} (Y_{ij} - \mu_i)\} \tag{5.16}$$

Using the *log* function, (5.16) can be simplified as

$$logP(Y = (Y_0, Y_1, ..., Y_N)|t, \mu = (\mu_0, \mu_1, ...\mu_i, ..., \mu_N)) = \sum_{i=0}^{N} log[(2\pi)^{-p/2} |\Sigma^{-1}|^{1/2}] +$$
$$\sum_{i=0}^{N} \{-\frac{1}{2} \sum_{j=t_i+1}^{t_{i+1}} (Y_{ij} - \mu_i)^T \Sigma^{-1} (Y_{ij} - \mu_i)\} \tag{5.17}$$

Substituting (5.16) into (5.12) leads to

$$c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)}) P(t) \frac{d}{d\mu_i} [\sum_{i=0}^{N} log[(2\pi)^{-p/2} |\Sigma^{-1}|^{1/2}] +$$
$$\sum_{i=0}^{N} \{-\frac{1}{2} \sum_{j=t_i+1}^{t_{i+1}} (Y_{ij} - \mu_i)^T \Sigma^{-1} (Y_{ij} - \mu_i)\}] = 0 \tag{5.18}$$

Since the first term of (5.18) does not contain μ_i, the expression in (5.18) is equivalent to

$$c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)}) P(t) \frac{d}{d\mu_i} [\sum_{i=0}^{N} \{-\frac{1}{2} \sum_{j=t_i+1}^{t_{i+1}} (Y_{ij} - \mu_i)^T \Sigma^{-1} (Y_{ij} - \mu_i)\}] = 0 \tag{5.19}$$

As this derivative is with respect to the *ith* term, we have

$$c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)}) P(t) \frac{d}{d\mu_i} [\{-\frac{1}{2} \sum_{j=t_i+1}^{t_{i+1}} (Y_{ij} - \mu_i)^T \Sigma^{-1} (Y_{ij} - \mu_i)\}] = 0$$
$$c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)}) P(t) (-\frac{1}{2}) [\sum_{j=t_i+1}^{t_{i+1}} -2\Sigma^{-1} (Y_{ij} - \mu_i)] = 0$$
$$c_1 \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)}) P(t) \Sigma^{-1} [\sum_{j=t_i+1}^{t_{i+1}} Y_{ij} - \sum_{j=t_i+1}^{t_{i+1}} \mu_i] = 0$$
$$\sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)}) P(t) \Sigma^{-1} \sum_{j=t_i+1}^{t_{i+1}} Y_{ij} = \sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)}) P(t) \Sigma^{-1} (t_{i+1} - t_i - 1) \mu_i \tag{5.20}$$

By solving (5.20), μ_i , $i = 0, 1, ..., N$ can be updated as

$$\mu_i^{(k+1)} = \{\sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)})P(t)\Sigma^{-1}(t_{i+1} - t_i - 1)\}^{-1}\{\sum_{t_1=1}^{n} ... \sum_{t_N=1}^{n} P(Y|t, \mu^{(k)})\times$$
$$P(t)\Sigma^{-1} \sum_{j=t_i+1}^{t_{i+1}} Y_{ij}\} \tag{5.21}$$

In (5.21), $P(Y|t, \mu^{(k)})$ can be expressed as

$$P(Y = (Y_0, Y_1, ..., Y_N)|t, \mu^{(k)} = (\mu_0^{(k)}, \mu_1^{(k)}, ...\mu_i^{(k)}, ..., \mu_N^{(k)})) = \prod_{i=0}^{N} (2\pi)^{-p/2}|\Sigma^{-1}|^{1/2}$$
$$exp\{-\frac{1}{2} \sum_{j=t_i+1}^{t_{i+1}} (Y_{ij} - \mu_i^{(k)})^T \Sigma^{-1}(Y_{ij} - \mu_i^{(k)})\} \tag{5.22}$$

and $P(t)$ is obtained using (3.18) in Chapter 3. So far, EM algorithm has been considered for parameter estimation without any constraint on parameters. In this section, constrained or restricted EM in the presence of change points will be reviewed first and then the above derived solution is generalised to a case where a set of constraints are to be satisfied.

5.2 Restricted EM in the Presence of Parameter Constraints

In this section, mean estimation problem is solved in the presence of change points and parameter constraints. Assume that in a network, the following constraints exist:

$$A\theta = a \tag{5.23}$$

where A is a known $h \times p$ matrix defining the constraints. Rank(A)=$h < p$ and a is a known $h \times 1$ vector. In the presence of restrictions, several methods exist in literature to solve the maximum likelihood estimation problem. In [55], a restricted maximum likelihood through a quadratic penalty function is used. On the other hand, Newton-Raphson iteration scheme is another way to solve the restricted EM problem. Two methods based on Newton-Raphson are proposed in [56] which are based on score function and information matrix. Lagrange multiplier approach is another approach to tackle the constrained EM problem. Applying Lagrange multiplier in M-step of EM results in a Lagrange function as

$$L(\theta, \lambda) = Q(\theta|\theta^{(k)}) + \lambda^T(a - A\theta) \tag{5.24}$$

where $\lambda = (\lambda_1, \lambda_2, ..., \lambda_h)$ are the coefficients defined for each constraint and Q-function is derived in E-step of the EM algorithm as explained in the previous section. The updated parameter can be obtained by finding the derivative of $L(\theta, \lambda)$ with respect to unknown parameters and λ. In other words, the following equations are to be solved to estimate the unknown parameters:

$$\nabla L(\theta, \lambda) = 0 \tag{5.25}$$

In the following, the mean estimation problem is solved through change point model structure using EM approach.

5.3 Mean Estimation for Change Point Detection in Presence of Constraints

In Section 5.1, the mean estimation problem was solved in the case of multiple shifts. Here, this problem is extended to constrained mean estimation problem using EM method. The problem is considered for a single shift detection through an illustrative example. The solution will then be extended to mean estimation in the presence of multiple change points.

5.3.1 Illustrative Example

Consider the process network shown in Figure 5.1 as in [7]. There are seven flow rates, $p = 7$, $y_t = (y_{1t}, y_{2t}, y_{3t}, y_{4t}, y_{5t}, y_{6t}, y_{7t})^T, t = 1, ..., n$ which must satisfy the mass balance constraints governing this network. Assuming no change occurs, the multivariate flow vector, y_t, consisting of y_{it}, $i = 1, ..., 7$ for $t = 1, ..., n$ can be represented by

$$y_t = \mu + \epsilon_t, \ , t = 1, ..., n \tag{5.26}$$

where $\mu = (\mu_1, ..., \mu_7)^T$ and $\epsilon = (\epsilon_1, ..., \epsilon_7)^T$ represent the multivariate mean and measurement noise respectively. The mean values, $\mu = (\mu_1, ..., \mu_7)^T$, should satisfy the network constraint as

$$A\mu = 0 \tag{5.27}$$

where A is a constant matrix related to process information. For the network in Figure 5.1, the mass balance constraints can be written as

$$\begin{aligned} \mu_1 + \mu_4 + \mu_6 - \mu_2 &= 0 \\ \mu_2 - \mu_3 &= 0 \\ \mu_3 - \mu_4 - \mu_5 &= 0 \\ \mu_5 - \mu_6 - \mu_7 &= 0 \end{aligned} \tag{5.28}$$

As a result, A can be written as

$$A = \begin{pmatrix} 1 & -1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix}$$

Assume that at an unknown time instant, $t = m$, a change occurs in the mean of data leading to

$$y_t = \mu' + \epsilon_t, \ , t = m + 1, ..., n \tag{5.29}$$

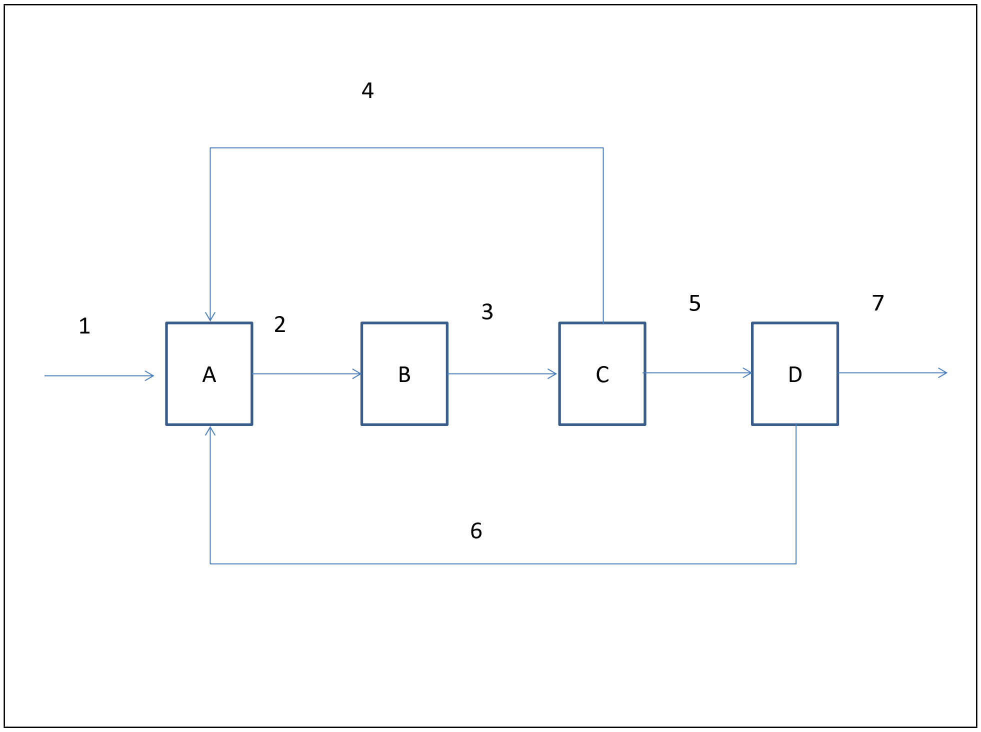

Figure 5.1: Process Network

where $\mu' = (\mu'_1, ..., \mu'_7)^T$ is the new mean vector which does not satisfy the constraint in (5.44), i.e., $A\mu' \neq 0$. Also, we assume that the covariance matrix of measurement noise remains constant. Therefore, the parameters to be estimated are $\theta = (\mu, \mu') = (\mu_1, ..., \mu_7, \mu'_1, ..., \mu'_7)$ under the constraints of (5.28). The Q-function has been derived in Section 5.1 for multiple change points. In the case of single change point, Q-function in E-step can be simplified as

