Spring 2017

Applications of outlier and change detection for longitudinal data

Yuxing Hou
University of Iowa

APPLICATIONS OF OUTLIER AND CHANGE DETECTION FOR LONGITUDINAL DATA

by

Yuxing Hou

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Industrial Engineering in the
Graduate College of
The University of Iowa

May 2017

Thesis Supervisor: Professor Yong Chen

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Yuxing Hou

has been approved by the Examining Committee for
the thesis requirement for the Doctor of Philosophy degree
in Industrial Engineering at the May 2017 graduation.

Thesis Committee: _____
Yong Chen, Thesis Supervisor

_____
Andrew Kusiak

_____
Amaury Lendasse

_____
Kate Cowles

_____
Joyee Ghosh

To my parents and grandparents; to my wife, Shan.

ACKNOWLEDGEMENTS

First I would like to express my sincere gratitude to my advisor and mentor, Professor Yong Chen, for all the time and effort he has invested in me throughout my PhD study. He has provided me not only the professional guidance on academic work, but also valuable ways of thinking and solving problems. His enthusiasm, patience and friendliness has made my PhD study an invaluable and unforgettable experience. It is my great fortune to have been his student.

I would also like to thank Professor Andrew Kusiak, Professor Amaury Lendasse, Professor Kate Cowles, and Professor Joyee Ghosh, for serving on my committee and offering valuable comments and suggestions on my research.

I thank all of the members of the Risk and Reliability Lab for their useful help. Special thanks to my previous colleague: Huan Yu, for her kindly help during my job search. Great appreciation for all of my friends in Iowa City. They have made my past five years a truly enjoyable part of my life.

Finally and most importantly, I would like to thank my parents and my wife, Shan, for their unconditional love, encouragement and support that make this possible. I dedicate this thesis to them.

ABSTRACT

The primary objective of this thesis is to study change detection problems and their applications in longitudinal and functional data. In particularly, two types of change detection problems for longitudinal data are considered. The first type of problems is on-line change detection for longitudinal data, where we focus on detecting changes within a single longitudinal data stream that arrive into the system sequentially. The other type of change detection problems that will be studied in this thesis concerns about detecting outliers from a set of longitudinal or functional data.

For the first type of change detection problems, we study two novel engineering applications. The first application is studied in Chapter 2, focusing on the on-line steady state detection. The goal is to identify the transition point between transient period and steady state period. We propose a novel on-line steady state detection algorithm based on a multiple change-point state space formulation and the sequential Monte Carlo methods. Compared to other existing methods, the main contribution of this work is its significantly improved computational efficiency by the use of the Rao-Blackwellization method, making it a much preferred method for many on-line applications where quick processing of the data in real time is critical. Additionally, the proposed method is shown to have more robust detection performance than existing methods when dealing with different types of signals.

In Chapter 3, we study the second application of change detection problems, which focuses on statistical process control for the short-run process. We propose new methods under the Bayesian framework to track the process mean and detect on-line if the process mean is beyond certain control limits or specification limits. Our model

modifies the original model proposed by Tsiamyrtzis and Hawkins (2005) and can be more flexible in handling linear trends of the process. Compared to the method proposed by Tsiamyrtzis and Hawkins (2005), the advantages of our method are two-folds. Firstly, the performance of our method is more robust to parameter misspecification and requires less knowledge of the process to make accurate estimations. Secondly, the resulted posterior inference of the process mean has a significantly reduced number of mixtures, leading to substantial saving of computational and memory cost.

The other type of change detection problems studied in this thesis concentrates on analysis of a set of longitudinal or functional data, which is discussed in Chapter 4. In particular, we focus on the outlier detection problem for functional data, where the outlier is defined as a curve that is generated from a different process compared to normal curves. Based on the use of the concept of data depth, we propose two new depth notions, the weighted band depth and the localized weighted band depth for detecting various outliers. Our main contribution is proposing a new idea called the shape distance, which makes our methods particularly effective in detecting outliers that have different shapes from normal curves.

PUBLIC ABSTRACT

In real practical engineering applications, we are often encountered with signals that consist of repeated measurements on the same object over different time points. Such signals are called the longitudinal signals. For example, in prognosis of some in-service units, the continuous change of their health status over time is a longitudinal trend; for some cutting tools, their gradual failure (tool wear) due to regular operations can be also considered as longitudinal signals. For such longitudinal signals, to timely detect their critical changes or occurrences of failure and anomalies is important to us since we can make corrective actions as quickly as possible. The main goal of this thesis is to introduce some new methods that can make efficient and robust detection on such change/anomalies of longitudinal signals observed in engineering applications.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER 1. INTRODUCTION 1
1.1 On-Line Steady State Detection 2
1.2 Short-Run Statistical Process Control 5
1.3 Statistical Depth and Outlier Detection for Functional Data 9

CHAPTER 2. ON-LINE STEADY STATE DETECTION BASED ON RAO-BLACKWELLIZED SEQUENTIAL MONTE CARLO 14
2.1 Piecewise Linear Model with Multiple Change-Points 15
2.2 Rao-Blackwellized Sequential Monte Carlo for On-line Steady State Detection 18
2.2.1 Review of Standard SMC for State Space Model 18
2.2.2 Rao-Blackwellized SMC Algorithm 22
2.2.3 Optimal Resampling 27
2.2.4 Steady State Detection Rule 30
2.3 Numerical Study 31
2.3.1 Simulated Signals and Model Parameter Setup 31
2.3.2 Illustration of the Steady State Detection 33
2.3.3 Performance Evaluation 35
2.3.4 Comparison with Existing On-Line Methods 36
2.3.5 Application to Steady State Detection in Micro/Nanoparticle Dispersion Process 39
2.4 Summary 40

CHAPTER 3. A NEW BAYESIAN ON-LINE INFERENCE METHOD FOR THE SHORT-RUN SPC 42
3.1 Motivation and Model Formulation 43
3.2 On-Line Exact Bayesian Inference of the Process Mean 46
3.2.1 On-Line Exact Bayesian Inference on the Posterior of Occasional Jumps 46
3.2.2 On-Line Inference for the Random Walk Plus Random Jump Model 48
3.2.3 On-Line Inference for the Piecewise Linear Model 50
3.2.4 On-Line Inference for Constant Noise Variance 53
3.3 Approximated On-Line Inference of the Process Mean 57
3.4 Simulation Study 59
3.4.1 Data Generated from the Random Walk Plus Random Jump Model 59
3.4.2 Piecewise Linear Data 66
3.4.3 Degradation Data for Fatigue Crack Propagation 68
3.5 Summary 71

CHAPTER 4. WEIGHTED BAND DEPTH APPROACHES FOR FUNCTIONAL DATA ANALYSIS 73
4.1 Review of the Band Depth for Functional Data 74
4.2 The Weighted Band Depth (WBD) and Localized Weighted Band Depth (LWBD) 77
4.2.1 The Shape Distance in WBD and its Motivation 77

4.2.2 The Localized WBD 77
4.3 Simulation Study 85
4.3.1 Models for Generating Simulated Data 86
4.3.2 Performance Evaluation 91
4.3.3 Simulation Results 92
4.4 Application 95
4.5 Summary 97

CHAPTER 5. CONCLUSION AND FUTURE WORK 99

APPENDIX A. PROOF OF LEMMA 2-1 IN SECTION 2.2.2 101

APPENDIX B. PROOF OF LEMMA 2-2 IN SECTION 2.2.2 102

APPENDIX C. THE STRATIFIED SAMPLING ALGORITHM IN SECTION 2.2.3 104

APPENDIX D. NORMAL APPROXIMATION OF RBSMC DETECTION RESULTS IN SECTION 2.3.4 105

APPENDIX E. PROOF OF LEMMA 3-1 IN SECTION 3.2.2 106

APPENDIX F. PROOF OF LEMMA 3-2 IN SECTION 3.2.3 107

APPENDIX G. CALCULATION OF $P(S, T)$ IN SECTION 3.2.3 108

APPENDIX H. COMPLETE ODA RESULTS OF HMD AND KFSD FOR SHMM1 – 8 IN SECTION 4.3.3 108

APPENDIX I. COMPLETE ODA RESULTS OF HMD AND KFSD FOR LOMM1 – 4 IN SECTION 4.3.3 110

APPENDIX J. COMPLETE ODA RESULTS OF WBD, LWBD, HMD AND KFSD FOR OUTLIER DETECTION IN SECTION 4.4 111

REFERENCES 112

LIST OF TABLES

Table 2-1. Four types of bias functions 32

Table 2-2. Equations and parameter values for three types of noises.....32

Table 2-3. Comparison of RBSMC, PF, SDM and VRT for w = 1 and noise type AR(0). The detection parameters are (1) RBSMC, $s_0 = 0.0021$; (2) PF, $s_0 = 0.0022$; (3) SDM, window size $m = 50$, threshold=8×10^{-5}; (4) VRT, $m = 98$, threshold=0.6.....38

Table 3-1. Simulation results of TH and EBKF using the correct parameter specifications with $S = 1000, N = 10$.....63

Table 3-2. Simulation results of TH and EBKF using different IMP settings and the TMP setting of $p = 0.1, \sigma_{RW}^2 = 0.05, m = 1$.....64

Table 3-3. Simulation results of different methods using data with two piecewise linear segments 68

Table 3-4. Simulation results of different methods using the fitted crack propagation path with different noise levels.....71

Table 4-1. Simulation results of ODA using WBD, LWBD, HMD, KFSD and BD for mixture models SHMM1 – 8 producing shape outliers (For HMD and KFSD, only results of the best average ODA are shown, See APPENDIX H for their complete results).....93

Table 4-2. Simulation results of ODA using WBD, LWBD, HMD, KFSD and BD for mixture models LOMM1 – 4 producing location outliers (For HMD and KFSD, only results of the best average ODA are shown, See APPENDIX I for their complete results).....94

Table 4-3. Best ODA of different depth approaches 97

LIST OF FIGURES

Figure 1-1. Illustration of the drawback of BD in detecting shape outliers. 11

Figure 2-1. Illustration of approximating nonlinear signals using piecewise linear model: (a) signal generated using exponential function and noise; (b) oscillating nonlinear function. 15

Figure 2-2. Illustration of the state space model. 17

Figure 2-3. Illustration of steady state detection for linear signal: simulated signal (top), estimated posterior probability of the latest change-point at different time steps (middle), and the detection index $\hat{\pi}_t$ (bottom). 34

Figure 2-4. Steady state detection for a signal with zero mean and exponentially decaying variance: simulated signal (top), and the detection index $\hat{\pi}_t$ (bottom). 35

Figure 2-5. The weighted standard detection error (WSDE) of the proposed RBSMC, PF, SDM, VRT as a function of the penalty weight w for AR(0), AR(1) and AR(2). 37

Figure 2-6. Steady state detection for the CNP signal with ultrasonic power 30W and 50W in the dispersion of 30 g Al_2O_3 40

Figure 3-1. Estimated mean tracking performance of EBPLUV and EBPLUV-EV methods. 56

Figure 3-2. Computational costs of different methods of 100 simulations with varying data size N. Each signal is generated using $\sigma^2 = 1, p = 0.1, \sigma_{RW}^2 = 0.05, \delta = \sigma_{RW}$. 65

Figure 3-3. Illustrations of the crack length-cumulative loading cycles curve with and without noise. 69

Figure 4-1. Band delimited by three solid curves. 75

Figure 4-2. (a) The band (solid lines) delimited by points (1, 3), (4, 6), (5, 4) in parallel coordinates and (b) the same three points and interval (the rectangle) in Cartesian coordinates. Point (2, 5) is inside the band, point (3, 7) is outside the band. 76

Figure 4-3. Illustration of the shape distance of the target curve f (solid line) to its enclosed band delimited by two dashed curves. 78

Figure 4-4. Illustration of the shape distance of the target curve f to the band when it is one of the boundary curves of the band. 79

Figure 4-5. Illustration of detecting location outliers. The dashed curve is the outlier curve, while two bolded curves (in the geometric center of their corresponding groups of normal curves) are curves with largest depths obtained by LWBD (both c_1 and c_2 are chosen to be 50-th percentile of

empirical distribution of $\left\{d(f_{i_1}, f_{i_2}) = \max_{t\in\mathcal{T}}|f_{i_1}(t) - f_{i_2}(t)|, 1 \le i_1 < i_2 \le n\right\}$). 82

Figure 4-6. Illustration of the location distance of the target curve f to the band when it is one of the boundary curves of the band. 83

Figure 4-7. Examples of simulated data sets generated by SHMM1 – 6. Bolded curves are outlier curves while others are normal curves. 88

Figure 4-8. Examples of simulated data sets generated by SHMM7 and SHMM8. Bolded curves are outlier curves while others are normal curves. 89

Figure 4-9. Examples of simulated data sets generated by LOMM1 – 4 ($K = 25$ for LOMM2, LOMM3 and LOMM4). Bolded curves are outlier curves while others are normal curves. 91

Figure 4-10. The near-infrared absorbance spectrum for both high and low fat content of meat samples. 97

CHAPTER 1. INTRODUCTION

Many engineering applications concern about longitudinal data. For example, in prognosis of the in-service units, the change of health status of the units over time is a longitudinal signal, which is also called the degradation signal in the engineering area. In process monitoring and control, usually the signal has to undergo a transient period before entering a steady period where its distribution becomes stable. Such signal also forms a longitudinal trend.

Longitudinal data can be also considered as a type of functional data (Ramsay and Silverman, 2005). In this thesis, we focus on change detection problems and its applications for longitudinal and functional data. We consider two types of change detection problems for longitudinal data. The first type of problems is on-line change detection for longitudinal data, where we consider on-line algorithms to detect changes in trend within a single longitudinal data stream with data observed sequentially. Recently, general on-line change detection for data with multiple change points is studied in statistics literature based on Bayesian methods. For example, Chopin (2007) studied the problem of on-line detecting multiple change-points in time series data under a Bayesian framework, and without any prior knowledge of the exact number of change-points. Fearnhead and Liu (2007) also developed an efficient Bayesian on-line inference method for multiple change-points problems. In this thesis, we study two novel engineering applications of on-line change detection. The first application is studied in Chapter 2, focusing on the on-line steady state detection, where the transition point between transient period and steady state period needs to be identified. The second application is in statistical process control for the short-run processes, which is studied in Chapter 3.

The goal is to track the process mean and detect if the process mean is beyond certain control limits or specification limits. The other type of change detection problems that will be studied in this thesis concerns about a set of longitudinal or functional data, which is discussed in Chapter 4. In particularly, we focus on developing useful methods based on the concept of *data depth* for detecting various outliers that are coming from a different process compared to normal curves. In the following subsections, we provide more information on the three problems studied in this thesis.