$$Q((\mu, \mu')|(\mu^{(k)}, \mu'^{(k)})) = E_{m|Y,\mu^{(k)},\mu'^{(k)}}\{logP(Y, m|\mu, \mu')\} \tag{5.30}$$

Following similar approach as in Section 5.1 leads to a Q-function similar to (5.7) as

$$\begin{aligned} Q(\mu, \mu'|\mu^{(k)}, \mu'^{(k)}) = c_1 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)})P(m)logP(Y|m, \mu, \mu')] + \\ c_1 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)})P(m)logP(m)] \end{aligned} \tag{5.31}$$

where $c_1 = \frac{1}{P(Y|\mu^{(k)},\mu'^{(k)})}$. Having derived the Q-function, in M-step, we can write the lagrange multiplier for $h = 4$ constraints as

$$\begin{aligned} L(\mu, \mu', \lambda) &= L(\mu_1, ..., \mu_7, \mu'_1, ..., \mu'_7, \lambda_1, \lambda_2, \lambda_3, \lambda_4) \\ &= Q(\mu, \mu'|\mu^{(k)}, \mu'^{(k)}) + \lambda_1(\mu_1 + \mu_4 + \mu_6 - \mu_2) + \lambda_2(\mu_2 - \mu_3) + \lambda_3(\mu_3 - \mu_4 - \mu_5) + \\ &\lambda_4(\mu_5 - \mu_6 - \mu_7) \end{aligned} \tag{5.32}$$

Substituting (5.31) into (5.32) yields

$$L(\mu, \mu', \lambda) = c_1 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)}) P(m) logP(Y|m, \mu, \mu')] +$$
$$c_1 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)}) P(m) logP(m)] + \lambda_1(\mu_1 + \mu_4 + \mu_6 - \mu_2) + \lambda_2(\mu_2 - \mu_3) +$$
$$\lambda_3(\mu_3 - \mu_4 - \mu_5) + \lambda_4(\mu_5 - \mu_6 - \mu_7) \tag{5.33}$$

Since the observations are assumed to be independent with Gaussian distribution, $logP(Y|m, \mu, \mu')$ can be written as

$$logP(Y|m, \mu, \mu') = \frac{-np}{2} log(2\pi) - \frac{n}{2} log(|\Sigma|) - \frac{1}{2} \sum_{j=1}^{m} [(Y_j - \mu)^T \Sigma^{-1} (Y_j - \mu)] -$$
$$\frac{1}{2} \sum_{j=m+1}^{n} [(Y_j - \mu')^T \Sigma^{-1} (Y_j - \mu')] \tag{5.34}$$

Hence, we have

$$P(Y|m, \mu^{(k)}, \mu'^{(k)}) = (2\pi)^{-np/2} |\Sigma^{-1}|^{n/2} exp\{ -\frac{1}{2} \sum_{j=1}^{m} (Y_j - \mu^{(k)})^T \Sigma^{-1} (Y_j - \mu^{(k)}) +$$
$$\sum_{j=m+1}^{n} (Y_j - \mu'^{(k)})^T \Sigma^{-1} (Y_j - \mu'^{(k)}) \} \tag{5.35}$$

In (5.33), $P(m)$ is the prior distribution of the change instant. If we assume uniform distribution for m, as in [7], one can write $P(m) = c_2, m = 1, 2, ..., n-1$. To find the updated parameters, (5.25) is applied to (5.33) leading to the following sets of equations:

$$\frac{dL}{d\mu_1} = c_1 c_2 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)}) \frac{dlogP(Y|m, \mu, \mu')}{d\mu_1}] + \lambda_1 = 0 \tag{5.36}$$

$$\frac{dL}{d\mu_2} = c_1 c_2 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)}) \frac{dlogP(Y|m, \mu, \mu')}{d\mu_2}] - \lambda_1 + \lambda_2 = 0 \tag{5.37}$$

$$\frac{dL}{d\mu_3} = c_1 c_2 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)}) \frac{dlogP(Y|m, \mu, \mu')}{d\mu_3}] - \lambda_2 + \lambda_3 = 0 \tag{5.38}$$

$$\frac{dL}{d\mu_4} = c_1 c_2 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)}) \frac{dlogP(Y|m, \mu, \mu')}{d\mu_4}] + \lambda_1 - \lambda_3 = 0 \tag{5.39}$$

$$\frac{dL}{d\mu_5} = c_1 c_2 \sum_{m=1}^{n-1} [P(Y|m, \mu^{(k)}, \mu'^{(k)}) \frac{dlogP(Y|m, \mu, \mu')}{d\mu_5}] - \lambda_3 + \lambda_4 = 0 \tag{5.40}$$

$$\frac{dL}{d\mu_6} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu_6}] + \lambda_1 - \lambda_4 = 0 \quad (5.41)$$

$$\frac{dL}{d\mu_7} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu_7}] - \lambda_4 = 0 \quad (5.42)$$

$$\frac{dL}{d\mu'_1} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu'_1}] = 0 \quad (5.43)$$

$$\frac{dL}{d\mu'_2} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu'_2}] = 0 \quad (5.44)$$

$$\frac{dL}{d\mu'_3} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu'_3}] = 0 \quad (5.45)$$

$$\frac{dL}{d\mu'_4} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu'_4}] = 0 \quad (5.46)$$

$$\frac{dL}{d\mu'_5} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu'_5}] = 0 \quad (5.47)$$

$$\frac{dL}{d\mu'_6} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu'_6}] = 0 \quad (5.48)$$

$$\frac{dL}{d\mu'_7} = c_1c_2\sum_{m=1}^{n-1}[P(Y|m,\mu^{(k)},\mu'^{(k)})\frac{dlogP(Y|m,\mu,\mu')}{d\mu'_7}] = 0 \quad (5.49)$$

$$\frac{dL}{d\lambda_1} = \mu_1 + \mu_4 + \mu_6 - \mu_2 = 0 \quad (5.50)$$

$$\frac{dL}{d\lambda_2} = \mu_2 - \mu_3 = 0 \quad (5.51)$$

$$\frac{dL}{d\lambda_3} = \mu_3 - \mu_4 - \mu_5 = 0 \quad (5.52)$$

$$\frac{dL}{d\lambda_4} = \mu_5 - \mu_6 - \mu_7 = 0 \quad (5.53)$$

Thus, (5.36) to (5.42) can be written as

$$\begin{aligned} &- c_1c_2\sum_{m=1}^{n-1}[mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_1(\bar{Y}_{1,1:m},...,\bar{Y}_{7,1:m})^T]+ \\ &c_1c_2[\sum_{m=1}^{n-1} mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_1(\mu_1,...,\mu_7)^T] + \lambda_1 = 0 \quad (5.54) \\ &- c_1c_2\sum_{m=1}^{n-1}[mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_2(\bar{Y}_{1,1:m},...,\bar{Y}_{7,1:m})^T]+ \\ &c_1c_2[\sum_{m=1}^{n-1} mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_2(\mu_1,...,\mu_7)^T] - \lambda_1 + \lambda_2 = 0 \quad (5.55) \end{aligned}$$

$$
\begin{aligned}
&-c_1c_2\sum_{m=1}^{n-1}[mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_3(\bar{Y}_{1,1:m},...,\bar{Y}_{7,1:m})^T]+ \\
&c_1c_2[\sum_{m=1}^{n-1} mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_3(\mu_1,...,\mu_7)^T]-\lambda_2+\lambda_3=0 \qquad (5.56) \\
&-c_1c_2\sum_{m=1}^{n-1}[mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_4(\bar{Y}_{1,1:m},...,\bar{Y}_{7,1:m})^T]+ \\
&c_1c_2[\sum_{m=1}^{n-1} mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_4(\mu_1,...,\mu_7)^T]+\lambda_1-\lambda_4=0 \qquad (5.57)
\end{aligned}
$$

$$
\begin{aligned}
&-c_1c_2\sum_{m=1}^{n-1}[mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_5(\bar{Y}_{1,1:m},...,\bar{Y}_{7,1:m})^T]+ \\
&c_1c_2[\sum_{m=1}^{n-1} mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_5(\mu_1,...,\mu_7)^T]-\lambda_3+\lambda_4=0 \qquad (5.58) \\
&-c_1c_2\sum_{m=1}^{n-1}[mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_6(\bar{Y}_{1,1:m},...,\bar{Y}_{7,1:m})^T]+ \\
&c_1c_2[\sum_{m=1}^{n-1} mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_6(\mu_1,...,\mu_7)^T]+\lambda_1-\lambda_4=0 \qquad (5.59) \\
&-c_1c_2\sum_{m=1}^{n-1}[mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_7(\bar{Y}_{1,1:m},...,\bar{Y}_{7,1:m})^T]+ \\
&c_1c_2[\sum_{m=1}^{n-1} mP(Y|m,\mu^{(k)},\mu'^{(k)})\gamma_7(\mu_1,...,\mu_7)^T]-\lambda_4=0 \qquad (5.60)
\end{aligned}
$$

where $\gamma_i, i = 1, ..., 7$ are row vectors corresponding to the ith row in matrix Σ^{-1}. Also $\bar{Y}_{i,1:m}, i = 1, ..., 7$ is the mean of variable i from sample 1 to m. The equations (5.54) to (5.60) can be simplified further as

$$
\begin{aligned}
\gamma_1(\alpha+\beta(\mu_1,...,\mu_7)^T)+\lambda_1=0 \qquad (5.61) \\
\gamma_2(\alpha+\beta(\mu_1,...,\mu_7)^T)-\lambda_1+\lambda_2=0 \qquad (5.62) \\
\gamma_3(\alpha+\beta(\mu_1,...,\mu_7)^T)-\lambda_2+\lambda_3=0 \qquad (5.63) \\
\gamma_4(\alpha+\beta(\mu_1,...,\mu_7)^T)+\lambda_1-\lambda_4=0 \qquad (5.64) \\
\gamma_5(\alpha+\beta(\mu_1,...,\mu_7)^T)-\lambda_3+\lambda_4=0 \qquad (5.65) \\
\gamma_6(\alpha+\beta(\mu_1,...,\mu_7)^T)+\lambda_1-\lambda_4=0 \qquad (5.66) \\
\gamma_7(\alpha+\beta(\mu_1,...,\mu_7)^T)+\lambda_1-\lambda_4=0 \qquad (5.67)
\end{aligned}
$$

where α is a constant vector and β is a constant scalar as

$$\alpha = -c_1c_2\sum_{m=1}^{n-1}[mP(Y|m,\mu^{(k)},\mu'^{(k)})(\bar{Y}_{1,1:m},...,\bar{Y}_{7,1:m})^T] \tag{5.68}$$

$$\beta = c_1c_2[\sum_{m=1}^{n-1} mP(Y|m,\mu^{(k)},\mu'^{(k)})] \tag{5.69}$$

Similarly, (5.43) to (5.49) can be written as

$$\gamma_1(\eta+\zeta(\mu'_1,...,\mu'_7)^T)=0 \tag{5.70}$$

$$\gamma_2(\eta+\zeta(\mu'_1,...,\mu'_7)^T)=0 \tag{5.71}$$

$$\gamma_3(\eta+\zeta(\mu'_1,...,\mu'_7)^T)=0 \tag{5.72}$$

$$\gamma_4(\eta+\zeta(\mu'_1,...,\mu'_7)^T)=0 \tag{5.73}$$

$$\gamma_5(\eta+\zeta(\mu'_1,...,\mu'_7)^T)=0 \tag{5.74}$$

$$\gamma_6(\eta+\zeta(\mu'_1,...,\mu'_7)^T)=0 \tag{5.75}$$

$$\gamma_7(\eta+\zeta(\mu'_1,...,\mu'_7)^T)=0 \tag{5.76}$$

where η is a constant vector and ζ is a constant scalar as

$$\eta = -c_1c_2\sum_{m=1}^{n-1}[(n-m-1)P(Y|m,\mu^{(k)},\mu'^{(k)})(\bar{Y}_{1,m+1:n},...,\bar{Y}_{7,m+1:n})^T] \tag{5.77}$$

$$\zeta = c_1c_2[\sum_{m=1}^{n-1}(n-m-1)P(Y|m,\mu^{(k)},\mu'^{(k)})] \tag{5.78}$$

Therefore, the set of equations (5.61) to (5.67), (5.70) to (5.76) and (5.50) to (5.53), i.e. 18 equations, altogether need to be satisfied. Denote $\alpha = (\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6,\alpha_7)^T$, and $\eta = (\eta_1,\eta_2,\eta_3,\eta_4,\eta_5,\eta_6,\eta_7)^T$. Every row in Σ^{-1} as mentioned before can be expressed as

$$\begin{aligned}\gamma_1 &= (\gamma_{11},\gamma_{12},....,\gamma_{17})\\ \gamma_2 &= (\gamma_{21},\gamma_{22},....,\gamma_{27})\\ \gamma_3 &= (\gamma_{31},\gamma_{32},....,\gamma_{37})\\ &.\\ &.\\ \gamma_7 &= (\gamma_{71},\gamma_{72},....,\gamma_{77})\end{aligned} \tag{5.79}$$

The solutions to equations (5.61) to (5.67), (5.70) to (5.76) and (5.50) to (5.53) give μ_1^{k+1}, μ_2^{k+1},..., μ_7^{k+1}, $\mu_1'^{k+1}$,..., $\mu_7'^{k+1}$ as functions of α, β, ζ and η which are functions of current parameters estimated in the last iteration, μ_1^k, μ_2^k,..., μ_7^k, $\mu_1'^k$,...,$\mu_7'^k$. Thus, the equations (5.61) to (5.67) and (5.50) to (5.53) can be solved as

$$\begin{pmatrix} \mu_1^{(k+1)} \\ \mu_2^{(k+1)} \\ \mu_3^{(k+1)} \\ \mu_4^{(k+1)} \\ \mu_5^{(k+1)} \\ \mu_6^{(k+1)} \\ \mu_7^{(k+1)} \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} =$$

$$= \frac{1}{\beta} \times \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \gamma_{15} & \gamma_{16} & \gamma_{17} & \frac{1}{\beta} & 0 & 0 & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & \gamma_{25} & \gamma_{26} & \gamma_{27} & -\frac{1}{\beta} & \frac{1}{\beta} & 0 & 0 \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34} & \gamma_{35} & \gamma_{36} & \gamma_{37} & 0 & -\frac{1}{\beta} & \frac{1}{\beta} & 0 \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} & \gamma_{45} & \gamma_{46} & \gamma_{47} & \frac{1}{\beta} & 0 & -\frac{1}{\beta} & 0 \\ \gamma_{51} & \gamma_{52} & \gamma_{53} & \gamma_{54} & \gamma_{55} & \gamma_{56} & \gamma_{57} & 0 & 0 & -\frac{1}{\beta} & \frac{1}{\beta} \\ \gamma_{61} & \gamma_{62} & \gamma_{63} & \gamma_{64} & \gamma_{65} & \gamma_{66} & \gamma_{67} & \frac{1}{\beta} & 0 & 0 & -\frac{1}{\beta} \\ \gamma_{71} & \gamma_{72} & \gamma_{73} & \gamma_{74} & \gamma_{75} & \gamma_{76} & \gamma_{77} & 0 & 0 & 0 & -\frac{1}{\beta} \\ 1 & -1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}^{-1} \times$$

$$\begin{pmatrix} -\gamma_{11}\alpha_1 - \gamma_{12}\alpha_2 - \gamma_{13}\alpha_3 - ... - \gamma_{17}\alpha_7 \\ -\gamma_{21}\alpha_1 - \gamma_{22}\alpha_2 - \gamma_{23}\alpha_3 - ... - \gamma_{27}\alpha_7 \\ -\gamma_{31}\alpha_1 - \gamma_{32}\alpha_2 - \gamma_{33}\alpha_3 - ... - \gamma_{37}\alpha_7 \\ -\gamma_{41}\alpha_1 - \gamma_{42}\alpha_2 - \gamma_{43}\alpha_3 - ... - \gamma_{47}\alpha_7 \\ -\gamma_{51}\alpha_1 - \gamma_{52}\alpha_2 - \gamma_{53}\alpha_3 - ... - \gamma_{57}\alpha_7 \\ -\gamma_{61}\alpha_1 - \gamma_{62}\alpha_2 - \gamma_{63}\alpha_3 - ... - \gamma_{67}\alpha_7 \\ -\gamma_{71}\alpha_1 - \gamma_{72}\alpha_2 - \gamma_{73}\alpha_3 - ... - \gamma_{77}\alpha_7 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and the equations (5.70) to (5.76) can be written as

$$\begin{pmatrix} \mu_1^{'(k+1)} \\ \mu_2^{'(k+1)} \\ \mu_3^{'(k+1)} \\ \mu_4^{'(k+1)} \\ \mu_5^{'(k+1)} \\ \mu_6^{'(k+1)} \\ \mu_7^{'(k+1)} \end{pmatrix} = \frac{1}{\zeta}\Sigma \begin{pmatrix} -\gamma_{11}\eta_1 - \gamma_{12}\eta_2 - \gamma_{13}\eta_3 - ... - \gamma_{17}\eta_7 \\ -\gamma_{21}\eta_1 - \gamma_{22}\eta_2 - \gamma_{23}\eta_3 - ... - \gamma_{27}\eta_7 \\ -\gamma_{31}\eta_1 - \gamma_{32}\eta_2 - \gamma_{33}\eta_3 - ... - \gamma_{37}\eta_7 \\ -\gamma_{41}\eta_1 - \gamma_{42}\eta_2 - \gamma_{43}\eta_3 - ... - \gamma_{47}\eta_7 \\ -\gamma_{51}\eta_1 - \gamma_{52}\eta_2 - \gamma_{53}\eta_3 - ... - \gamma_{57}\eta_7 \\ -\gamma_{61}\eta_1 - \gamma_{62}\eta_2 - \gamma_{63}\eta_3 - ... - \gamma_{67}\eta_7 \\ -\gamma_{71}\eta_1 - \gamma_{72}\eta_2 - \gamma_{73}\eta_3 - ... - \gamma_{77}\eta_7 \end{pmatrix}$$