1.1 On-Line Steady State Detection

Steady state of a system in many applications is one of the most important requirements to evaluate the performance of the process or to trigger the next actions in the process control. Timely detection of whether a data stream reaches the steady state (i.e., mean and variance unchanged) has been found critical in various fields, such as process control (Mahuli et al., 1992; Cao and Rhinehart, 1995; Jiang et al., 2003; Wu et al., 2013), data reconciliation (Narasimhan et al., 1986; Bagajewicz and Jiang, 2000; Bhat and Saraf, 2004; Korbel et al., 2014), fault detection and diagnosis (FDD) (Kim et al., 2008), and process optimization (Mhamdi et al., 2010). We can categorize the steady state detection problem into two types: *off-line* and *on-line*. Most of the well-developed methods in the literature correspond to the off-line detection problems arising from discrete-event simulations, where it is usually very difficult to start the simulation directly from the steady state because the steady state of the system is typically unknown. Data collected during the transient period (or *warm-up period*) prior to the steady state causes estimation bias, which is called the *initialization bias*, in the steady state parameter estimation (Kelton and Law, 1983; Gallagher et al., 1996; Fishman, 2001;

Hoad et al., 2010). To solve this problem, usually the simulation is first run long enough to guarantee the simulation output has reached the steady state. Then off-line methods are typically used to identify the starting point of the steady state in the simulation outputs so that the data from the transient period can be removed. These methods can be further classified into five different types (Robinson, 2007): graphical methods, heuristic approaches, statistical methods, initialization bias tests, and hybrid methods.

Compared to the off-line methods which have been extensively investigated in the simulation literature, studies for *on-line* steady state detection are limited. The main challenge of on-line detection is that it has to be done *in real-time*, which justifies the need for detection procedures to timely update estimations as the latest observations become available. The existing on-line methods typically utilize a moving data window, based on which some test statistics are developed to decide if the signal has entered the steady state. Examples of such methods include polynomial interpolation test (PIT) (Savitzky and Golay, 1964; Roux et al., 2008), variance ratio test (VRT) (Crow et al., 1960; Cao and Rhinehart, 1995), slope detection method (SDM) (Holly et al., 1989; Bethea and Rhinehart, 1991; Wu et al., 2013), and *t*-test (Narasimhan et al., 1987). However, the performance of all these methods is highly dependent on the selection of the data window size. Too small or too large the size may significantly increase either the *false alarm rates* (FAR) or detection delays. Also, the appropriate window size is very sensitive to noise levels as well as types of signals. Usually one specific window size cannot perform consistently well for various types of signals with different noise levels. Therefore, it is urgent to develop methods with more flexibility and robustness to handle various situations.

Naturally the on-line steady state detection problem can be approached from the Bayesian perspective, where the prior information of model parameters and transition point to the steady state, combined with the latest observations, are used to dynamically update the knowledge of the current state of the process. The main disadvantage of the Bayesian method is that it usually leads to intractable posterior distributions. However, recent advance in the field of computational statistics makes the computations of posterior distributions feasible. The most common computational Bayesian methods is the *Markov Chain Monte Carlo* (MCMC) (Liu, 1998; Robert and Casella, 2004) algorithm. However, MCMC algorithms are typically not appropriate for on-line applications because of its fast increasing computational cost over time when data size gets larger and larger. On the other hand, the *sequential Monte Carlo* (SMC) methods, including the *particle filtering* (PF) technique (Doucet et al., 2001; Arulampalam et al., 2002), can be used. In contrast to the standard MCMC approaches, the sequential structure of the SMC methods allows for updating estimations sequentially for each newly arriving observation in a computationally efficient way, which is very useful for on-line inference.

Recently, a particle filtering (PF) method has been proposed by Wu et al. (2015) for on-line steady state detection. In their method, the target signal is approximated by a multiple change-point model and particle filtering techniques are used to estimate the posterior distribution of the latest change-point and other model parameters (slope, intercept and noise variance). Some improvement strategies including the stratified sampling, partial Gibbs resample-move techniques, and timeliness improvement strategy are developed to overcome the particle degeneracy and impoverishment problems and

reduce the computational cost. The PF method is shown to be much more effective and robust in steady state detection compared with moving data window based methods. However, the major disadvantage of the PF method is that it requires a large number of particles to achieve accurate posterior distribution approximations, due to the high dimension of model parameters, which leads to high computational cost and may limit its applications in many on-line detection problems that require quick responses.

In Chapter 2, a novel SMC method is proposed using a Rao-Blackwellization technique, combined with a resampling method called the *Optimal Resampling*, to substantially reduce the computational cost and improve the detection robustness on noisy data. The main contribution of our algorithm is its significant improvement in computational efficiency by taking advantage of the Rao-Blackwellization technique, while still achieving comparable or even better detection performance compared with the PF method. The significant reduction of computational costs makes our method a much more preferred method for many on-line applications where quick steady state detection is critical.

1.2 Short-Run Statistical Process Control

In the area of statistical process control (SPC), the process parameters such as process mean are subject to changes over time and we are concerned with detecting on-line whether the process mean has deviated from certain in-control region and fallen into some out-of-control region. Standard methods such as Shewhart control charts (Shewhart, 1931) and cumulative sum (CUSUM) control charts (Page, 1961) have been popularly used based on a long phase I data gathering.

However, in recent decades, as a major consequence of the popularity of the just-in-time (JIT) manufacturing, industries are gradually abandoning the mass production and in favor of the lean production in order to be more competitive in the market (Womack et al., 1990). Because of this, there has been an inevitable trend to produce smaller lot sizes or production runs. Unlike the classical SPC, in short production runs, prior to the beginning of production the control limits for a standard control chart and the phase I data needed to estimate the in-control process parameters are often unavailable. Therefore, instead of collecting a large number of in-control data in phase I analysis, as required by traditional control chart methods, new methods should be developed to estimate and timely update the process parameters from the very beginning of a short-run process.

So far, substantial research works have emerged focusing on the short-run SPC. Many methods are developed following the traditional SPC setup, where the process is assumed to have a *stationary in-control state* (i.e., the in-control process parameters such as process mean are constant). For example, Hawkins (1987) introduced the 'self-starting' CUSUM charts, which do not rely on estimating process parameters from a large Phase I data set prior to monitoring the short-run process. Quesenberry (1990, 1991) proposed the 'self-starting' Q-charts, which are constructed based on sequentially updating the estimates of the in-control process mean and variance. However, as pointed out by Del Castillo and Montgomery (1992), in some cases the Q-charts do not exhibit adequate average run length (ARL) performance. To obtain a better ARL performance, Del Castillo and Montgomery (1992) proposed a first-order adaptive Kalman filtering method assuming the measurement noise variance is unknown, while the knowledge of

the process mean is needed or not too different from real values. In addition, Woodward and Naylor (1993) proposed a Bayesian method for the process with piecewise constant means subject to some jumps. However, rather than random jumps, the jumps are assumed to occur at some known times.

Opposed to the stationary in-control state, in many short-run processes such as a process subject to tool wear or system degradations, signals may exhibit continuous drifts in the in-control state. The current existing literature for tackling such *in-control state with drifting* for the short-run process is mainly based on the assumption that the in-control process mean follows a normal random walk. For example, Crowder and Eshleman (2001) proposed a Bayesian adaptive filtering approach where the process mean follows purely a random walk. The presented method focused only on parameter estimation, while not considering the test of hypothesis whether the process mean has shifted to the out-of-control state. Tsiamyrtzis and Hawkins (2005) proposed a Bayesian method based on the assumption that the process mean follows a normal random walk with some random jumps of fixed size at random times. The random jumps can be used to model random step changes of processes due to, say, tool chipping (Tsiamyrtzis and Hawkins, 2005). This method is effective in detecting the shift of the process mean into the out-of-control state under the assumed model. However, the assumption of known and fixed jump sizes may not be realistic in many applications. For example, the amount of material loss due to tool chipping is usually random and unknown. The method of Tsiamyrtzis and Hawkins (2005) also suffers from intensive computational burden with exponentially increasing number of mixtures in the posterior distribution of the process mean. Most recently, Apley (2012) introduced a Bayesian approach that utilizes the so-

called posterior distribution charts for graphically exploring the change of the process mean. The proposed model in Apley (2012) can handle any Markov model for the process mean. The major limitation is that the numerical approach proposed by Apley (2012) for approximating the posterior distributions is difficult to handle cases with more than one unknown parameters.

In Chapter 3, we propose a novel and flexible framework for short run SPC based on a Bayesian model. The major contributions of the proposed models and methods in Chapter 3 are summarized as follows: (1) Compared with existing methods such as the methods proposed by Tsiamyrtzis and Hawkins (2005) and Apley (2012), our proposed methods require much less knowledge of the process parameters. Using non-informative priors, the proposed model can handle situations where little knowledge is available about the jump size or the relation of the process means before and after a jump. (2) The proposed methods are much more computationally efficient, which makes our methods more applicable in many practical situations where the information of model parameters is limited and timely detection of the out-of-control state is crucial. By assuming the process mean before and after random jumps are independent in our newly proposed model, the *exact Bayesian* method (Fearnhead and Liu, 2007) for the multiple change-point model can be used to reduce the computational costs from being exponential as in Tsiamyrtzis and Hawkins (2005) to be quadratic over time. We also propose an approximate Bayesian inference method based on the particle filter method to further reduce the computation costs to be linear over time. (3) When a process is subject to multiple phases of degradations, the trend of the process can be modeled with a piecewise linear model (Chen and Tsui, 2013). And a nonlinear degradation curve can also be

approximated by a piecewise linear trend. Following this idea, in Chapter 3 we further propose to extend the step changes in the models of Tsiamyrtzis and Hawkins (2005) and Apley (2012) to a piecewise linear model that can be applied to general degradation signals. In many degradation models, a failure limit is defined so that a failure occurs if the degradation path crosses the failure limit (Lu and Meeker, 1993; Gebraeel et al., 2005; Chen and Tsui, 2013; Si et al., 2016). If the control limit of our model is set as the failure limit of a degradation signal, the proposed procedure can be applied to detect failures based on a noisy degradation signal.

1.3 Statistical Depth and Outlier Detection for Functional Data

Functional data analysis (FDA) has been a growing interest in recent decades (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006) due to the technological advances in areas such as engineering, biology, medicine, finance and environmetrics which allow us to have random samples of curves. For functional data, each sample consists of values that are assumed to be generated from a stochastic function at different time points in a given interval of finite length. FDA is superior to the standard multivariate analysis in dealing with such functional samples for mainly two reasons. Firstly, since the number of dimensions is typically much higher than the number of samples, standard multivariate methods are computationally infeasible due to the curse of dimensionality. Secondly, the standard multivariate analysis is usually not able to capture some natural properties of the functional data, such as shape.

One of the major tasks in FDA is to obtain an ordering set of functional curves, which is an extension of the univariate order statistics. A natural way of ordering such curves is to measure the degree of centrality of each curve within the underlying

population distribution or given samples. This is the idea of *statistical depth*, whose aim is to provide a center-outward ranking within the sample curves. For example, the curve with maximum depth is a median function that can be considered as the center of the functional distribution, while curves of significantly low depth are expected to be those curves being far away from the center of the data (i.e., outliers). Therefore, the depth is particularly useful for outlier detection and defining some robust location estimates of the sample curves, such as the trimmed mean, or other L-estimates (linear combinations of order statistics).

The statistical depth was originally proposed in the multivariate framework. The well-known examples include the Tukey's halfspace depth (Tukey, 1975), the Oja's depth (Oja, 1983), the simplicial depth (Liu, 1990), the spatial depth (Chaudhuri, 1996), the zonoid depth (Koshevoy and Mosler, 1997), the Mahalanobis depth (Zuo and Serfling, 2000), and the projection depth (Zuo, 2003). Zuo and Serfling (2000) also provides an extensive study of the definitions, properties, as well as some applications of multivariate depths. However, direct extension of using statistical depths from multivariate data to functional data encounters some difficulties. As discussed in Cuevas et. al (2007), some of the proposed multivariate depths have reasonable computational performance in lower dimensions (two or three) but will be computationally intractable in higher dimensional cases. Additionally, the multivariate depths typically do not take into consideration some natural properties of the functional curves, such as shape (López-Pintado and Romo, 2009).

Therefore, different depth notions for analyzing functional data with more effectiveness and robustness are developed, which are called the *functional data depths*.

Most of the existing functional data depths are *global-oriented*, in which the degree of the centrality of a curve depends equally on the remaining curves in the data. After the pioneering work proposed by Fraiman and Muniz (2001), several alternatives have been also proposed, including the projection-based functional depth in Curvas et al. (2007), the *band depth* in López-Pintado and Romo (2009), the *half-region depth* in López-Pintado and Romo (2011), and the *functional spatial depth* in Chakraborty and Chaudhuri (2014). In contrast, some research works also focus on developing the *local-oriented functional depth*, such as the *h-modal depth* (HMD) in Curvas et al. (2006), and the *kernel-based functional spatial depth* (KFSD) proposed by Sguera el al. (2014). These local-oriented depth approaches are particularly useful in detecting location outliers (outliers that are very distant from the mean of normal curves), as a result of considering reduced contributions of curves to the depth of the target curve (the curve whose depth is to be calculated) when their distances to the target curve increase.

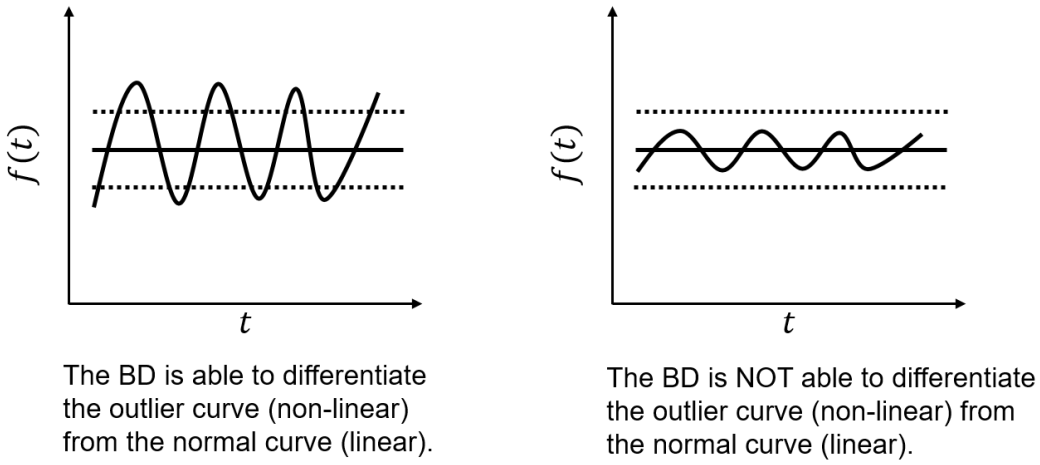


Figure 1-1. Illustration of the drawback of BD in detecting shape outliers.