Starting μ_1^{k+1}, μ_2^{k+1},..., μ_7^{k+1}, $\mu_1'^{k+1}$,..., $\mu_7'^{k+1}$ with initial values, these update equations repeat until no changes in the estimated parameters are observed. The results show that the mean of data has changed at $t = 22$. The true mean before the change satisfies the constraints in (5.28) and is $\mu = [7, 9, 9, 1, 8, 1, 7]^T$. After $t = 22$, the true mean changes to $\mu' = [7.5, 10, 9, 1, 8, 1, 7]^T$. A total $n = 50$ samples are generated with the covariance matrix as

$$\Sigma = \begin{pmatrix} 1 & 0.05 & 0.02 & 0 & 0.03 & 0.05 & 0 \\ 0.05 & 1 & 0.01 & 0.01 & 0.01 & 0.15 & 0.05 \\ 0.02 & 0.01 & 0.4 & 0.01 & 0.16 & 0.01 & 0.2 \\ 0 & 0.01 & 0.01 & 0.8 & 0.01 & 0.01 & 0.2 \\ 0.03 & 0.01 & 0.16 & 0.01 & 0.5 & 0.01 & 0.01 \\ 0.05 & 0.15 & 0.01 & 0.01 & 0.01 & 1 & 0.02 \\ 0 & 0.05 & 0.2 & 0.2 & 0.01 & 0.02 & 0.5 \end{pmatrix}$$

The free parameters are chosen arbitrarily as $c_1 = 2, c_2 = 0.5$. The estimated results of constrained and unconstrained $\mu = (\mu_1, ..., \mu_7)$ and unconstrained $\mu' = (\mu_1', ..., \mu_7')$ with respect to iteration number are shown in Figures 5.2 and 5.3 respectively. As we can see, starting with some initial values, in 10 iterations, the mean values converge to the true ones indicating fast convergence of EM algorithm. The comparison of restricted μ values and unrestricted parameters shows the difference at convergence. Obviously, the constraints are satisfied before the mean shift occurs. In Table 5.1, the minimum, maximum, mean and standard deviation of estimation error of 14 estimated parameters are given for all iterations of EM. The estimation error is approximately of zero mean except for μ_3' which is 0.22. For μ_3', the true value is 9 while the estimated value is 9.2. In addition, the standard deviations of most of the parameters are smaller than 0.23 except μ_1' and μ_5' for which the standard error is about 0.5. These standard errors compared with true values are 10% , in the worst case.

This problem can also be generalized to the cases where covariance is unknown and changing. All the assumptions made in Chapters 3 and 4 can be used here and the only difference is addition of constraints to the first segment of the data, i.e before change occurs. In the presence of unknown covariance, one can add this parameter to the lists of unknowns and estimate the parameters at every iteration of EM.

In next section, a more realistic simulation example with nonlinear constrains is taken into account for mean estimation.

5.4 Example 2: Mean Estimation in Presence of Nonlinear Constraints

In this section, the mean estimation problem is extended to nonlinear constraints. The CSTR problem is a benchmark example in chemical engineering. This system was studied for the purpose of change point detection in previous chapters. The nonlinear dynamic

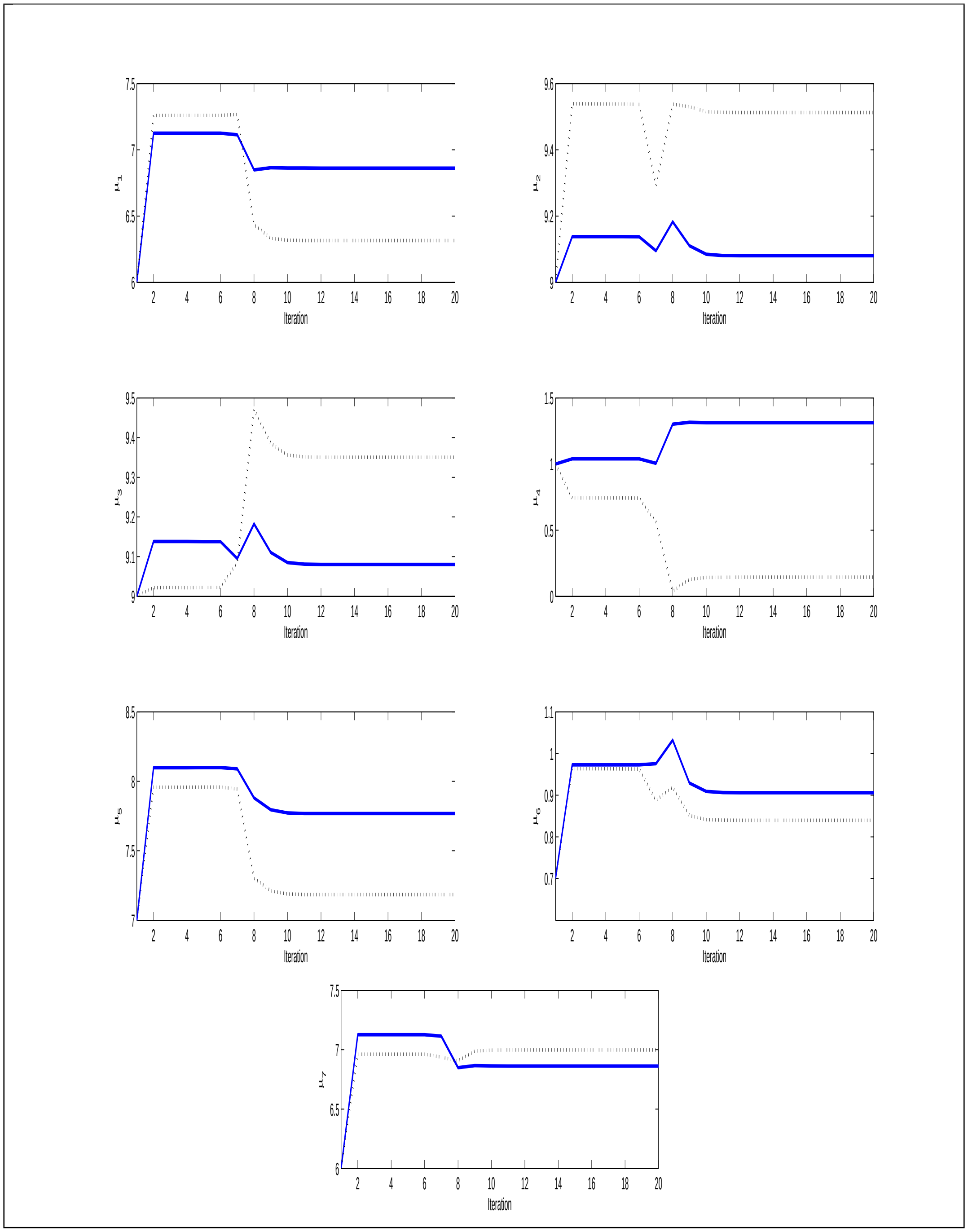

Figure 5.2: Estimated Mean Values Before the Change Point $\mu = (\mu_1, ..., \mu_7)$ Using EM, Unconstrained Solution (Dashed Line) and Constrained Solution (Solid Line)