Among the existing work in functional depth, the band depth (BD) in López-Pintado and Romo (2009) is shown to be effective in detecting shape outliers – outliers

exhibiting different shape patterns from normal curves or very irregular in a set of smooth curves. However, the ability of BD in detecting shape outliers is still very limited. BD is defined based on all possible bands delimited by the graphs on the plane of at least two curves, and whether the target curve is inside these bands. The more bands that the target curve falls inside, the larger depth of the target curve would have. As a consequence, BD is only sensitive to those shape outliers which are not fully inside the bands, while being incapable of differentiating those shape outliers that are completely inside the bands. As illustrated in Figure 1-1, for the first case, BD is able to detect the outlier since it is not completely inside the band (delimited by two black curves); however, for the second case, BD cannot distinguish the outlier from the normal curve since they are both inside the band.

To overcome such drawback and improve the robustness in detecting various types of outliers, in Chapter 4 we propose a new functional data depth called the weighted band depth (WBD) and its localized version for functional data. The main contributions of our work are two-folds. Firstly, by introducing the idea called *shape distance*, which is defined based on measuring differences between curves with respect to their shapes, WBD has substantially improved strength of detecting shape outliers. Secondly, by considering the relative location information between curves, the *localized* version of WBD (LWBD) can be also very sensitive to detect location outliers. LWBD is shown to be robust in detecting both location outliers and shape outliers, making it a preferred depth approach when different types of outliers are present.

In summary, the remainder of this thesis is organized as follows. In Chapter 2, we study the on-line steady state detection problem. A novel on-line steady state detection

algorithm under the Bayesian framework based on a multiple change-point state space model formulation and the sequential Monte Carlo methods is presented. In Chapter 3, an efficient Bayesian method is introduced for on-line inference of the process mean in the short-run process of the statistical process control. In Chapter 4, we propose the weighted band depth and its localized version for functional data in detecting various types of outliers with more effectiveness and robustness.

CHAPTER 2. ON-LINE STEADY STATE DETECTION BASED ON RAO-BLACKWELLIZED SEQUENTIAL MONTE CARLO

In this chapter, we introduce a novel on-line steady state detection algorithm under the Bayesian framework based on a multiple change-point state space formulation and the sequential Monte Carlo methods. A Rao-Blackwellization technique is proposed to substantially reduce the variance of Monte Carlo estimation and greatly enhance the computational efficiency. In addition, a resampling scheme called the Optimal Resampling is used for eliminating duplicate samples and the robustness of steady state detection is significantly improved by using the information of the particles more efficiently. Numerical studies based on simulated signals and application to a real data set are used to evaluate the performance of the proposed method and compare with other existing methods from the literature. The proposed method is shown to establish a more robust performance than other methods. And it is much more computationally efficient than the standard sequential Monte Carlo method.

The remainder of this chapter is organized as follows. A multiple change-point model formulation of this problem is introduced in Section 2.1. In Section 2.2, we give a detailed description of the proposed SMC algorithm. The numerical examples and application to real signals are presented in Section 2.3. And a summary is given in Section 2.4.

2.1 Piecewise Linear Model with Multiple Change-Points

Given a noisy signal $\mathbf{y}_{0:N} = (y_0, y_1, \ldots, y_N)$, this chapter targets detecting the steady state of the signal using multiple change-point models. The basic idea is that we approximate any signal, linear or nonlinear, using a piecewise linear model that allows for local linear representations of the signal, as shown in Figure 2-1. The parameters (slope, intercept and noise variance) are assumed to be independent across different segments. When the latest line segment is sufficiently 'flat', the signal is considered to be in the steady state.

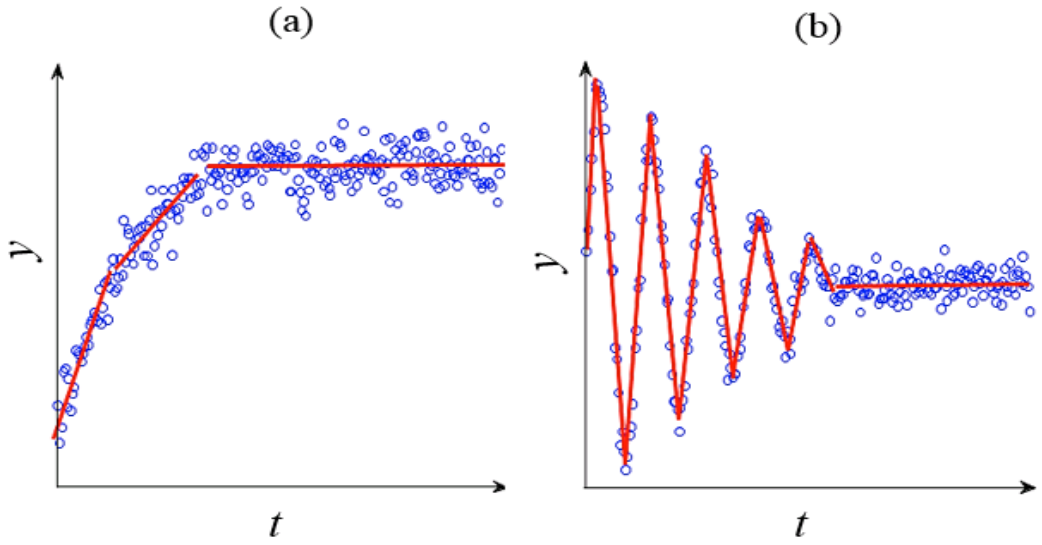


Figure 2-1. Illustration of approximating nonlinear signals using piecewise linear model: (a) signal generated using exponential function and noise; (b) oscillating nonlinear function.

Suppose at time t the model parameters are $\boldsymbol{\xi}_t = (\beta_{0t}, \beta_{1t}, \sigma_t^2)$, where β_{0t} is the intercept of the current line segment, β_{1t} is the slope, and σ_t^2 is the unknown noise variance. The model parameters $\boldsymbol{\xi}_t$ are subject to change at unknown change-points $\mathbf{C}_m =$

$(c_1, c_2, \dots, c_m)$ while remaining constant otherwise. This multiple change-point model can be written as:

$$\boldsymbol{\xi}_t = \begin{cases} \boldsymbol{\theta}_1 & \text{if } 0 \le t < c_1 \\ \boldsymbol{\theta}_2 & \text{if } c_1 \le t < c_2 \\ \vdots & \vdots \\ \boldsymbol{\theta}_m & \text{if } c_{m-1} \le t < c_m \\ \boldsymbol{\theta}_{m+1} & \text{if } c_m \le t \le N \end{cases} \tag{2-1}$$

where $\boldsymbol{\theta}_i \in \mathcal{R}^3$, $i = 1, 2, \dots, m+1$ are the values of model parameters at $m+1$ different line segments.

Within the Bayesian framework, we assign appropriate priors to the change-points between two segments, as well as other model parameters. Their posterior distributions can then be sequentially updated and the steady state can be inferred based on the posterior distributions of model parameters (e.g., slope) of the current line segment. In model (2-1), the $\boldsymbol{\theta}_i$'s and c_i's are unknown. We assign a prior distribution $q_{\boldsymbol{\theta}}(\cdot)$ for $\boldsymbol{\theta}_i$'s. Let τ_t be the latest change-point up to time t. It is easy to see that $\tau_t = c_{i-1}$ iff $c_{i-1} \le t < c_i$. Therefore τ_t's and c_i's contain equivalent information. We assign a prior transition probability for τ_t given τ_{t-1} as $P(\tau_t|\tau_{t-1})$. If a change occurs at time t, we have $\tau_t = t$, otherwise $\tau_t = \tau_{t-1}$. In this chapter we assume $P(\tau_t = t|\tau_{t-1}) = p$, which corresponds to a geometric distribution with probability p for the random duration of each line segment.

To facilitate application of the sequential Monte Carlo (SMC) method for online inference of model parameters, we first reformulate the piecewise linear model in Eq. (2-1) into a state space model, where at each time t the distribution of the observation y_t depends on a hidden state vector denoted by $\mathbf{x}_t$. Defining the state vector as $\mathbf{x}_t = (\boldsymbol{\xi}_t, \tau_t)$, the state space model is given as:

$$\mathbf{x}_t = (\boldsymbol{\xi}_t, \tau_t)|\mathbf{x}_{t-1} = \begin{cases} (\boldsymbol{\theta}^*, t) & \text{with probability } p \text{ (change occurs at } t) \\ \mathbf{x}_{t-1} = (\boldsymbol{\xi}_{t-1}, \tau_{t-1}) & \text{with probability } 1-p \text{ (no change at } t) \end{cases} \quad (2\text{-}2)$$

$$y_t = \beta_{0t} + \beta_{1t} t + \varepsilon_t, \varepsilon_t \sim N(0, \sigma_t^2)$$

where $\boldsymbol{\theta}^* \sim q_{\boldsymbol{\theta}}(\cdot)$ and is independent of $\boldsymbol{\xi}_{t-1}$, and ε_t is the Gaussian noise. The state transition of the state space model is illustrated in Figure 2-2. It can be seen that the state space model in Eq. (2-2) equivalently represents the piecewise linear model in Eq. (2-1). Based on Eq. (2-2), at any time t, with probability p a change occurs ($\tau_t = t$) and a new line segment is started. Since $\boldsymbol{\xi}_t, t \geq 0$ contain all the information on $\boldsymbol{\theta}_i$'s in Eq. (2-1) and $\tau_t, t \geq 0$ contain all the information on c_i's in Eq. (2-1), the inference of the parameters in the piecewise linear model in Eq. (2-1) is equivalent to the inference of the state vector $\mathbf{x}_t = (\boldsymbol{\xi}_t, \tau_t)$ in Eq. (2-2). Consequently, we can focus on the state space model in Eq. (2-2) and the inference of $(\boldsymbol{\xi}_t, \tau_t)$ to develop the steady state detection method.

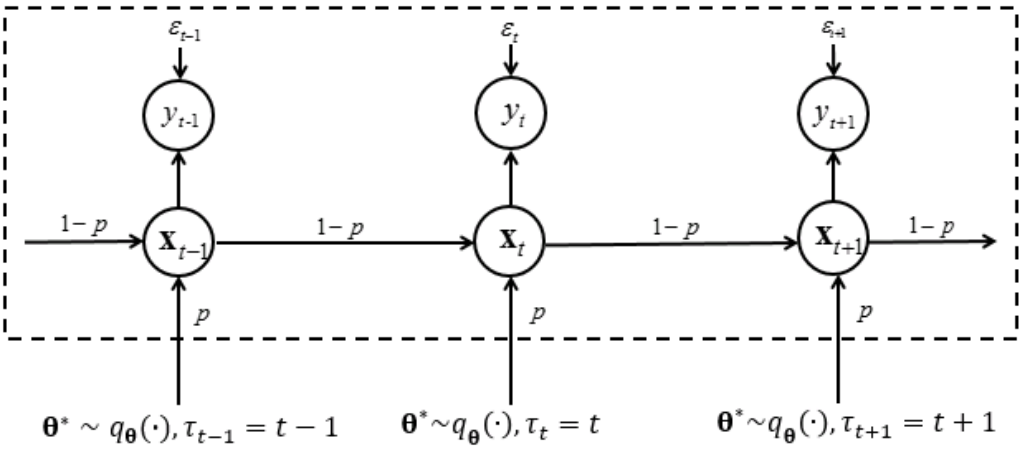


Figure 2-2. Illustration of the state space model.

If $\tau_t = 0$ for all t, that is, if there is no change-point, the above model reduces to a linear Gaussian system, in which the *Kalman filter* (KF) (Kalman, 1960) provides a closed-form solution for efficient state estimations. However, because of the existence of

unknown change-points, the state space model is a nonlinear model which cannot be solved directly by KF. On the other hand, the PF techniques are common and effective for nonlinear state space models. Recently, a PF method has been proposed by Wu et al. (2015) for on-line steady state detection. A disadvantage of the method in Wu et al. (2015) is that it requires a large number of particles because each parameter in the state vector $\mathbf{x}_t$ needs to be sampled. In the following section, we propose a more efficient sequential Monte Carlo algorithm based on a variance reduction method called *Rao-Blackwellization*.

2.2 Rao-Blackwellized Sequential Monte Carlo for On-line Steady State Detection

2.2.1 Review of Standard SMC for State Space Model

In this section, the general framework of standard SMC algorithm for state space model is reviewed. The main idea for SMC method is to use a recursive importance sampling strategy to get a set of properly weighted samples, which are used to approximate the desired posterior density of the state vector in a state space model. For the state space model in Eq. (2-2), let $\mathbf{x}_{0:t} = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_t\}$ be the set of all state vectors up to the current time t, and $\mathbf{y}_{0:t} = (y_0, y_1, \dots, y_t)$ be the observations up to the current time t. On-line inference of $\mathbf{x}_{0:t}$ is of our central interest; that is, at time t, we want to estimate the posterior density $p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$. Usually $p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$ is of an intractable density form where direct simulations cannot be implemented.

In such case, we often resort to the *Importance Sampling* (IS) approach in which we bypass the intractable distribution by considering some *importance distribution* where direct simulations can be easily performed. Suppose samples (particles) $\{\mathbf{x}_{0:t}^{(i)}\}_{i=1}^{n}$ are

generated from an importance distribution $\pi(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$. By associating the *importance weight*:

$$w_t^{(i)} = \frac{p(\mathbf{x}_{0:t}^{(i)}|\mathbf{y}_{0:t})}{\pi(\mathbf{x}_{0:t}^{(i)}|\mathbf{y}_{0:t})} \tag{2-3}$$

to the sample $\mathbf{x}_{0:t}^{(i)}$, the posterior density of interest can be approximated as:

$$\hat{p}(\mathbf{x}_{0:t}|\mathbf{y}_{0:t}) = \sum_{i=1}^{n} w_t^{(i)} \delta\left(\mathbf{x}_{0:t} - \mathbf{x}_{0:t}^{(i)}\right) \tag{2-4}$$

where δ denotes the Dirac function and the weights are normalized such that $\sum_{i=1}^{n} w_t^{(i)} = 1$. Therefore $\hat{p}(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$ is a discrete weighted approximation of the true posterior $p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$. The pairs $\{\mathbf{x}_{0:t}^{(i)}, w_t^{(i)}\}_{i=1}^{n}$ are a collection of *properly weighted sample* with respect to the posterior distribution $p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$. An important observation is that $\mathbf{x}_t^{(i)}$ is also properly weighted by $w_t^{(i)}$ with respect to the marginal posterior distribution $p(\mathbf{x}_t|\mathbf{y}_{0:t})$.