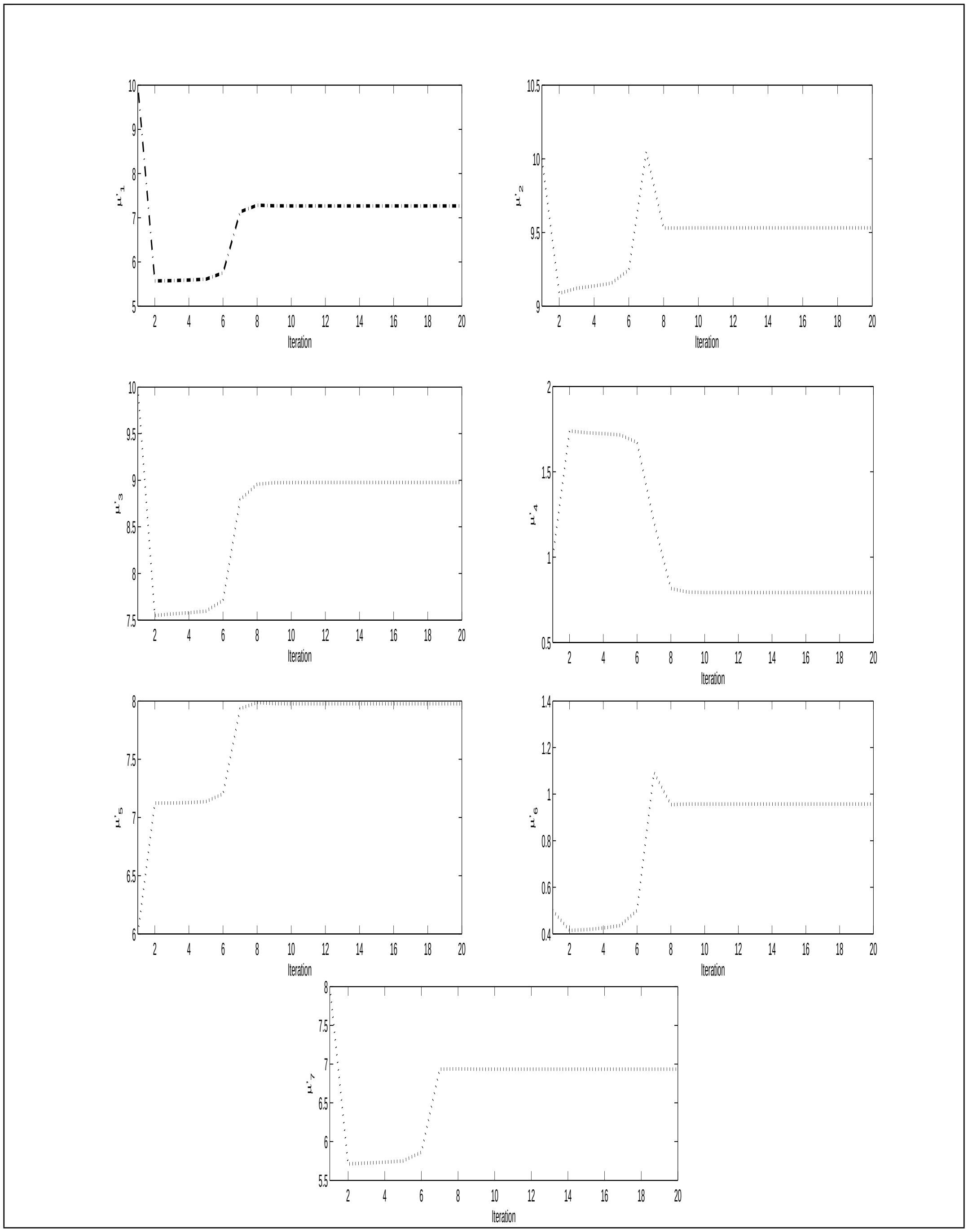

Figure 5.3: Estimated Mean Values (Unconstrained) After the Change Point $\mu' = (\mu'_1, ..., \mu'_7)$ Using EM

Table 5.1: The Estimation Error Characteristics Derived for 14 Parameters

Estimation Error	Min	Max	Mean	Std
Er(μ_1)	-1	0.0366	-0.0997	0.2112
Er(μ_2)	0	0.2814	0.0707	0.1049
Er(μ_3)	0	0.2814	0.0707	0.1049
Er(μ_4)	-0.0191	0.2145	0.0278	0.0930
Er(μ_5)	-1	0.0672	-0.0047	0.2283
Er(μ_6)	-0.3	0.1173	0.0807	0.0939
Er(μ_7)	-1	0.0366	-0.0997	0.2112
Er(μ'_1)	-0.1206	2.5	0.0676	0.5576
Er(μ'_2)	-0.2328	0	-0.0937	0.0621
Er(μ'_3)	0.1290	1	0.2200	0.2049
Er(μ'_4)	-0.0384	0.3712	0.0365	0.1425
Er(μ'_5)	-2	0.6021	-0.0118	0.5044
Er(μ'_6)	-0.5	0.1743	-0.1628	0.1501
Er(μ'_7)	-0.5	0.1743	-0.1628	0.1501

equations of CSTR can be written as

$$\dot{C_A}(t) = \frac{q(t)}{V}(C_{A0}(t) - C_A(t)) - k_0 C_A(t) exp(-\frac{E}{RT(t)}) \tag{5.80}$$

$$\begin{aligned} \dot{T}(t) = \frac{q(t)}{V}(T_0(t) - T(t)) + \frac{\Delta H k_0 C_A(t)}{\rho C_p} exp(-\frac{E}{RT(t)}) + \\ \frac{\rho_c C_{pc}}{\rho C_p V} q_c(t)\{1 - exp(\frac{-hA}{q_c(t)\rho C_p})\}(T_{c0} - T(t)) \end{aligned} \tag{5.81}$$

As discussed in previous chapters, the input is q_c which drives the process to different operating modes. In Chapter 3, the process input is changed and corresponding to that change, the system outputs, i.e. product concentration and temperature, could change accordingly. In this chapter, however, the objective is to estimate the mean of outputs before and after the change point assuming that the change point is unknown. The difference between mean estimation in this section and the one in previous section, i.e. linear process network example, is that here the measurements must satisfy the nonlinear constraints not only before the change but also after the change. The outputs are corrupted by measurement noise. The measurement noise added to each output is of Gaussian distribution with mean zero and standard deviation 10% of the states. The input is selected as $q_c = 97\ L/min$ before the change occurs and then it changes to $q_c = 97\ L/min$. The mean estimation using EM algorithm in presence of constraints can be solved using Lagrange multiplier. Under the steady state condition, the constraints are of the form

$$f_1 = \frac{q(t)}{V}(C_{A0}(t) - C_A(t)) - k_0 C_A(t) exp(-\frac{E}{RT(t)}) = 0 \tag{5.82}$$

$$\begin{aligned} f_2 = \frac{q(t)}{V}(T_0(t) - T(t)) + \frac{\Delta H k_0 C_A(t)}{\rho C_p} exp(-\frac{E}{RT(t)}) + \\ \frac{\rho_c C_{pc}}{\rho C_p V} q_c(t)\{1 - exp(\frac{-hA}{q_c(t)\rho C_p})\}(T_{c0} - T(t)) = 0 \end{aligned} \tag{5.83}$$

Assume that before the change point we have $\mu = [C_A, T]^T$ and after the change occurs, the mean vector is $\mu' = [C'_A, T']^T$. The Lagrange function can be written as

$$\begin{aligned} L(\mu, \mu', \lambda) &= L(C_A, T, C'_A, T', \lambda_1, \lambda_2) \\ &= Q(\mu, \mu' | \mu^{(k)}, \mu'^{(k)}) + \lambda_1(f_1) + \lambda_2(f_2) \end{aligned} \tag{5.84}$$

Differentiating the Lagrange function with respect to μ and setting them equal to zero, we have

$$\frac{dL}{dC_A} = \frac{dQ}{dC_A} + \lambda_1 \frac{df_1}{dC_A} + \lambda_2 \frac{df_2}{dC_A} = 0 \tag{5.85}$$

$$\frac{dL}{dT} = \frac{dQ}{dT} + \lambda_1 \frac{df_1}{dT} + \lambda_2 \frac{df_2}{dT} = 0 \tag{5.86}$$

These equations can be further simplified as

$$\lambda_1 A_{11} + \lambda_2 A_{12} = -\gamma_1(\alpha + \beta(C_A, T)^T) \tag{5.87}$$

$$\lambda_1 A_{13} + \lambda_2 A_{14} = -\gamma_2(\alpha + \beta(C_A, T)^T) \tag{5.88}$$

where

$$\frac{dQ}{dC_A} = \gamma_1(\alpha + \beta(C_A, T)^T) \tag{5.89}$$

$$\frac{dQ}{dT} = \gamma_2(\alpha + \beta(C_A, T)^T) \tag{5.90}$$

$\gamma_1 = (\gamma_{11}, \gamma_{12})$ corresponds to the first row of Σ^{-1}, the measurement covariance matrix, and $\gamma_2 = (\gamma_{21}, \gamma_{22})$ corresponds to the second row of Σ^{-1}. α and β are the same as equations (5.68) and (5.69). Also,

$$\begin{aligned} A_{11} &= \frac{df_1}{dC_A} = -\frac{q(t)}{V} - k_0 exp(-\frac{E}{RT(t)}) \\ A_{12} &= \frac{df_1}{dT} = \frac{\Delta H k_0}{\rho C_p} exp(-\frac{E}{RT(t)}) \\ A_{13} &= \frac{df_2}{dC_A} = -k_0 C_A(t)(\frac{E}{RT^2}) exp(-\frac{E}{RT(t)}) \\ A_{14} &= \frac{df_2}{dT} = -\frac{q(t)}{V} + \frac{\Delta H k_0 C_A(t)}{\rho C_p}(\frac{E}{RT^2}) exp(-\frac{E}{RT(t)}) - \frac{\rho_c C_{pc}}{\rho C_p V} q_c(t)\{1 - exp(\frac{-hA}{q_c(t)\rho C_p})\} \end{aligned} \tag{5.91}$$

Thus, in order to find the mean vector before the change point, one needs to solve four equations (5.82), (5.83), (5.85) and (5.86) with respect to C_A, T,λ_1 and λ_2 by selecting $q_c = 97$.