To implement Monte Carlo techniques for the on-line estimation problem, $\hat{p}(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$ with respect to $p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})$ needs to be sequentially computed. Since the state equation in our system follows the Markovian structure, we can implement the importance sampling recursively, which forms the basis of SMC methods (Liu and Chen, 1998; Doucet et al., 2000). Based on the Bayes' theorem, weights can be sequentially updated as follows:

$$w_t^{(i)} = \frac{p\left(\mathbf{x}_{0:t}^{(i)}\middle|\mathbf{y}_{0:t}\right)}{\pi\left(\mathbf{x}_{0:t}^{(i)}\middle|\mathbf{y}_{0:t}\right)} \propto \frac{p\left(\mathbf{x}_{0:t}^{(i)}\middle|\mathbf{y}_{0:t-1}\right)p\left(y_t\middle|\mathbf{x}_{0:t}^{(i)},\mathbf{y}_{0:t-1}\right)}{\pi\left(\mathbf{x}_{0:t-1}^{(i)}\middle|\mathbf{y}_{0:t-1}\right)\pi\left(\mathbf{x}_t^{(i)}\middle|\mathbf{x}_{0:t-1}^{(i)},\mathbf{y}_{0:t}\right)}$$
$$= w_{t-1}^{(i)} \frac{p\left(\mathbf{x}_t^{(i)}\middle|\mathbf{x}_{0:t-1}^{(i)},\mathbf{y}_{0:t-1}\right)p\left(y_t\middle|\mathbf{x}_{0:t}^{(i)},\mathbf{y}_{0:t-1}\right)}{\pi\left(\mathbf{x}_t^{(i)}\middle|\mathbf{x}_{0:t-1}^{(i)},\mathbf{y}_{0:t}\right)} \tag{2-5}$$

Applying Eq. (2-5) to the state space model in Eq. (2-2), we have:

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{f_{t-1}\left(\mathbf{x}_t^{(i)}\middle|\mathbf{x}_{t-1}^{(i)}\right)g_{t-1}\left(y_t\middle|\mathbf{x}_t^{(i)}\right)}{\pi\left(\mathbf{x}_t^{(i)}\middle|\mathbf{x}_{0:t-1}^{(i)},\mathbf{y}_{0:t}\right)} \tag{2-6}$$

where $f_{t-1}(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the probability density function (pdf) of $(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $g_{t-1}(y_t|\mathbf{x}_t)$ is the pdf of y_t given $\mathbf{x}_t$.

If we want to estimate the expectation of a function of $\mathbf{x}_{0:t}$, say $m(\mathbf{x}_{0:t})$, conditioning on $\mathbf{y}_{0:t}$, we have:

$$E(m(\mathbf{x}_{0:t})|\mathbf{y}_{0:t}) = \int m(\mathbf{x}_{0:t})p(\mathbf{x}_{0:t}|\mathbf{y}_{0:t})\,d\mathbf{x}_{0:t} \simeq \sum_{i=1}^{n} w_t^{(i)} m\left(\mathbf{x}_{0:t}^{(i)}\right) \tag{2-7}$$

There are two important issues regarding the design and implementation of the SMC algorithm. One is the selection of the importance density $\pi\left(\mathbf{x}_t^{(i)}\middle|\mathbf{x}_{0:t-1}^{(i)},\mathbf{y}_{0:t}\right)$ in Eq. (2-6). It is often convenient to choose it to be the prior $\pi\left(\mathbf{x}_t^{(i)}\middle|\mathbf{x}_{0:t-1}^{(i)},\mathbf{y}_{0:t}\right) = f_{t-1}\left(\mathbf{x}_t^{(i)}\middle|\mathbf{x}_{t-1}^{(i)}\right)$, which greatly simplifies the weights update in Eq. (2-6) as:

$$w_t^{(i)} \propto w_{t-1}^{(i)} g_{t-1}\left(y_t\middle|\mathbf{x}_t^{(i)}\right) \tag{2-8}$$

The other important issue is the use of *resampling*. The standard SMC algorithm suffers from the *degeneracy* phenomenon, where after a few iterations, all but one of the weights are very close to zero (Doucet et al., 2000). This degeneracy indicates the

majority of computational work for updating samples and weights are wasted. The resampling scheme is used to reduce such effect by eliminating particles with low weights and concentrating on those with large weights, so that all of the resampled particles can contribute significantly to the importance sampling estimates. A simple and common choice of resampling procedure is the *multinomial resampling*, which involves generating a new set of particles $\{\mathbf{x}_t^{(j)}\}_{j=1}^{n}$ by resampling with replacement n times from $\{\mathbf{x}_t^{(i)}\}_{i=1}^{n}$ according to their weights $\{w_t^{(i)}\}_{i=1}^{n}$, and then assigning equal weights $1/n$ to the new set of particles $\{\mathbf{x}_t^{(j)}\}_{j=1}^{n}$.

The above procedure of standard SMC algorithm with resampling is usually called the *particle filtering* algorithm and shown in Algorithm 2-1. In the following sections, we propose two improvements of the standard SMC algorithm. In section 2.2.2, a variance reduction method called Rao-Blackwellization is used to substantially lower the computational cost of the standard SMC algorithm. In section 2.2.3, we introduce a new resampling scheme to achieve better detection estimation of the steady state.

Algorithm 2-1. Standard SMC Algorithm

Initialization, for $i = 1, \ldots, n$:

- Sample particles $\mathbf{x}_0^{(i)} = (\boldsymbol{\xi}_0^{(i)}, \tau_0^{(i)})$, where $\boldsymbol{\xi}_0^{(i)} \sim q_{\boldsymbol{\xi}}(\cdot)$ and $\tau_0^{(i)} = 0$.
- Assign $w_0^{(i)} = 1/n$ to each particle $\mathbf{x}_0^{(i)}$.

For $t = 1, \ldots, T$:

- For $i = 1, \ldots, n$:
  - Draw samples $\mathbf{x}_t^{(i)}$ from the importance distribution $f_{t-1}\left(\mathbf{x}_t \middle| \mathbf{x}_{t-1}^{(i)}\right)$.
- Updating weights:

- Compute weights $w_t^{(i)}$ according to Eq. (2-8) and normalize the weights such that $\sum_{i=1}^{n} w_t^{(i)} = 1$.

♦ Resampling: resample $\{\mathbf{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^{n}$ to generate n equally weighted particles $\{\tilde{\mathbf{x}}_t^{(j)}, \frac{1}{n}\}_{j=1}^{n}$, and set $\{\mathbf{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^{n} = \{\tilde{\mathbf{x}}_t^{(j)}, \frac{1}{n}\}_{j=1}^{n}$.

2.2.2 Rao-Blackwellized SMC Algorithm

In this subsection, a Rao-Blackwellized version of the SMC algorithm is proposed. Its main idea is to marginalize out the parameters associated with each line segment and achieves substantial variance reduction of Monte Carlo estimates. The parameters of each line segment can be integrated out and estimated by taking advantage of the results from Bayesian linear regression. With Rao-Blackwellization (Casella and Robert, 1996; Doucet et al., 2001), the algorithm is much more efficient than the standard SMC algorithm.

To make proper inference of the state vector $\mathbf{x}_t$ which consists of four parameters, the common way for standard SMC algorithm is to obtain estimates based on their joint posterior distribution, namely $p(\boldsymbol{\beta}_t, \sigma_t^2, \tau_t | \mathbf{y}_{0:t})$, where $\boldsymbol{\beta}_t = (\beta_{0t}, \beta_{1t})$ (Wu et al., 2015). Although such method is effective and robust to detect the steady state, it is at the expense of high computational cost since a large number of particles have to be generated at each time step t to approximate the joint distribution of four variables. However, under appropriate choice of priors, $\boldsymbol{\beta}_t$ and σ_t^2 of each line segment can actually be integrated out, resulting in a particularly efficient algorithm with reduced variance of the estimates. More specifically, the unknown posterior density of interest can be factorized as follows:

$$p(\boldsymbol{\beta}_t, \sigma_t^2, \tau_t | \mathbf{y}_{0:t}) = p(\boldsymbol{\beta}_t, \sigma_t^2 | \tau_t, \mathbf{y}_{0:t}) P(\tau_t | \mathbf{y}_{0:t}) \tag{2-9}$$

in which the conditional posterior density $p(\boldsymbol{\beta}_t, \sigma_t^2 | \tau_t, \mathbf{y}_{0:t})$ can be solved analytically if we use the conjugate priors for $\boldsymbol{\beta}_t$ and σ_t^2. Consequently, estimating the joint distribution $p(\boldsymbol{\beta}_t, \sigma_t^2, \tau_t | \mathbf{y}_{0:t})$ requires Monte Carlo samples from only a one dimensional distribution $P(\tau_t | \mathbf{y}_{0:t})$, which can dramatically reduce the number of particles needed to reach a given estimation accuracy. In the remainder of this section, we will first discuss how to obtain the analytical solution for the conditional joint posterior density $p(\boldsymbol{\beta}_t, \sigma_t^2 | \tau_t, \mathbf{y}_{0:t})$. Then we will discuss how to sample τ_t based on another application of Rao-Blackwellization.

To derive the conditional joint posterior density $p(\boldsymbol{\beta}_t, \sigma_t^2 | \tau_t, \mathbf{y}_{0:t})$, we first factorize it as:

$$\begin{aligned} p(\boldsymbol{\beta}_t, \sigma_t^2 | \tau_t, \mathbf{y}_{0:t}) &= p(\boldsymbol{\beta}_t | \sigma_t^2, \tau_t, \mathbf{y}_{0:t}) p(\sigma_t^2 | \tau_t, \mathbf{y}_{0:t}) \\ &= p(\boldsymbol{\beta}_t | \sigma_t^2, \tau_t, \mathbf{y}_{\tau_t:t}) p\left(\sigma_t^2 \middle| \tau_t, \mathbf{y}_{\tau_t:t}\right) \end{aligned} \tag{2-10}$$

The last equation in Eq. (2-10) is due to the assumption that the model parameters between different line segments are independent. Using the conjugate prior, we assume that $\boldsymbol{\beta}_t | \sigma_t^2$ follows a normal distribution with its mean vector $\boldsymbol{\mu}_0$ and covariance matrix $\sigma_t^2 \boldsymbol{\Sigma}_0$, where $\boldsymbol{\Sigma}_0$ is a 2×2 symmetric positive definite matrix. The prior distribution of σ_t^2 is chosen to be the inverse gamma density with shape parameter $a_0 (a_0 > 1)$ and scale parameter $b_0 (b_0 > 0)$. Due to the conjugacy of normal and inverse gamma distributions and based on the results from Bayesian linear regression (O'Hagan, 1994), we have:

$$(\boldsymbol{\beta}_t | \sigma_t^2, \tau_t, \mathbf{y}_{\tau_t:t}) \sim N\left(\boldsymbol{\mu}_{\tau_t,t}, \sigma_t^2 \boldsymbol{\Sigma}_{\tau_t,t}\right)$$

$$\left(\sigma_t^2 \middle| \tau_t, \mathbf{y}_{\tau_t:t}\right) \sim IG(a_{\tau_t,t}, b_{\tau_t,t})$$

and the parameters $\boldsymbol{\mu}_{\tau_t,t}, \boldsymbol{\Sigma}_{\tau_t,t}, a_{\tau_t,t}, b_{\tau_t,t}$ are updated according to the following equations:

$$\begin{aligned} \boldsymbol{\Sigma}_{\tau_t,t} &= \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{D}_{\tau_t,t}^T \mathbf{D}_{\tau_t,t}\right)^{-1} \\ \boldsymbol{\mu}_{\tau_t,t} &= \boldsymbol{\Sigma}_{\tau_t,t}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{D}_{\tau_t,t}^T \mathbf{y}_{\tau_t:t}\right) \\ a_{\tau_t,t} &= a_0 + \frac{t-\tau_t+1}{2} \\ b_{\tau_t,t} &= b_0 + \frac{1}{2}\left(\mathbf{y}_{\tau_t:t}^T \mathbf{y}_{\tau_t:t} + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_{\tau_t,t}^T \boldsymbol{\Sigma}_{\tau_t,t}^{-1} \boldsymbol{\mu}_{\tau_t,t}\right) \end{aligned} \tag{2-11}$$

where $\mathbf{D}_{\tau_t,t} = \begin{bmatrix} 1 & \tau_t \\ 1 & \tau_t+1 \\ \vdots & \vdots \\ 1 & t \end{bmatrix}$.

When approaching the steady state, the signal can be characterized by a sustained 'flat' line segment. For this reason, the posterior knowledge of $\boldsymbol{\beta}_t$, especially that of the slope parameter β_{1t}, is very important for steady state detection. The marginal posterior distribution of $\boldsymbol{\beta}_t$ after integrating out σ_t^2 can be computed based on Lemma 2-1 as follows (the proof is included in APPENDIX A).

Lemma 2-1. $(\boldsymbol{\beta}_t|\tau_t, \mathbf{y}_{\tau_t:t})$ follows a bivariate non-standardized student's t distribution with degrees of freedom $2a_{\tau_t,t}$, location parameter $\boldsymbol{\mu}_{\tau_t,t}$, and scale matrix $\frac{b_{\tau_t,t}}{a_{\tau_t,t}}\boldsymbol{\Sigma}_{\tau_t,t}$, which is denoted as $(\boldsymbol{\beta}_t|\tau_t, \mathbf{y}_{\tau_t:t}) \sim t_2\left(\boldsymbol{\mu}_{\tau_t,t}, \frac{b_{\tau_t,t}}{a_{\tau_t,t}}\boldsymbol{\Sigma}_{\tau_t,t}, 2a_{\tau_t,t}\right)$.

To approximate the desired posterior density $p(\boldsymbol{\beta}_t, \sigma_t^2, \tau_t|\mathbf{y}_{0:t})$ using Eq. (2-9), we also need to generate Monte Carlo samples of τ_t, the latest change-point at the current time t, to approximate $P(\tau_t|\mathbf{y}_{0:t})$. Intuitively, we can generate each sample $\tau_t^{(i)}$ from its prior transition probability $P\left(\tau_t^{(i)}\middle|\tau_{t-1}^{(i)}\right)$ based on the idea of the standard SMC

algorithm. However, a more efficient way to sample $\tau_t^{(i)}$ can be done as follows, which is based on the application of Rao-Blackwellization to τ_t (Chopin, 2007):

- For each particle $\tau_{t-1}^{(i)}$ and its associated weight $w_{t-1}^{(i)}$, $i = 1, \dots, n$, create its two possible descendants at time t, with weights:

$$\begin{aligned} &\tau_t^{(i,1)} = \tau_{t-1}^{(i)}, w_t^{(i,1)} = w_{t-1}^{(i)} P\left(\tau_t^{(i)} = \tau_{t-1}^{(i)} \middle| \tau_{t-1}^{(i)}\right) p\left(y_t \middle| \tau_t^{(i)} = \tau_{t-1}^{(i)}, \mathbf{y}_{\tau_t^{(i)}:t-1}\right); \\ &\tau_t^{(i,2)} = t, w_t^{(i,2)} = w_{t-1}^{(i)} P\left(\tau_t^{(i)} = t \middle| \tau_{t-1}^{(i)}\right) p\left(y_t \middle| \tau_t^{(i)} = t, \mathbf{y}_{\tau_t^{(i)}:t-1}\right) \end{aligned} \tag{2-12}$$

The predictive density $p\left(y_t \middle| \tau_t^{(i)}, \mathbf{y}_{\tau_t^{(i)}:t-1}\right)$ in Eq. (2-12) can be calculated based on the following Lemma 2-2 (see APPENDIX B for the proof).