In order to find the mean vector after the change point, the same equations are solved with respect to C'_A, T' and new coefficients defined as λ'_1 and λ'_2 but with $q_c = 99$ as

$$\begin{aligned}\lambda'_1 A'_{11} + \lambda'_2 A'_{12} &= -\gamma_1(\eta + \zeta(C'_A, T')^T)\\ \lambda'_1 A'_{13} + \lambda'_2 A'_{14} &= -\gamma_2(\eta + \zeta(C'_A, T')^T)\\ f_1 &= 0\\ f_2 &= 0 \qquad (5.92)\end{aligned}$$

where η and ζ can be derived as in (5.77) and (5.78). A'_{11}, A'_{12}, A'_{13} and A'_{14} are derived as (5.91) but with $q_c = 99$ as the input value.

As we can see, since the equations are nonlinear, at every iteration of EM, we need to solve 8 nonlinear equations. The solutions will be nonlinear functions of α, β, η and ζ which all contain the current mean vectors $\mu^{(k)}, \mu'^{(k)}$. Using $fsolve$ function, these nonlinear functions are evaluated in Matlab. In Figure 5.4, the measurements are shown. The free parameters are chosen randomly as $c_1 = 2, c_2 = 0.1$. The initial values are set as $\mu^{(0)} = [C_A, T]^T = [0.001, 700]^T$,$\mu'^{(0)} = [C'_A, T']^T = [0.002, 550]^T$. The true mean values before the mean shift are $\mu^{(true)} = [C_A, T]^T = [0.0799, 436.9345]^T$ and after the change point, the means are changed to $\mu'^{(true)} = [C'_A, T']^T = [0.0846, 443.4214]^T$. The covariance matrix can be calculated using the measurements.

The mean estimates for concentration, C_A and temperature, T before and after the change point are illustrated in Figure 5.5 and 5.6 respectively.

As we can see, even if the initial condition is far from the true mean values, in the second iteration, EM converges to true mean values. From these Figures, the estimated mean before the change point is $\mu^{(*)} = [C_A, T]^T = [0.0793, , 443.5109]^T$ and the mean after the change point is $\mu'^{(*)} = [C'_A, T']^T = [0.0851, 441.9877]^T$. Apparently, the estimated means are very close to true values as mentioned before.

Similarly, in the presence of multiple change points, we have more nonlinear equations representing the process constraints at every operational mode. Corresponding to every change point we need to solve additional four nonlinear equations to determine the mean.

5.4.1 Conclusion

In this chapter, the mean estimation in change point detection problem is investigated. Through Expectation Maximization (EM) method, the means of data before and after the change points are estimated without knowing where the change point is. This problem is also extended to a more complex case where these mean values, i.e. unknown parameters, must satisfy certain constraints. Using Lagrange multiplier, this problem is solved for a process

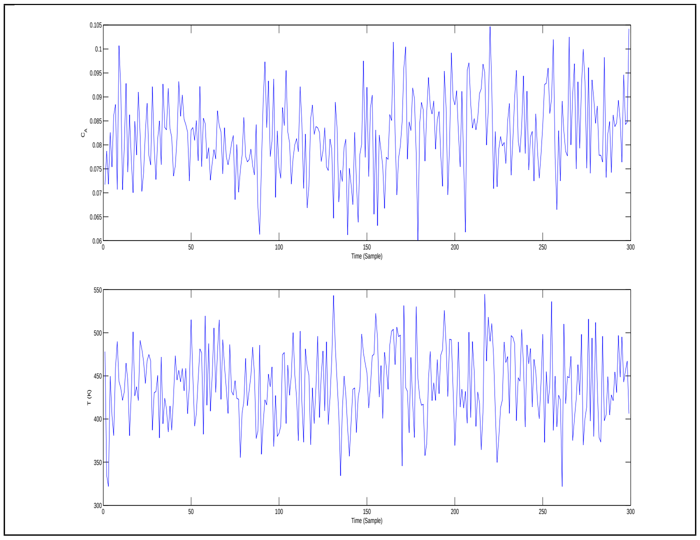

Figure 5.4: Product Concentration (C_A) and Reactor Temperature (T)

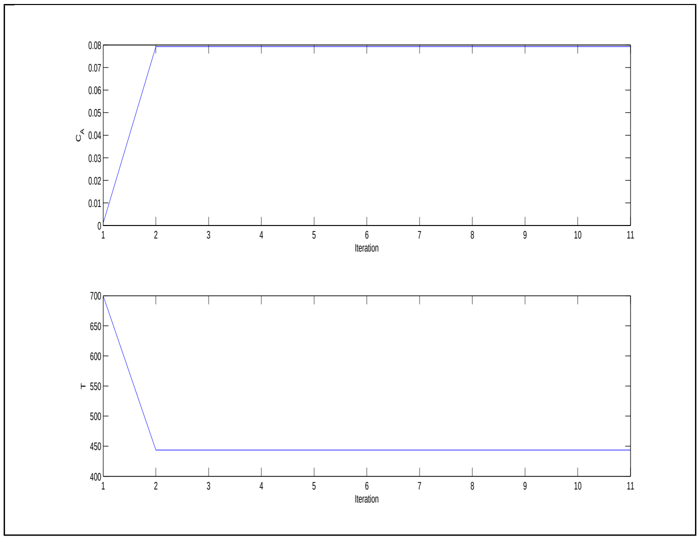

Figure 5.5: Mean Estimation for Product Concentration (C_A, Upper Plot) and Reactor Temperature (T, Lower Plot) Before the Change Point

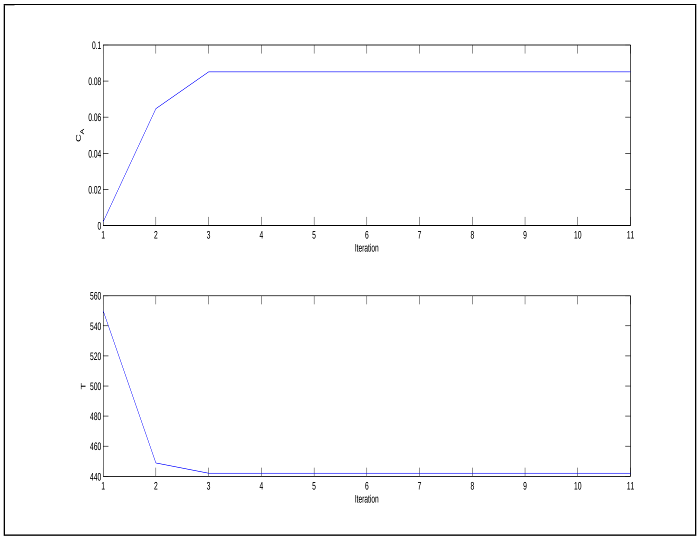

Figure 5.6: Mean Estimation for Product Concentration (C_A, Upper Plot) and Reactor Temperature (T, Lower Plot) After the Change Point

network and a more realistic problem in chemical engineering. The results demonstrate successful estimation of unknown parameters in the presence of unknown change points.

Chapter 6

Conclusion and Directions for Future Work

6.1 Summary and Conclusion

This thesis is focused mostly on change point detection under various scenarios. Change point detection can be found in different applications including hydrology, signal processing, finance, economics, pharmacology, environmental studies, meteorology and etc. Detection of changes can assist us in various ways; one may use these techniques in identification of bias in instruments in the form of mean shift detection. One application of change detection is in identification of process variability. Another application is in process abnormality detection and also detection of mode of operation. Depending on the application, one can take advantage of these techniques to preprocess or analyse the data. In this thesis, EM approach is used for change point detection under various scenarios. The reason is, EM is an alternative solution to maximum likelihood estimation. In this method, defining new sets of variables as hidden variables, one can find their expectation using the data and current estimate of unknown parameters and then maximize this expectation with respect to unknown parameters. It is proved that at every iteration of EM, the likelihood of observed data increases. It was also shown that compared with Bayesian inference, EM is not sensitive to priors. Like other optimisation techniques, EM can be also sensitive to initial selection of unknown parameters resulting in local maxima.

Having introduced the motivation and background for change point detection problem in Chapter 1, in Chapter 2, this problem is solved for univariate data and single change in the mean of data. In fact, the problem is solved using three approaches including Bayesian, EM and SEM. The performance of these methods are compared through simulation example and the overall power of algorithms in correct identification of change point is calculated through Monte Carlo runs.

In Chapter 3, change point detection was solved for multivariate data in the presence of known covariance matrix. Single and multiple changes detection problems are solved using Bayesian and EM methods. Through simulation and experimental data, the performance of proposed algorithms are evaluated. EM algorithms outperforms Bayesian method in the case of wrong priors. Convergence of EM is fast in terms of the number of iterations required.

In Chapter 4, change point detection is further extended to unknown and changing covariance matrix. EM solution for this problem is derived and through simulation and experimental data, the detection performance is studied. The results show successful performance of the proposed method.

In any parameter estimation, problem, there may be constraints on unknown parameters. In Chapter 5, these process constraints are taken into account. The mean estimation in multivariate data is elaborated through EM algorithm using change point models. The update equations for parameters in multiple changes are derived. As an extension of this derivation, the problem was also solved through EM imposing constraints such as mass balance. The performance was evaluated using an illustrative example of process network. The results show that in the presence of process constraints, EM convergent is fast and the estimation is accurate.