Lemma 2-2. Denote $\mathbf{X}_t = [1 \quad t]$, then:

(a). $\left(y_t | \tau_t, \mathbf{y}_{\tau_t:t-1}\right) \sim t_1\left(\mathbf{X}_t \boldsymbol{\mu}_{\tau_t,t-1}, \frac{b_{\tau_t,t-1}}{a_{\tau_t,t-1}}\left(1 + \mathbf{X}_t \boldsymbol{\Sigma}_{\tau_t,t-1} \mathbf{X}_t^T\right), 2a_{\tau_t,t-1}\right)$, if $\tau_t = \tau_{t-1}$;

(b). $\left(y_t | \tau_t, \mathbf{y}_{\tau_t:t-1}\right) \sim t_1\left(\mathbf{X}_t \boldsymbol{\mu}_0, \frac{b_0}{a_0}(1 + \mathbf{X}_t \boldsymbol{\Sigma}_0 \mathbf{X}_t^T), 2a_0\right)$, if $\tau_t = t$.

The resulted set of $2n$ particles after Rao-Blackwellzation of $\tau_t^{(i)}, i = 1, \dots, n$ can be resampled according to the weights $w_t^{(i,1)}, w_t^{(i,2)}, i = 1,2, \dots, n$ to obtain n resampled particles to avoid an exponentially increasing number of particles. By doing exact calculations on weights of all possible values of $\tau_t^{(i)}$, i.e., $\tau_t^{(i)} = \tau_{t-1}^{(i)}$ and $\tau_t^{(i)} = t$, the randomness inherent in the simulation of $\tau_t^{(i)}$ based on $P\left(\tau_t^{(i)} \middle| \tau_{t-1}^{(i)}\right)$ is removed, which leads to further variance reduction.

At time t, similar to the standard SMC algorithm, the resulted $2n$ particles can be used to approximate the true posterior distribution $P(\tau_t | \mathbf{y}_{0:t})$. Suppose

$\{\tau_t^{(i,j)}, w_t^{(i,j)}\}_{i=1}^{n}, j = 1,2$ are the $2n$ properly weighted samples with respect to $P(\tau_t|\mathbf{y}_{0:t})$ after the Rao-Blackwellization of $\tau_t^{(i)}$, then the posterior density $p(\boldsymbol{\beta}_t|\mathbf{y}_{0:t})$ can be approximated as:

$$\hat{p}(\boldsymbol{\beta}_t|\mathbf{y}_{0:t}) = \sum_{i=1}^{n}\sum_{j=1}^{2} w_t^{(i,j)}\, p\left(\boldsymbol{\beta}_t|\tau_t^{(i,j)}, \mathbf{y}_{\tau_t^{(i,j)}:t}\right) \tag{2-13}$$

where $p\left(\boldsymbol{\beta}_t|\tau_t^{(i,j)}, \mathbf{y}_{\tau_t^{(i,j)}:t}\right)$ can be obtained from Lemma 2-1.

It is well-known that the student's t distribution in Lemma 2-2 can be well-approximated by normal distribution with the same mean and variance when its degrees of freedom $2a_{\tau_t,t-1}$ (or $2a_0$) ≥ 30 (Li and Moor, 1999) to reduce the computational cost. In Section 2.3, we will use both the exact calculations for the pdf of student's t distribution and their normal approximations to study the detection performance.

In summary, the proposed Rao-Blackwellized version of our SMC algorithm is given in Algorithm 2-2 as follows.

Algorithm 2-2. Rao-Blackwellized SMC Algorithm

Initialization:

- Sample particles $\{\tau_0^{(i)}, w_0^{(i)}\}_{i=1}^{n}$ with $\tau_0^{(i)} = 0$ and $w_0^{(i)} = \frac{1}{n}$.

For $t = 1, \ldots, T$:

- For $i = 1, \ldots, n$:
  - Create two possible descendants $\tau_t^{(i,1)} = \tau_{t-1}^{(i)}$, $\tau_t^{(i,2)} = t$ of $\tau_{t-1}^{(i)}$, calculate their associated weights $w_t^{(i,1)}, w_t^{(i,2)}$ based on Eq. (2-12) and normalize the weights such that $\sum_{i=1}^{n}\sum_{j=1}^{2} w_t^{(i,j)} = 1$.

- Parameter estimation: estimate $p(\boldsymbol{\beta}_t|\mathbf{y}_{0:t})$ based on Eq. (2-13).
- Resampling: resample $\{\tau_t^{(i,j)}, w_t^{(i,j)}\}_{i=1}^{n}, j = 1,2$ to generate n equally weighted particles $\{\tilde{\tau}_t^{(k)}, \frac{1}{n}\}_{k=1}^{n}$, and set $\{\tau_t^{(i)}, w_t^{(i)}\}_{i=1}^{n} = \{\tilde{\tau}_t^{(k)}, \frac{1}{n}\}_{k=1}^{n}$.

2.2.3 Optimal Resampling

A simple resampling method for the Rao-Blackwellized SMC algorithm discussed in previous section is the multinomial resampling method. However, the resampled particles using multinomial resampling suffer from a significant loss of diversity (with many duplicate particles) because the particles with significant importance weights $w_t^{(i)}$ are repeatedly selected many times. Due to the discrete nature of the change points τ_t in our model, having duplicate particles is wasteful as they contain exactly the same information as a single particle with its weight equal to the sum of the weights of the duplicate particles.

Therefore, at each time t we will combine duplicate particles into a *single* particle so that we only have distinctive particles before the resampling step. Then we will apply the *Optimal Resampling* (OR) method (Fearnhead and Clifford, 2003) for the set of distinctive particles. Suppose at time $t-1$ we have n_d distinctive particles after resampling. Then after Rao-Blackwellization of $\tau_t^{(i)}$ at time t, there will be $2n_d$ particles $\{\tau_t^{(i,j)}, w_t^{(i,j)}\}_{i=1}^{n_d}, j = 1,2$, where $\tau_t^{(i,1)} = \tau_{t-1}^{(i)}$, $\tau_t^{(i,2)} = t$ and $\sum_{i=1}^{n_d}\sum_{j=1}^{2} w_t^{(i,j)} = 1$. By combining the duplicate particles $\{\tau_t^{(i,2)} = t, w_t^{(i,2)}\}_{i=1}^{n_d}$, we will have $n_d + 1$ distinctive particles $\{\tau_t^{(i)}, w_t^{(i)}\}_{i=1}^{n_d+1}$, where $\tau_t^{(i)} = \tau_t^{(i,1)}, w_t^{(i)} = w_t^{(i,1)}, i = 1,2,\dots,n_d; \tau_t^{(n_d+1)} = t, w_t^{(n_d+1)} = \sum_{i=1}^{n_d} w_t^{(i,2)}$. The following Algorithm 2-3, which is a special case of the OR method, can be used to resample n_d distinctive particles from the $n_d + 1$ distinctive

particles $\{\tau_t^{(i)}, w_t^{(i)}\}_{i=1}^{n_d+1}$.

The stratified sampling scheme used in the last step ensures each distinctive particle is resampled at most once so that the resulting n_d particles are distinctive. It is shown in Fearnhead and Clifford (2003) that the resampled n_d samples $\{\tilde{\tau}_t^{(i)}, \tilde{w}_t^{(i)}\}_{i=1}^{n_d}$ are still properly weighted with respect to the desirable posterior distribution $P(\tau_t|\mathbf{y}_{0:t})$. The OR method has the computational complexity of $O(n_d)$. And it is optimal (Fearnhead and Clifford, 2003) since it minimizes the expected squared error $E[\sum_{i=1}^{n_d+1}(W_t^{(i)} - w_t^{(i)})^2]$, where $W_t^{(i)}$ is the random weight of a particle after resampling ($W_t^{(i)}$equals $w_t^{(i)}$, $1/c$ or 0).

Algorithm 2-3. OR Algorithm

- Step 1: calculate the unique solution c to the equation: $\sum_{i=1}^{n_d+1} \min\left(1, cw_t^{(i)}\right) = n_d$.
- Step 2: for $i = 1, \ldots, n_d + 1$, if $w_t^{(i)} \geq 1/c$ then particle $\tau_t^{(i)}$ is kept with its original weight $w_t^{(i)}$. Assume we keep q particles.
- Step 3: apply the stratified sampling algorithm in APPENDIX C to resample $n_d - q$ particles from the remaining $n_d + 1 - q$ particles, each resampled particle is assigned a weight $1/c$.

Note that when t is small (e.g., t =1, 2…), only t distinctive particles are needed, corresponding to $\tau_t = 1, \ldots, t$, respectively. When t is large, we need to limit the maximum number of distinctive particles after resampling at each time t. Let n_d denote this maximum limit and n_t ($n_t \leq n_d + 1$) be the number of distinctive particles after

Rao-Blackwellization of $\tau_t^{(i)}$ at time t, with the properly weighted samples $\{\tau_t^{(i)}, w_t^{(i)}\}_{i=1}^{n_t}$ with respect to $P(\tau_t|\mathbf{y}_{0:t})$, the posterior density $p(\boldsymbol{\beta}_t|\mathbf{y}_{0:t})$ at time t can be approximated as:

$$\hat{p}(\boldsymbol{\beta}_t|\mathbf{y}_{0:t}) = \sum_{i=1}^{n_t} w_t^{(i)} p\left(\boldsymbol{\beta}_t|\tau_t^{(i)}, \mathbf{y}_{\tau_t^{(i)}:t}\right) \tag{2-14}$$

The Rao-Blackwellized SMC algorithm with the OR method is summarized in Algorithm 2-4.

Algorithm 2-4. Rao-Blackwellized SMC with Optimal Resampling

Set $n_t = 1$, $\tau_1^{(1)} = 1$ and $w_1^{(1)} = 1$.

For $t = 2, \ldots, T$:

- For $i = 1, \ldots, n_t$:
  - Create two possible descendants $\tau_t^{(i,1)} = \tau_{t-1}^{(i)}$, $\tau_t^{(i,2)} = t$ of each $\tau_{t-1}^{(i)}$, calculate their associated weights $w_t^{(i,1)}, w_t^{(i,2)}$ based on Eq. (2-12) and normalize the weights such that $\sum_{i=1}^{n_t}\sum_{j=1}^{2} w_t^{(i,j)} = 1$.
- Combine duplicate particles into $n_t + 1$ distinctive particles, the resulted samples are $\{\tau_t^{(i)}, w_t^{(i)}\}_{i=1}^{n_t+1}$, where $\tau_t^{(i)} = \tau_t^{(i,1)}, w_t^{(i)} = w_t^{(i,1)}, i = 1,2, \ldots, n_t; \tau_t^{(n_t+1)} = t, w_t^{(n_t+1)} = \sum_{i=1}^{n_t} w_t^{(i,2)}$. Set $n_t = n_t + 1$.
- Parameter estimation: estimate $p(\boldsymbol{\beta}_t|\mathbf{y}_{0:t})$ based on Eq. (2-14).
- Resampling: if $n_t = n_d + 1$, using the OR method in Algorithm 2-3 to obtain $\{\tilde{\tau}_t^{(i)}, \tilde{w}_t^{(i)}\}_{i=1}^{n_d}$; set $\{\tau_t^{(i)}, w_t^{(i)}\}_{i=1}^{n_t-1} = \{\tilde{\tau}_t^{(i)}, \tilde{w}_t^{(i)}\}_{i=1}^{n_d}$ and $n_t = n_d$.

2.2.4 Steady State Detection Rule

As the steady state can be characterized by a sustained 'flat' line segment in the piecewise linear model of signals, we can develop our detection rule based on the slope parameter of the last line segment of the piecewise linear model. Therefore, we define the steady state detection index as $\pi_t \equiv \Pr(|\beta_{1t}| \le s_0|\mathbf{y}_{0:t})$ where s_0 is the slope threshold. At time t, the detection index can be estimated as:

$$\hat{\pi}_t = \sum_{i=1}^{n_t} w_t^{(i)} \Pr\left(|\beta_{1t}| \le s_0 \middle| \tau_t^{(i)}, \mathbf{y}_{\tau_t^{(i)}:t}\right) \tag{2-15}$$

Based on Lemma 2-1, the marginal posterior distribution of β_{1t} follows a non-standardized student's t distribution. However, computing cumulative probability for the non-standardized student's t distribution is time-consuming. Note that the conditional posterior distribution of β_{1t} given σ_t^2 is:

$$\beta_{1t}\,|\sigma_t^2, \tau_t^{(i)}, \mathbf{y}_{\tau_t^{(i)}:t} \sim N\left(\mu^{(2)}_{\tau_t^{(i)},t}, \sigma_t^2 \boldsymbol{\Sigma}^{(2,2)}_{\tau_t^{(i)},t}\right)$$

where $\mu^{(2)}_{\tau_t^{(i)},t}$ is the second element of $\boldsymbol{\mu}_{\tau_t^{(i)},t}$, $\boldsymbol{\Sigma}^{(2,2)}_{\tau_t^{(i)},t}$ is the second diagonal element of $\boldsymbol{\Sigma}_{\tau_t^{(i)},t}$, and $\boldsymbol{\mu}_{\tau_t^{(i)},t}$ and $\boldsymbol{\Sigma}_{\tau_t^{(i)},t}$ are obtained using Eq. (2-11). If we replace σ_t^2 with its posterior mean value $E\left(\sigma_t^2|\tau_t^{(i)}, \mathbf{y}_{\tau_t^{(i)}:t}\right) = \frac{b_{\tau_t^{(i)},t}}{a_{\tau_t^{(i)},t}-1}$, where $a_{\tau_t^{(i)},t}$, $b_{\tau_t^{(i)},t}$ are obtained using Eq. (2-11), the marginal posterior distribution of $\beta_{1t}^{(i)}$ can be approximated by the normal distribution as:

$$\beta_{1t}|\tau_t^{(i)}, \mathbf{y}_{\tau_t^{(i)}:t} \dot{\sim} N\left(\mu^{(2)}_{\tau_t^{(i)},t}, \frac{b_{\tau_t^{(i)},t}}{a_{\tau_t^{(i)},t}-1} \boldsymbol{\Sigma}^{(2,2)}_{\tau_t^{(i)},t}\right)$$

Based on this approximation, the corresponding probabilities of each particle can be calculated much faster. From our experience $\hat{\pi}_t$ typically increases quickly and becomes close to one after the data reach the steady state. So the detection result is pretty robust to small errors in approximating $\hat{\pi}_t$.

At each time t, if $\hat{\pi}_t$ is larger than a given threshold value, π_0, we stop the algorithm and consider that a steady state is detected at time t. Similarly, since $\hat{\pi}_t$ typically increases quickly and becomes close to one after the data reach the steady state, the detection results are robust to small change of π_0. In this following of this chapter, we set π_0 to be 0.9 for all the examples.

2.3 Numerical Study

2.3.1 Simulated Signals and Model Parameter Setup

Simulations based on artificially generated signals are first conducted to evaluate the performance of our proposed steady state detection algorithm. A signal is generated based on the superposition of a bias functions and the noise. This chapter uses four types of bias functions: linear, quadratic, exponential and oscillating, as shown in Table 2-1. These are the most commonly tested functions for off-line steady state detection algorithms in the discrete-event simulation literature. For the bias direction of the first three functions, without loss of generality, we use the negative bias which represents the biased data starting below the steady state mean (Hoad et al., 2010). A linearly decreasing function is chosen for the amplitude of the oscillating signals. For the noises, we use three types of autoregressive model: random Gaussian error (AR(0)), AR(1) and AR(2), as shown in the following Table 2-2. The signal, $y(t)$, is generated based on the addition of the bias function $B(t)$ in Table 2-1 and the noise r_t in Table 2-2, namely

$y(t) = B(t) + r_t, t = 1, \dots, N.$

Table 2-1. Four types of bias functions

Signal	*Bias function*
Linear	$B(t) = \begin{cases} \frac{t}{T_0}h, & t = 1, \dots, T_0 \\ h, t = T_0 + 1, \dots, N \end{cases}$
Quadratic	$B(t) = \begin{cases} h\left(1 - \frac{(t-T_0)^2}{(T_0-1)^2}\right), & t = 1, \dots, T_0 \\ h, & t = T_0 + 1, \dots, N \end{cases}$
Exponential	$B(t) = \begin{cases} h\left(1 - 10^{-\frac{1-t}{T_0-1}}\right), & t = 1, \dots, T_0 \\ y(T_0), & t = T_0 + 1, \dots, N \end{cases}$
Oscillating	$B(t) = \begin{cases} h\frac{T_0-t}{T_0-1} sin\left(\frac{\pi t}{f}\right), & t = 1, \dots, T_0 \\ 0, & t = T_0 + 1, \dots, N \end{cases} (f = \frac{T_0}{10})$

For all signals T_0 is the true transition point to the steady state. In selecting the prior parameters, a non-informative prior with $\boldsymbol{\mu}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\boldsymbol{\Sigma}_0 = \begin{bmatrix} 10000 & 0 \\ 0 & 10000 \end{bmatrix}$ is used for $\boldsymbol{\beta}_t$. For the noise variance σ^2, its prior parameters a_0 and b_0 are set to be 10 and 0.1, respectively. For the prior transition probability p, any value between 0.1 and 0.5 can work well and here we use $p = 0.2$. The maximum number of distinctive particles after resampling is chosen to be $n_d = 16$ for all simulations in this chapter.

Table 2-2. Equations and parameter values for three types of noises

Types	*Equation*	*Parameter values*
AR(0)	$r_t = \varepsilon_t$	$\varepsilon_t \sim N(0, \sigma_t^2)$
AR(1)	$r_t = \varphi_1 r_{t-1} + \varepsilon_t$	$\varphi_1 = 0.6$
AR(2)	$r_t = \varphi_2 r_{t-1} + \varphi_3 r_{t-2} + \varepsilon_t$	$\varphi_2 = -0.25, \varphi_3 = 0.5$

2.3.2 Illustration of the Steady State Detection

We illustrate our proposed algorithm on steady state detection by using a simulated linear signal with noise level $\sigma = 0.06, h = 1$ and $T_0 = 200$. The slope threshold s_0 is set to be 0.0025. Figure 2-3 shows the corresponding detection process. It can be seen that $\hat{\pi}_t$ jumps abruptly to large values close to 1 shortly after T_0 (the true steady state transition point). Besides, at time $t = 100, 200$ and 500, the estimate of the posterior probability of the latest change-point, $\hat{P}(\tau_t|\mathbf{y}_{0:t})$, is almost concentrated near the true change-points: 1 ($t \leq 200$) or 200 ($t > 200$). Processing each signal with 500 time steps by the proposed algorithm only takes an average of 0.9 seconds in MATLAB running on a 3.40 GHz Intel processor, which is much lower than that of the PF method in Wu et al. (2015) (12 seconds for 500 time steps). Meanwhile, by sampling only one variable based on the Rao-Blackwellization method and employing the efficient OR resampling algorithm, our method uses at most 16 particles for each time step, which is substantially lower than the 1000 particles used for each time step by the PF method of Wu et al. (2015). In addition, using normal approximations for calculating the pdf of student's t distribution leads to further reduction of computational time to an average of 0.6 seconds and similar detection performance. This example shows that our algorithm can detect the change-point with timeliness and high computational efficiency.

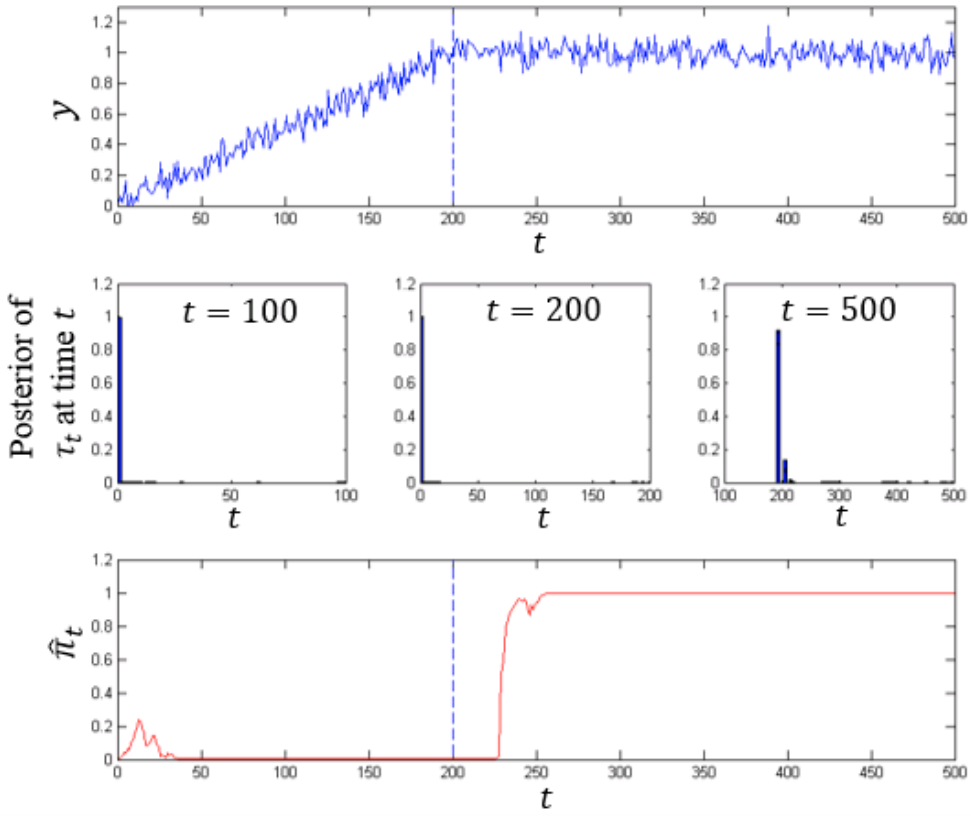


Figure 2-3. Illustration of steady state detection for linear signal: simulated signal (top), estimated posterior probability of the latest change-point at different time steps (middle), and the detection index $\hat{\pi}_t$ (bottom).

Many signals in practice have a decaying variance with a fixed mean. When the signal enters the steady state, the variance is small and stable. To see how well our algorithm performs to detect the transition to steady state for such signals, we simulate the signal with zero mean and the noise amplitude as follows:

$$\sigma(t) = \begin{cases} 30^{(T_0-t)/(T_0-1)}\sigma_0, \text{if } t \le T_0 \\ \sigma_0 \qquad\qquad\qquad\quad , \text{if } t > T_0 \end{cases}$$

where $\sigma_0 = 0.1$. The following Figure 2-4 shows that our detection algorithm is also well-performed for such a signal.

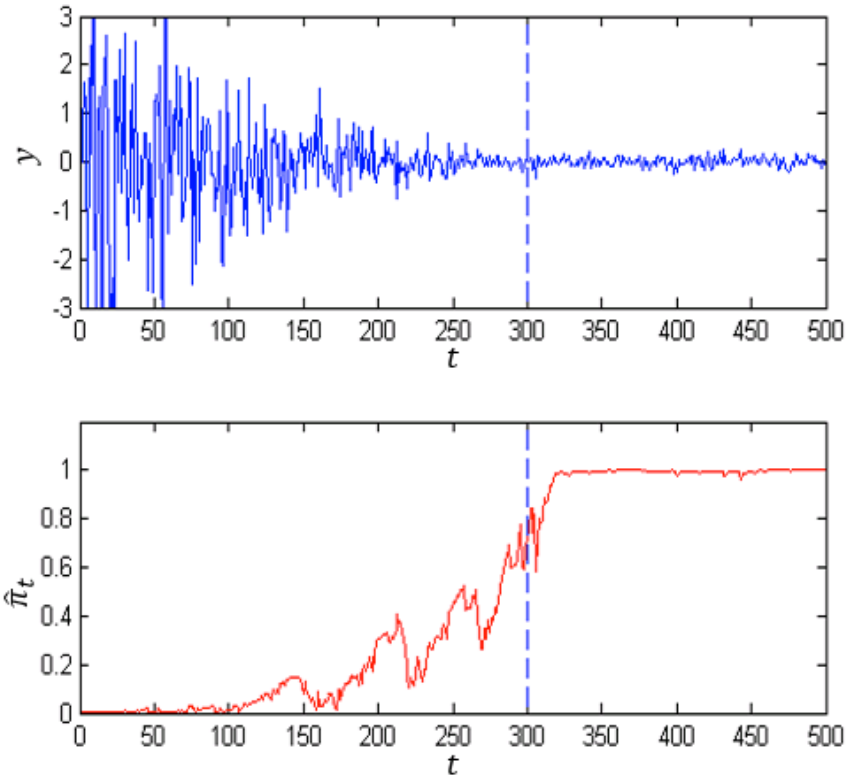


Figure 2-4. Steady state detection for a signal with zero mean and exponentially decaying variance: simulated signal (top), and the detection index $\hat{\pi}_t$ (bottom).

2.3.3 Performance Evaluation

Following Wu et al. (2015), we consider the *false alarm rates* (FAR) and deviation of the estimated steady state transition point $\hat{\tau}$ from the true transition point T_0 (also called *detection bias*) to evaluate the performance of the proposed on-line steady state detection algorithm. The FAR is the probability that $\hat{\tau} < T_0$. It is considered because in some situations the cost of early detection is higher than that of delayed detection. The FAR usually serves as an auxiliary evaluation metric. The main evaluation metric in this chapter is the *weighted standard detection error* (WSDE), as defined in Wu et al. (2015):

$$WSDE = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s} w(\hat{\tau}_i)\,(\hat{\tau}_i - T_0)^2} \qquad w(\hat{\tau}_i) = \begin{cases} w \in (0,1], \hat{\tau}_i \geq T_0 \\ 1, \qquad \hat{\tau}_i < T_0 \end{cases} \qquad (2\text{-}16)$$

where $\hat{\tau}_i$ denotes the estimated change-point in the i^{th} replication and N_s is the total number of replications used for each type of signals in the simulation. When w = 1, WSDE is the root mean square deviation (RMSD) of $\hat{\tau}_i$ from T_0. When w < 1, larger penalty is given to early detection than late detection in WSDE.

2.3.4 Comparison with Existing On-Line Methods

In this section, we compare the proposed method in this chapter with three other existing on-line steady state detection methods: the PF method (Wu et al., 2015), the SDM method (Holly et al., 1989; Bethea and Rhinehart, 1991; Wu et al., 2013), and the VRT method (Crow et al., 1960; Cao and Rhinehart, 1995). The PF method (Wu et al., 2015) is the most recently developed on-line steady state detection method in the literature. It is based on the standard SMC, or the particle filter algorithm, and incorporates several improvement strategies such as the partial Gibbs resample-moves technique. The other two methods, the SDM and VRT, both incorporate a moving data window based on which they estimate either the slope or the variance to determine if the signal enters the steady state. In this section, we refer to the proposed Rao-Blackwellized SMC method in this chapter as RBSMC method.

Each type of signal is generated by combining a bias function in Table 2-1 and a noise type in Table 2-2. For the bias function, we set $h = 1$ and test two values of T_0: $T_0 = 200$ and $T_0 = 300$. Three noise amplitudes $\sigma_t = 0.06, 0.1, 0.14$ are used for AR(0) and $\sigma_t = 0.06, 0.1$ are used for AR(1) and AR(2). As in Wu et al. (2015), we use

$N_s = 500$ replications for each type of signals and the detection parameters (window size, thresholds) of all methods are selected to minimize the WSDE with $w = 1$. For our proposed method, we set slope threshold $s_0 = 0.0021$ for all signals.

Figure 2-5 shows the WSDE as functions of the penalty weight w for each noise autoregressive type. As we can see, the proposed RBSMC method and PF are much more accurate than SDM, VRT in terms of the WSDE in all penalty weights. Besides, the proposed RBSMC method is slightly better comparing with the PF method in most of the penalty weights.

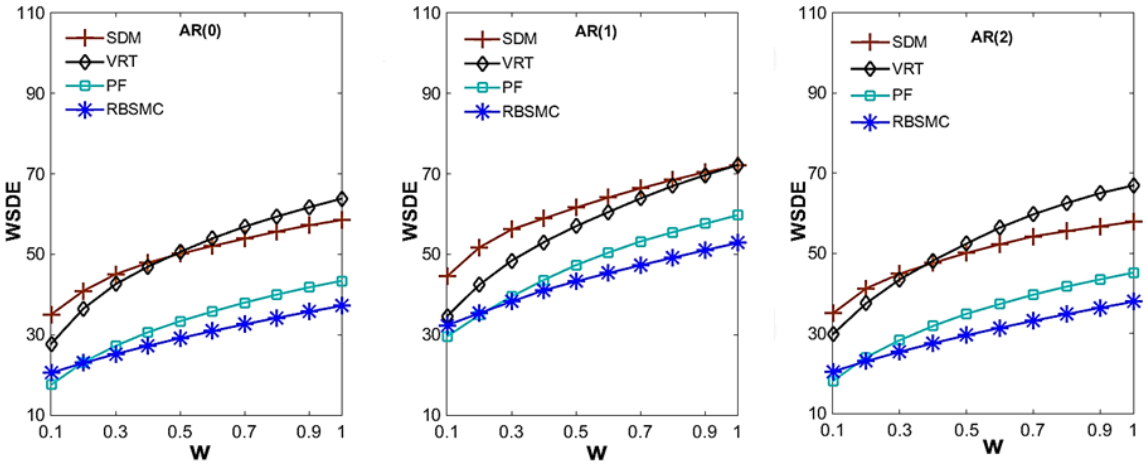


Figure 2-5. The weighted standard detection error (WSDE) of the proposed RBSMC, PF, SDM, VRT as a function of the penalty weight w for AR(0), AR(1) and AR(2).