6.2 Directions for Future Work

Throughout this thesis, we assume that the number of change point is already known as priori. Extending EM formulation to the case where the number of change point can be determined during iterations of EM can be regarded as future direction. The estimation of the number of change points can be solved for different assumptions for mean and covariance of data.

Another direction can be online detection of change points without adding much more complexity at each sample instant compared with [23]. Derivation of EM solution to change point detection problem where there is dependency or correlation in observations can also be regarded as another future work.

Bibliography

[1] S. Narasimhan, C. Jordache, Data Reconciliation and Gross Error Detection: An Intelligent Use of Process Data, Houston, TX:Gulf Publishing Company, 2000.

[2] M. Basseville, I. V. Nikiforov, Detection of Abrupt Changes: Theory and Application, Prentice-Hall, 1993.

[3] D. Lu, P. Mausel, E.Brondizio, E. Moran, Change Detection Techniques, International Journal of Remote Sensing, Vol. 25, No.12, PP. 2365-2407, 2004.

[4] J. Reeves, J. Chen, X. L. wang, R. Lund, Q. Lu A Review and Comparison of Change Point Detection Techniques for Climate Data, Journal of Applied Meteorology and Climatology, Vol.46, PP. 900-915, 2007.

[5] A. C. Tamhane, C. Iordache, R. S. H. Mah, A Bayesian Approach to Gross Error Detection in Chemical Process Data. Part I: Model Development, Chemometrics and Intelligent Laboratory Systems, Vol. 4, PP. 33-45, 1988.

[6] A. C. Tamhane, C. Iordache, R. S. H. Mah, A Bayesian Approach to Gross Error Detection in Chemical Process Data. Part II: Simulation Results, Chemometrics and Intelligent Laboratory Systems, Vol. 4, PP. 131-146, 1988.

[7] S. Devanathan, S. B. Vardeman, D. K. Rollins, Likelihood and Bayesian Methods for Accurate Identification of Measurement Biases in Pseudo Steady-State Processes, Chemical Engineering Research and Design, Vol. 83, PP. 1391-1398, 2005.

[8] L. Perreault, E. Parent, J. Bernier, B. Bobee, M. Slivitzky, Retrospective Multivariate Bayesian Change-Point Analysis: A Simultaneous Single Change in the Mean of Several Hydrological Sequences, Stochastic Environmental Reseach and Risk Assessment, Vol. 14, PP. 243-261, 2000.

[9] L. Perreault, E. Parent, J. Bernier, B. Bobee, M. Slivitzky, Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited, Journal of Hydrology, Vol. 235, PP. 221-241, 2000.

[10] A. M. Djafari, O. Feron, Bayesian Approach to Change Point Detection in Time Series, Wiley Periodicals Inc., Vol. 16, PP. 215-221, 2007.

[11] K. D. Zambadouglas, M. Hawkins, A Multivariate Change Point Model for Statistical Process Control, Technometrics, Vol. 48, No. 4, PP. 539-549, 2006.

[12] Y. S. Son, S.W. Kim, Bayesian Single Change Point Detection in a Sequence of Multivariate Normal Observations, Statistics: A Journal of Theoretical and Applied Statistics, Vol. 39, No.5, PP.373-387, 2005.

[13] R. J. Karunamuni, S. Zhang, Empirical Bayes Detection of a Change in Distribution, Ann. Inst. Statis. Math., Vol. 48, No. 2, pp. 229-246, 1996.

[14] Venter, J.H., Steel, S.J., Finding Multiple Abrupt Change Points. Computational Statistics and Data Analysis Vol. 22, PP.481-504, 1996.

[15] Hawkins, D.M., Finding Multiple Change Point Models to Data. Computational Statistics and Data Analysis, Vol. 37, PP. 323-341, 2001.

[16] Barry D., Hartigan J.A., A Bayesian Analysis for Change Point Problems. Journal of the American Statistical Association Vol.88, PP.309-319, 1993.

[17] Crowley, E.M., Product Partition Models for Normal Means. Journal of the American Statistical Association, Vol. 92, PP. 192-198, 1997.

[18] M. Lavielle, Detection of Multiple Change-Points in Multivariate Time Series, Lithuanian Mathematical Journal, Vol. 46, No. 3, pp. 287-306, 2006.

[19] S. Cheon, J. Kim, Multiple Change-Point Detection of Multivariate Mean Vectors With the Bayesian Approach, Computational Statistics and Data Analysis, Vol. 54, PP. 406-415, 2010.

[20] R.H. Loschi, F.R.B. Cruz, R.H.C. Takahashi, P.L. Iglesias, R. B. Arellano, J. MacGregor Smith, A Note on Bayesian Identification of Change Points in Data Sequence, Computers and Operations Research, Vol. 35, PP. 156- 170, 2008.

[21] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian Data Analysis, Chapman and Hall/CRC, 2004.

[22] G. J. MacLachlan, T. Krishnan, The EM algorithm and extensions, John Wiley and Sons, 1997.

[23] S. Yildirim, S. S. Singh, A. Doucet, An Online Expectation-Maximization Algorithm for Change point Models, Journal of Computational and Graphical Statistics, In press.

[24] N.K. Bansal, H. Du, G. Hamedani, An Application of EM Algorithm to Change Point Problem, Communication in Statistics-Theory and Methods, Vol. 37, PP. 2010-2021, 2008.

[25] Statisticat LLC, Bayesian Inference, Statistical Consulting and Support, Available at: http://www.bayesian-inference.com

[26] A.P. Dempster, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society,, Vol. 39, No. 1, PP. 1-38, 1977.

[27] X.L. Meng, On the rate of convergence of the ECM algorithm, Annals of Statistics, Vol. 22, PP. 326-339, 1994.

[28] R.J.A. Little and D.B. Rubin, Statistical Analysis with Missing Data, Wiley, New York, second edition, 2002.

[29] C.F.J. Wu, On the convergence properties of the EM algorithm, Annals of Statistics, Vol. 11, PP. 95-103, 1983.

[30] D. Nettleton, Convergence properties of the EM algorithm in constrained parameter spaces, Canadian Journal of Statistics, Vol. 27, PP. 639-648, 1999.

[31] M.J. Lindstrom , D.M. and Bates, Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data, Journal of the American Statistical Association, Vol. 83, PP. 1014-1022, 1988.

[32] Bohning D., Dietz E., Schaub R., Schlattmann P. and Lindsay B., The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family, Annals of the Institute of Statistical Mathematics, Vol. 46, PP.373- 388, 1994.

[33] X. Xuan, Bayesian Inference on Change Point Problems, M.Sc. Thesis in Computer Science, University of British Colombia, March 2007.

[34] M. Keshavarz, B. Huang, Expectation Maximization Approach to Gross Error and Change Point Detection, $10th$ IEEE International Conference on Control and Automation, June 12-14, Hangzhou, China, 2013.

[35] D. Karlis, E. Xekalaki, Choosing Initial Values for the EM Algorithm for Finite Mixtures, Computational Statistics and Data Analysis, Vol. 41, PP. 577-590, 2003.

[36] Leroux, B.G., Consistent Estimation of a Mixing Distribution, Ann. Statist., Vol. 20, PP. 1350- 1360, 1992.

[37] Woodward, W., Parr, W., Schucany, R., Lindsey, H., A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion, J. Amer. Statist. Assoc, Vol. 79, PP. 590- 598, 1984.

[38] McLachlan, G.J., On the choice of initial values for the EM algorithm in fitting mixture models. The Statistician Vol. 37, PP. 417- 425, 1988.

[39] X. Jin. Multiple ARX Model Based Identication for Switching Nonlinear Systems with EM Algorithm. Masters thesis, University of Alberta, 2010.

[40] K. Tsuda, D. Mignone, G. Ferrari-Trecate and M. Morari. Reconfiguration Strategies for Hybrid Systems, Proceedings of the American Control Conference, Arlington, VA June 25-27, 2001.

[41] M. Keshavarz, B. Huang, Bayesian and Expectation Maximization Methods for Multivariate Change Point Detection, submitted to Computers and Chemical Engineering Journal, 2013.

[42] D.R. Kuehn, H. Davidson, Computer Control II, Mathematics of Control, Chemical Engineering Progress, vol 57, pp.44-47, 1961.

[43] C.M. Crow, Y.A. Garcia Campos, A. Hyrmak, Reconciliation of Process Flow Rates by Matrix Projection. Part I. Linear Case, AIChE Journal, vol. 29, No. 6, pp. 881-888, 1983.

[44] J.C. Knepper, J.W. Gorman, Statistical Analysis of Constrained Data Sets, AIChE Journal, vol. 26, pp.260-264, 1980.

[45] S. Narasimhan, P. Harikumar, A Method to Incorporate Bounds in Data Reconciliation and Gross Error Detection. The Bounded Data Reconciliation Problem, Computers and Chemical Engineering, vol 17, No.11, pp.1115-1120, 1993.

[46] I.B. Tjoa, L.T. Biegler, Simultaneous Strategies for Data Reconciliation and Gross Error Detection of Nonlinear Systems, Computers and Chemical Engineering, vol. 15, No. 10, pp.679-690, 1991.

[47] D.B. Ozyurt, R.W. Pike, Theory and Practice of Simultaneous Data Reconcilation and Gross Error Detection for Chemical Process, Computers and Chemical Engineering, vol. 28,2004.

[48] R. Serth, W. Heenan, Gross Error Dectection and Data Reconciliation in Steam-Metering Systems, AIChE Journal, vol. 32, pp. 733-742, 1986.

[49] I.W. Kim, M.S. Park, T.F. Edgar, Robust Data Reconciliation and Gross error Detection: the Modified MIMT Using NLP, Computers and Chemical Engineering, vol. 21, No. 7, pp. 775-782, 1997.

[50] W. Wongrat, T. Srinophakun, P. Srinophakun,Modified Genetic Algorithm for Nonlinear Data Reconciliation, vol.29, pp.1059-1067, 2005.

[51] D.M. Prata,M.Schwaab,E.D. Lima,J.N.Pinto, Nonlinear Dynamic Data Reconciliation and Parameter Estimation Through Particle Swarm Optimization:Application for an Industrial Polypropylene Reactor. Chem. Eng. Sci.,vol. 64, pp.3953-3967,2009.

[52] D.M. Prata,,J.N.Pinto,E.D. Lima, Simultaneous Robust Data Reconciliation and Gross Error Detection Through Particle Swarm Optimization for an Industrial Polypropylene Reactor, Chem. Eng. Sci., vol. 65, pp. 4943-4954, 2010.

[53] E.D. Valdetaro,R. Schirru, simutaneous Model selection, Robust Data Reconciliation and Outlier Detection with Swarm Intelligence in a Thermal Reactor Power Calculation, Annals of Nuclear Energy, vol. 38, pp.1820-1832, 2011.

[54] H. Pakravesh, A. Shojaei, Optimization of Industrial CSTR for Vinyl Acetate Polymerization Using Novel shuffled Frog Leaping Based Hybrid Algorithms and Dynamic Modeling,Computers and Chemical Engineering,vol. 35, pp.2351-2365, 2011.

[55] H. Nyquist, Restricted Estimation of Genelalized Linear Models, Applied Statistics, Vol. 40, PP.133-141, 1991.

[56] D.K. Kim, J.M.G Taylor, The Restricted EM Algorithm for Maximum Likelihood Estimation Under Linear Constraint Restrictions on the Parameters, Journal of American Statistical Association, Vol. 90, No. 430, PP. 708-716, 1995.

.1 Appendix

In Bayesian derivation, the joint probability in (3.10) is proportional to the posterior probability of change point given the data. This joint probability can be written as the product of two terms, F_1 and F_2. In the following, the Bayesian solution is given. We have

$$P(D, \mu_1, \mu_2, \Sigma, \beta_0, m) = F_1 \times F_2 \tag{1}$$

where

$$F_1 = C_1 exp\{-\frac{1}{2}\{\sum_{i=1}^{m}[(y_i - \mu_1)^T \Sigma^{-1}(y_i - \mu_1)] + (\mu_1 - \mu_1^0)^T \Sigma_{01}^{-1}(\mu_1 - \mu_1^0)\}\} \tag{2}$$

$$F_2 = C_2 exp\{-\frac{1}{2}\{\sum_{i=m+1}^{n}[(y_i - \mu_2)^T \Sigma^{-1}(y_i - \mu_2)] + (\mu_2 - \mu_2^0)^T \Sigma_{02}^{-1}(\mu_2 - \mu_2^0)\}\} \tag{3}$$

We can write

$$\begin{aligned}
F_1 &= C_1 exp\{-\frac{1}{2}\{\sum_{i=1}^{m}[(y_i - \mu_1)^T \Sigma^{-1}(y_i - \mu_1)] + (\mu_1 - \mu_1^0)^T \Sigma_{01}^{-1}(\mu_1 - \mu_1^0)\}\} \\
&= C_1 exp\{-\frac{1}{2}\left(\mu_1^T\left(m\Sigma^{-1} + \Sigma_{01}^{-1}\right)\mu_1 - \mu_1^T\left(m\Sigma^{-1}\bar{y} + \Sigma_{01}^{-1}\mu_1^0\right) - \left(m\Sigma^{-1}\bar{y} + \Sigma_{01}^{-1}\mu_1^0\right)^T\mu_1\right)\}\times \\
&exp\{-\frac{1}{2}\{{\mu_1^0}^T \Sigma_{01}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i\}\}
\end{aligned} \tag{4}$$

where

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i \tag{5}$$

In order to complete the square term, define $A = m\Sigma^{-1} + \Sigma_{01}^{-1}$ and $B = m\Sigma^{-1}\bar{y} + \Sigma_{01}^{-1}\mu_1^0$. We have

$$F_1 = C_1 exp\{-\frac{1}{2}\left(\mu_1^T A\mu_1 - \mu_1^T B - B^T\mu_1\right)\} * exp\{-\frac{1}{2}({\mu_1^0}^T \Sigma_{01}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i)\} \tag{6}$$

Further algebraic manipulation yields

$$\begin{aligned}
F_1 = C_1 exp\{-\frac{1}{2}&\left(\mu_1^T A\mu_1 - \mu_1^T B - B^T\mu_1 + B^T A^{-1}B - B^T A^{-1}B\right)\}\times \\
&exp\{-\frac{1}{2}({\mu_1^0}^T \Sigma_{01}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i)\} \\
&= C_1 exp\{-\frac{1}{2}\left(\mu_1^T A\mu_1 - \mu_1^T B - B^T\mu_1 + B^T A^{-1}B\right)\} * \\
&exp\{-\frac{1}{2}({\mu_1^0}^T \Sigma_{01}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T \Sigma^{-1} y_i - B^T A^{-1}B)\}
\end{aligned} \tag{7}$$

Note that A is the weighted sum of two symmetric and full rank covariance matrices; thus A is symmetric and invertible. We have $I = A^{-1}A = AA^{-1}$ and (A.7) can be further rewritten as

$$F_1 = C_1 exp\{-\frac{1}{2}\left(\mu_1^T A\mu_1 - \mu_1^T AA^{-1}B - B^T A^{-1}A\mu_1 + B^T A^{-1}AA^{-1}B\right)\}\times$$

$$exp\{-\frac{1}{2}(\mu_1^{0^T}\Sigma_{01}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T\Sigma^{-1}y_i - B^T A^{-1}B)\} \tag{8}$$

Let $\Delta_n = A^{-1}$ and $\Omega_n = A^{-1}B$, then (A.8) can be written as

$$F_1 = C_1 exp\{-\frac{1}{2}\left(\mu_1^T\Delta_n^{-1}\mu_1 - \mu_1^T\Delta_n^{-1}\Omega_n - \Omega_n^T\Delta_n^{-1}\mu_1 + \Omega_n^T\Delta_n^{-1}\Omega_n\right)\}*$$

$$exp\{-\frac{1}{2}(\mu_1^{0^T}\Sigma_{01}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T\Sigma^{-1}y_i - B^T A^{-1}B)\}$$

$$= C_1 exp\{-\frac{1}{2}\left(\mu_1 - \Omega_n\right)'\Delta_n^{-1}\left(\mu_1 - \Omega_n\right)\} * exp\{-\frac{1}{2}(\mu_1^{0^T}\Sigma_{01}^{-1}\mu_1^0 + \sum_{i=1}^{m} y_i^T\Sigma^{-1}y_i - B^T A^{-1}B)\} \tag{9}$$

So, F_1 is derived as in (A.8) where $\Delta_n = A^{-1} = (m\Sigma^{-1} + \Sigma_{01}^{-1})^{-1}$ and $\Omega_n = A^{-1}B = (m\Sigma^{-1} + \Sigma_{01}^{-1})^{-1}(m\Sigma^{-1}\bar{y} + \Sigma_{01}^{-1}\mu_1^0)$.
Following the same procedure, F_2 is derived as

$$F_2 = C_2 exp\{-\frac{1}{2}\left(\mu_2 - \Psi_n\right)'\Lambda_n^{-1}\left(\mu_2 - \Psi_n\right)\} \times exp\{-\frac{1}{2}(\mu_2^{0^T}\Sigma_{02}^{-1}\mu_2^0 + \sum_{i=m+1}^{n} y_i^T\Sigma^{-1}y_i - D^T C^{-1}D)\} \tag{10}$$

where $\bar{y} = \frac{1}{n-m}\sum_{i=m+1}^{n} y_i$, $C = (n-m)\Sigma^{-1} + \Sigma_{02}^{-1}$, $D = (n-m)\Sigma^{-1}\bar{y} + \Sigma_{02}^{-1}\mu_2^0$, $\Lambda_n = C^{-1} = ((n-m)\Sigma^{-1} + \Sigma_{02}^{-1})^{-1}$ and $\Psi_n = C^{-1}D = ((n-m)\Sigma^{-1} + \Sigma_{02}^{-1})^{-1}((n-m)\Sigma^{-1}\bar{y} + \Sigma_{02}^{-1}\mu_2^0)$.
As a result, the joint probability distribution is

$$P(D, \mu_1, \mu_2, \Sigma, \beta_0, m) = F_1 \times F_2 \tag{11}$$