The detailed results for the AR(0) noise with $w = 1$ are shown in Table 2-3. From Table 2-3, it can be seen that the two SMC methods (RBSMC and PF) offer the most competitive detection results in terms of the WSDE and FAR with signals of various bias functions. Both of them have consistently better performance than the moving-window based methods (SDM and VRT).

Table 2-3. Comparison of RBSMC, PF, SDM and VRT for w = 1 and noise type AR(0). The detection parameters are (1) RBSMC, $s_0 = 0.0021$; (2) PF, $s_0 = 0.0022$; (3) SDM, window size $m = 50$, threshold=8×10^{-5}; (4) VRT, $m = 98$, threshold=0.6

Signal		σ	WSDE				FAR			
			RBS-MC	PF	SDM	VRT	RBS-MC	PF	SDM	VRT
Linear	h=1, T_0=200	0.06	52.6	42.0	59.8	78.6	0	0	0	0
		0.10	55.7	53.9	60.7	66.1	0	0	0	0
		0.14	59.6	64.5	57.6	60.4	0	0	0	0
	h=1, T_0=300	0.06	46.7	40.9	58.0	70.9	0	0	0	0
		0.10	49.9	53.1	55.9	56.6	0	0	0	0.04
		0.14	53.5	70.4	58.6	121.6	0.01	0.01	0.03	0.75
Quadratic	h=1, T_0=200	0.06	17.6	12.1	33.6	37.6	0.01	0.11	0	0
		0.10	22.8	21.2	31.9	26.7	0.03	0.04	0.02	0.04
		0.14	26.3	33.7	28.2	18.5	0.06	0.06	0.14	0.36
	h=1, T_0=300	0.06	20.0	33.8	22.4	16.3	0.89	1	0.11	0.33
		0.10	24.4	28.1	23.9	37.5	0.81	0.93	0.45	0.83
		0.14	29.9	22.4	34.5	72.8	0.71	0.62	0.62	0.93
Exponential	h=1, T_0=200	0.06	23.3	16.6	45.9	44.4	0.01	0.12	0	0
		0.10	28.1	26.3	40.5	23.2	0.07	0.08	0.03	0.30
		0.14	31.7	35.0	34.8	26.1	0.12	0.06	0.17	0.73
	h=1, T_0=300	0.06	37.6	61.7	35.5	27.2	0.93	1	0.04	0.51
		0.10	45.4	49.4	32.2	67.3	0.85	0.98	0.38	0.96
		0.14	51.1	40.1	55.7	107.8	0.85	0.88	0.82	0.99
Oscillating	h=1, T_0=200	0.06	23.1	27.1	94.9	74.4	0	0.01	1	0
		0.10	29.3	27.6	90.2	61.5	0	0.04	0.99	0
		0.14	38.7	26.6	94.0	54.1	0	0.06	1	0
	h=1, T_0=300	0.06	9.4	23.0	156	63.6	0.39	0.04	1	0
		0.10	19.5	25.7	156	49.8	0.21	0.2	0.99	0
		0.14	32.6	29.4	152	40.6	0.07	0.55	1	0
Overall			37.3	39.1	59.9	60.3	0.25	0.28	0.38	0.28

The advantage of the SMC-based methods over the moving-window based methods can be understood intuitively as follows: The SMC-based methods are based on the piecewise linear model in Eq. (2-1) with multiple unknown change-points. Therefore they behave as methods with adaptive ‘window’ sizes. Compared with the moving-window based methods with a fixed ‘window’ size, SMC-based methods are very

flexible in adjusting their 'window' sizes based on the observed signals to give much more robust detection performance.

Comparing the RBSMC and PF methods, the RBSMC has slightly better overall performance in terms of smaller WSDE (37.3 vs. 39.1) and smaller FAR (0.25 vs. 0.28), and using normal approximations for calculating the pdf of student's t distribution in RBSMC leads to the same overall performance (see APPENDIX D for detailed results). Most importantly, the main advantage of the RBSMC method to the PF method is in the substantial saving of computational cost. This significant saving of computational cost makes the RBSMC method a much preferred method for many on-line applications where quick processing of the data in real time is critical.

2.3.5 Application to Steady State Detection in Micro/Nanoparticle Dispersion Process

In this section, we apply the RBSMC algorithm to real signals called *cavitation noise power* (CNP) signals from the ultrasonic-cavitation based nanoparticle dispersion process. Currently, the micro/nanoparticle research has attracted intense scientific interest because of its potential applications in biomedical, optical and electrical fields. In these applications, micro/nanoparticles need to be dispersed evenly into the base materials before use to improve the material properties. However, the particles often cluster together as a result of high surface energy and large surface-to-volume ratio. The ultrasonic cavitation method can be used for effective dispersion of micro/nanoparticles. The dispersion process can be monitored by detecting the steady state of CNP signals based on the fact that the steady state of CNP signals corresponds to the maximum dispersion extent at the ultrasonic power level. Please refer to Wu et al. (2013) for details

on the ultrasonic cavitation process, the experimental setup used to collect the data, and the method to obtain the CNP signals.

Figure 2-6 shows the detection results for the CNP signals with ultrasonic power 30W and 50W, respectively, in the dispersion of 30 g Al_2O_3. The off-line method EWMA-MSER (Wu et al., 2013) is used as a benchmark method to evaluate the proposed method. It can be seen that the detection results (red solid line) of the proposed RBSMC method are quite close to those of the off-line method (black dashed line).

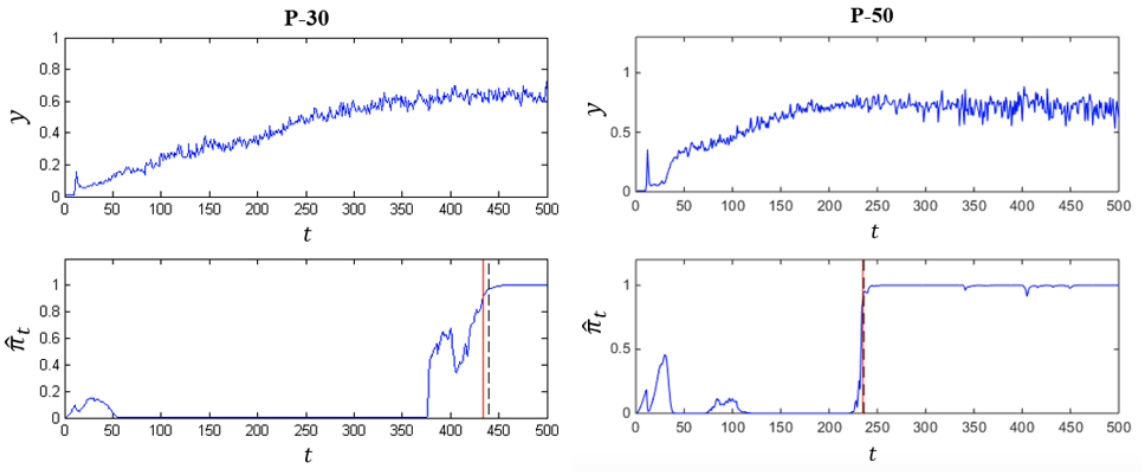


Figure 2-6. Steady state detection for the CNP signal with ultrasonic power 30W and 50W in the dispersion of 30 g Al_2O_3.

2.4 Summary

In this chapter, we study the problem of on-line steady state detection using a multiple change-point model and sequential Monte Carlo methods. A piecewise linear model is used to approximate the signal. Within the Bayesian framework, the posterior densities of model parameters can be sequentially updated given the latest observations. The stopping criterion for detecting the steady state is established based on the fact that the steady state can be characterized by a sustained 'flat' line segment in the piecewise

linear model.

The main contribution of our proposed algorithm is its high computational efficiency based on the Rao-Blackwellization technique. By solving analytically for the conditional distribution of the model parameters, estimating the joint posterior distribution of four variables in the state vector requires sampling from only a one dimensional posterior distribution. This leads to substantial variance reduction of the Monte Carlo estimates and the use of significantly smaller number of particles than the standard SMC algorithm, while achieving comparable or better estimation accuracy. In addition, by applying the so-called Optimal Resampling method and eliminating duplicate particles, the robustness and timeliness of steady state detection is significantly improved by using the information of the particles more efficiently.

The performance of our proposed method is evaluated through both artificially simulated signals and a real data example from the ultrasonic-cavitation based nanoparticle dispersion process. Results demonstrate the robustness of the proposed algorithm for various types of signals with different levels of noises, and much faster computational time compared to the standard PF method.

CHAPTER 3. A NEW BAYESIAN ON-LINE INFERENCE METHOD FOR THE SHORT-RUN SPC

In short production runs of the statistical quality control process, the major challenge is that there exists little or no historical in-control data. The data necessary to accurately estimate the process parameters and the control limits for standard control charts are usually not available prior to a production run. In this chapter, we propose new methods under the Bayesian framework to detect on-line whether the mean of the short-run process has exceeded a critical threshold. Our model modifies the original model proposed by Tsiamyrtzis and Hawkins (2005) and can be more flexible in handling linear trends of the process. Under the modified modeling framework, the exact Bayesian method based on Fearnhead and Liu (2007) can be used to obtain a significantly reduced number of mixtures in the posterior distribution of process mean, which requires substantially less computational and memory cost. Numerical studies based on simulated signals and application to degradation data are used to evaluate the performance of the proposed method and compare with the method proposed by Tsiamyrtzis and Hawkins (2005). Our proposed method is shown to establish a more robust estimation performance while being less sensitive to parameter misspecifications and much more computationally efficient.

The remainder of this chapter is organized as follows. A new model formulation of the short-run process is introduced in Section 3.1. Sections 3.2 and 3.3 give a detailed description on the exact Bayesian method and the corresponding approximate method based on particle filter for efficient detection of the out-of-control state of the process

mean. Extensive simulation studies as well as application to degradation data are presented in Section 3.4. A summary is provided in Section 3.5.

3.1 Motivation and Model Formulation

In traditional control charts, such as the CUSUM control chart, we are mainly concerned with whether the process remains in control at the target value θ_0 or shifts to some out-of-control value of the mean $\theta_1(\theta_1 \neq \theta_0)$. Assuming that the in-control state is stable, we have a sequence of hypothesis tests at different time points:

$$\begin{cases} H_0: \theta_t = \theta_0 \\ H_1: \theta_t = \theta_1 \end{cases}, t = 1, 2, \ldots$$

where θ_t is the process mean at time t. In standard control chart methods, the value of θ_0 is estimated through a long phase I data gathering. The control limits can be obtained based on θ_0 and θ_1 to help decide if the process is out-of-control at some time point.

However, for the unstable in-control state where the in-control process mean may have continuous small changes over time, the aforementioned framework can no longer be used. As an extension, Tsiamyrtzis and Hawkins (2005) considered making a sequence of more generic decisions between:

$$\begin{cases} H_0: \theta_t \leq M \\ H_1: \theta_t > M \end{cases}, t = 1, 2, \ldots$$

where M is the pre-specified upper threshold of θ_t. Thus our concern at each time point becomes whether the mean of the short-run process has exceeded a critical threshold so that some corrective action is needed. Although this is a one-sided test assuming the jumps are upward, it can be easily adapted to the case where the jumps are downward.

Suppose the noisy signal is denoted as $\mathbf{y}_{1:N} = (y_1, y_2, \dots, y_N)$, where N is the total length of the signal. To provide a theoretical basis that helps make the decision at each time, Tsiamyrtzis and Hawkins (2005) introduced a random walk plus a fixed size random jump model (referred to as *TH model* in the rest of this chapter), which is:

$$\theta_t|\theta_{t-1} \sim \begin{cases} N(\theta_{t-1}, \sigma_{RW}^2) & \text{with probability } 1-p \\ N(\theta_{t-1}+\delta, \sigma_{RW}^2) & \text{with probability } p \end{cases}, \quad (3\text{-}1)$$

$$y_t = \theta_t + \varepsilon_t, \varepsilon_t \sim N(0, \sigma^2)$$

The TH model suggests that the mean of the short-run process drifts according to a normal random walk while being subject to some occasional jumps with probability p. σ_{RW}^2 is the random walk variance and δ is the fixed jump size that is assumed as known. σ^2 quantifies the measurement variability of the process. Under the Bayesian framework, this model leads to a closed-form solution of the posterior distribution $p(\theta_t|\mathbf{y}_{1:t})$, based on which a decision rule is obtained. It is shown to be effective in detecting whether the mean exceeds the critical threshold in the short-run process.

However, the TH model suffers from two major limitations. Firstly, it is computationally intensive. Tsiamyrtzis and Hawkins (2005) derived the posterior distribution at each time t as $p(\theta_t|\mathbf{y}_{1:t}) = \sum_{i=0}^{2^t-1} \alpha_i^{(t)} N\left(\theta_i^{(t)}, \hat{\sigma}_t^2\right)$, which is a mixture of 2^t normal distributions. Please refer to Theorem 1 of Tsiamyrtzis and Hawkins (2005) for the detailed expressions for $\alpha_i^{(t)}$, $\theta_i^{(t)}$, and $\hat{\sigma}_t^2$. The exponentially increasing number of mixtures is a result of the dependence of θ_t before and after each of the occasional jumps in model (3-1). It requires demanding computational and memory cost. For example, even when t is moderately small, say $t = 25$, implementing the TH method can cause a computer to run out of memory. Secondly, it assumes the jump size δ is known. And

whether the jump size is correctly specified can actually lead to very different estimation performance. In real practice, however, such assumption is usually too restrictive since we seldom have the exact knowledge of the jump size.

To overcome these drawbacks, in this chapter we propose a new model that leads to highly efficient calculations of the posterior distribution $p(\theta_t|\mathbf{y}_{1:t})$, while requiring no exact knowledge of the jump size. Specifically, we assume:

$$\theta_t|\theta_{t-1}\sim\begin{cases}N(\theta_{t-1},\sigma_{RW}^2) & \text{with probability } 1-p\\ N(\mu_J,\sigma_J^2) & \text{with probability } p\end{cases}, \qquad (3\text{-}2)$$

$$y_t=\theta_t+\varepsilon_t,\varepsilon_t\sim N(0,\sigma^2)$$

Rather than imposing a fixed and known jump size, our model in Eq. (3-2) makes a more general assumption on the jump size by considering it follows a normal distribution with mean μ_J and a large variance σ_J^2 that are independent of θ_{t-1}. The assignment of large values to σ_J^2 represents the vagueness of our knowledge about the jump, which is more close to the reality since in practice the magnitude of an occasional jump is typically unknown. Additionally, based on our model, it can be seen that the process mean before and after an occasional jump are independent, which enables us to take advantage of the exact Bayesian method (Fearnhead and Liu, 2007) to dramatically reduce the computational cost in the posterior inference of the process mean.

As discussed previously, when a process is subject to degradations, the data often exhibit some linear trends, or piecewise linear trends. Therefore, we generalize our model (3-2) to a more flexible model that can also handle linear trends on the basis of a piecewise linear formulation, which is expressed as:

$$(\boldsymbol{\beta}_t, \sigma_t^2)|(\boldsymbol{\beta}_{t-1}, \sigma_{t-1}^2) = \begin{cases} (\boldsymbol{\beta}_{t-1}, \sigma_{t-1}^2) & \text{with probability } 1-p \\ (\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) & \text{with probability } p \end{cases}, \quad (3\text{-}3)$$

$$y_t = \theta_t + \varepsilon_t = \mathbf{x}_t \boldsymbol{\beta}_t + \varepsilon_t, \varepsilon_t \sim N(0, \sigma_t^2)$$

where $\boldsymbol{\beta}_t = \begin{bmatrix} \beta_{1t} \\ \beta_{0t} \end{bmatrix}$, $\theta_t = \mathbf{x}_t \boldsymbol{\beta}_t$, $\mathbf{x}_t = [t \quad 1]$, and $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ are random samples generated independently from some prior distributions $\pi_{\boldsymbol{\beta}}(\cdot)$ and $\pi_{\sigma}(\cdot)$, respectively. In model (3-3), the observation y_t follows a piecewise linear trend while being subject to some occasional jumps with randomly changing slopes, intercepts and variance. The β_{1t} and β_{0t} correspond to the slope and intercept at time t. The measurement noise variance σ_t^2 at time t varies across different line segments (between occurrence of occasional jumps), while being unchanged within each line segment. If little knowledge is available for $\tilde{\sigma}^2$ and $\tilde{\boldsymbol{\beta}}$ after an occasional jump, non-informative priors can be used for $\pi_{\boldsymbol{\beta}}(\cdot)$ and $\pi_{\sigma}(\cdot)$.

Based on our newly proposed model (3-2) and model (3-3), in the following sections, we will have detailed discussions on the exact Bayesian method (Fearnhead and Liu, 2007) for efficient inference of the process parameters and make decisions on whether the process mean crosses a critical threshold.

3.2 On-Line Exact Bayesian Inference of the Process Mean

3.2.1 On-Line Exact Bayesian Inference on the Posterior of Occasional Jumps

In this subsection, we will focus on making inference of the process mean θ_t under the Bayesian framework, which facilitates our sequential decisions on whether the process mean has exceeded a critical threshold. Specifically, to make decisions at time t, first we need to obtain the posterior distribution of the process mean $p(\theta_t|\mathbf{y}_{1:t})$.

Conditioning on possible values of τ_t, the location of the latest occasional jumps up to time t, we have:

$$\begin{aligned} p(\theta_t|\mathbf{y}_{1:t}) &= \sum_{s=1}^{t} P(\tau_t = s|\mathbf{y}_{1:t})\, p(\theta_t|\tau_t = s, \mathbf{y}_{1:t}) \\ &= \sum_{s=1}^{t} P(\tau_t = s|\mathbf{y}_{1:t})\, p(\theta_t|\tau_t = s, \mathbf{y}_{s:t}) \end{aligned} \tag{3-4}$$

The last equation in Eq. (3-4) is due to our model assumption that process mean before and after an occasional jump are independent. If an occasional jump occurs at time t, we have $\tau_t = t$; otherwise $\tau_t = \tau_{t-1}$. Based on our model (3-2) and model (3-3), it is easy to see that the prior transition probability of τ_t is $P(\tau_t = t|\tau_{t-1}) = p$.

As indicated by Eq. (3-4), we need to know the terms $P(\tau_t = s|\mathbf{y}_{1:t})$ and $p(\theta_t|\tau_t = s, \mathbf{y}_{s:t})$ in order to obtain $p(\theta_t|\mathbf{y}_{1:t})$. The first term $P(\tau_t = s|\mathbf{y}_{1:t}),\ s = 1, \dots, t$ can be calculated based on Fearnhead and Liu (2007), who studied multiple change-point detection problems under the assumption that model parameters before and after the change-point are *independent*. An exact Bayesian (EB) method is developed in Fearnhead and Liu (2007) to find posterior distributions of the latest change-points. Since both our model (3-2) and model (3-3) satisfy the independent assumption in Fearnhead and Liu (2007), their method can be applied for inference of τ_t. Denote observations from time s to t as $\mathbf{y}_{s:t} = (y_s, y_{s+1}, \dots, y_t)$ and the posterior probabilities of τ_t at time t as $w_t^s = P(\tau_t = s|\mathbf{y}_{1:t}), s = 1, \dots, t$. Based on the results from Fearnhead and Liu (2007), each w_t^s can be sequentially updated as:

$$w_t^s \propto \begin{cases} \dfrac{P(s,t)}{P(s,t-1)} \cdot (1-p) \cdot w_{t-1}^s & \text{if } s < t \\ P(t,t) \cdot p & \text{if } s = t \end{cases} \tag{3-5}$$

where $P(s,t) = P(\mathbf{y}_{s:t}|\mathbf{y}_{s:t}$ are in the same line segment) and its calculation will be discussed later. For the second term $p(\theta_t|\tau_t = s, \mathbf{y}_{s:t})$ involved in Eq. (3-4), detailed discussions on finding its analytical form are presented in the following subsections based on model (3-2) and model (3-3), respectively.

3.2.2 On-Line Inference for the Random Walk Plus Random Jump Model

For our model (3-2), we assume the measurement noise variance σ^2 is known and is constant over time. This is the same assumption made in the TH method. The densities $p(\theta_t|\tau_t = s, \mathbf{y}_{s:t}), s = 1, \dots, t$ can be obtained via the Kalman filter (KF) (Kalman, 1960), which is shown in the following Lemma 3-1 (the proof can be also found in APPENDIX E).

Lemma 3-1. For model (3-2), let $\mu_{\theta_{s,t}}, \sigma^2_{\theta_{s,t}}, k_{s,t}$ be parameters associated with the posterior distribution of the process mean for observations $\mathbf{y}_{s:t}$. Then:

(1) If $s < t$: $p(\theta_t|\tau_t = s, \mathbf{y}_{s:t}) \sim N\left(\mu_{\theta_{s,t}}, \sigma^2_{\theta_{s,t}}\right)$ with $\mu_{\theta_{s,t}} = k_{s,t-1}\mu_{\theta_{s,t-1}} + (1 - k_{s,t-1})y_t$ and $\sigma^2_{\theta_{s,t}} = \sigma^2(1 - k_{s,t-1})$, where $k_{s,t-1} = \frac{\sigma^2}{\sigma^2_{\theta_{s,t-1}} + \sigma^2_{RW} + \sigma^2}$; $\frac{P(s,t)}{P(s,t-1)} = p(y_t|\mathbf{y}_{s:t-1}) \sim N\left(\mu_{\theta_{s,t-1}}, \sigma^2_{RW} + \sigma^2 + \sigma^2_{\theta_{s,t-1}}\right)$.

(2) If $s = t$: $p(\theta_t|\tau_t = t, y_t) \sim N\left(\mu_{\theta_{t,t}}, \sigma^2_{\theta_{t,t}}\right)$ with $\mu_{\theta_{t,t}} = k_{t,t}\mu_J + (1 - k_{t,t})y_t$ and $\sigma^2_{\theta_{t,t}} = \sigma^2(1 - k_{t,t})$, where $k_{t,t} = \frac{\sigma^2}{\sigma_J^2 + \sigma^2_{RW} + \sigma^2}$; $P(t,t) = p(y_t) \sim N(\mu_J, \sigma^2_{RW} + \sigma^2 + \sigma_J^2)$.

where the terms $\frac{P(s,t)}{P(s,t-1)}$ and $P(t,t)$ are used for updating w_t^s in Eq. (3-5).

In this chapter, we refer to our exact Bayesian method using KF as the EBKF method. For the EBKF method, based on Lemma 3-1, Eq. (3-4) becomes:

$$p(\theta_t|\mathbf{y}_{1:t}) = \sum_{s=1}^{t} w_t^s N\left(\mu_{\theta_{s,t}}, \sigma^2_{\theta_{s,t}}\right) \tag{3-6}$$

Eq. (3-6) suggests that the posterior distribution of the process mean in the EBKF contains a mixture of only t normal distributions at each time t, which substantially lowers the computational and memory cost compared to the exponentially increasing number of mixtures in the TH method. Therefore, the EBKF is much more preferable in the situation where the run length of the process is not very small.

To help make the decision on whether the process mean has exceeded the critical threshold M at time t, we adopt the same decision rule as in the TH method where we calculate the posterior probability $P_t = \Pr(\theta_t > M|\mathbf{y}_{1:t})$ and accept the alternative hypothesis iff $P_t \geq \alpha$ (α is a cutoff value between 0.5 and 1). The process is continued until the null hypothesis is rejected, at which time we stop the process and take some corrective actions. Mathematically, based on Eq. (3-6), the decision rule in the EBKF can be obtained as:

$$P_t = 1 - \Pr(\theta_t \leq M|\mathbf{y}_{1:t}) = 1 - \sum_{s=1}^{t} w_t^s \Phi\left(\frac{M - \mu_{\theta_{s,t}}}{\sigma_{\theta_{s,t}}}\right) \tag{3-7}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

In summary, our proposed EBKF method based on model (3-2) for detecting on-line whether the process man has exceeded a critical threshold is given in the following Algorithm 3-1 as follows.

Algorithm 3-1. The EBKF for Detecting On-Line Whether θ_t has Crossed a Critical Threshold

1. At $t = 1$, initializing $P_1 = 0$ and $P(\tau_1 = 1|y_1) = 1$.
2. For $t = 2, \cdots, N$
   - For $s = 1, \cdots, t$

     Calculate the un-normalized w_t^s based on Lemma 3-1 and Eq. (3-5).

     End
   - Normalize w_t^s $(s = 1, \cdots, t)$ such that $\sum_{s=1}^{t} w_t^s = 1$.
   - Calculate the detection index P_t based on Eq. (3-7); if $P_t \geq \alpha$, the process mean is considered exceeding the critical threshold and the algorithm is terminated.

   End

3.2.3 On-Line Inference for the Piecewise Linear Model

To handle linear trends in a process, we use the piecewise linear model formulation in Eq. (3-3), which satisfies the independent assumption made in Fearnhead and Liu (2007). So the exact Bayesian method can be applied for posterior inference in Eq. (3-3). Since the measurement noise variance σ_t^2 is assumed as unknown in Eq. (3-3), we refer to this method as EBPLUV. In EBPLUV, each linear segment is assigned a common conjugate prior distribution of σ_t^2 and $\boldsymbol{\beta}_t$ using inverse Gamma and Gaussian distribution as:

$$\sigma_t^2 \sim \pi_\sigma = IG\left(\frac{v}{2}, \frac{\gamma}{2}\right)$$

$$\boldsymbol{\beta}_t | \sigma_t^2 \sim \pi_{\boldsymbol{\beta}} = N(\boldsymbol{\mu}_0, \sigma_t^2 \boldsymbol{\Sigma}_0)$$

where $v, \gamma, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ are specified prior parameters. If little knowledge is available for σ_t^2 and $\boldsymbol{\beta}_t$ at the beginning of each new line segment, the prior parameters can be selected to make the prior distributions relatively non-informative.

Define the design matrix $\mathbf{X}_t^s = \begin{bmatrix} s & 1 \\ s+1 & 1 \\ \vdots & \vdots \\ t & 1 \end{bmatrix}$ and

$$\begin{aligned} \mathbf{M}_t^s &= ((\mathbf{X}_t^s)^T \mathbf{X}_t^s + \boldsymbol{\Sigma}_0^{-1})^{-1} \\ \mathbf{N}_t^s &= \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + (\mathbf{X}_t^s)^T \mathbf{y}_{s:t}^T \\ \boldsymbol{\mu}_t^s &= \mathbf{M}_t^s \mathbf{N}_t^s \\ d_t^s &= t - s + v + 1 \ (d_t^s > 2) \\ H_t^s &= \mathbf{y}_{s:t} \mathbf{y}_{s:t}^T + \gamma + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 - (\mathbf{N}_t^s)^T \mathbf{M}_t^s \mathbf{N}_t^s \end{aligned} \tag{3-8}$$

The densities $p(\theta_t | \tau_t = s, \mathbf{y}_{s:t}), s = 1, \ldots, t$ in Eq. (3-4) can be obtained analytically based on the following Lemma 3-2 (the proof is included in APPENDIX F).

Lemma 3-2. Suppose $\sigma_t^2 \sim IG\left(\frac{v}{2}, \frac{\gamma}{2}\right)$ and $\boldsymbol{\beta}_t | \sigma_t^2 \sim N(\boldsymbol{\mu}_0, \sigma_t^2 \boldsymbol{\Sigma}_0)$, we have:

(1) $(\boldsymbol{\beta}_t | \sigma_t^2, \tau_t = s, \mathbf{y}_{s:t}) \sim N(\boldsymbol{\mu}_t^s, \sigma_t^2 \mathbf{M}_t^s)$ and $(\sigma_t^2 | \tau_t = s, \mathbf{y}_{s:t}) \sim IG\left(\frac{d_t^s}{2}, \frac{H_t^s}{2}\right)$.

(2) $(\theta_t | \tau_t = s, \mathbf{y}_{s:t})$ follows a non-standardized student's t distribution with degrees of freedom d_t^s, mean $\mathbf{x}_t \boldsymbol{\mu}_t^s$ and variance $\frac{H_t^s}{d_t^s} \mathbf{x}_t \mathbf{M}_t^s \mathbf{x}_t^T$, which is denoted by $(\theta_t | \tau_t = s, \mathbf{y}_{s:t}) \sim t\left(\mathbf{x}_t \boldsymbol{\mu}_t^s, \frac{H_t^s}{d_t^s} \mathbf{x}_t \mathbf{M}_t^s \mathbf{x}_t^T, d_t^s\right)$.

where $\mathbf{M}_t^s, \boldsymbol{\mu}_t^s, H_t^s$ and d_t^s are defined in Eq. (3-8). Additionally, $P(s,t)$ can be

calculated as:

$$P(s,t) = \pi^{-(t-s+1)/2}\left(\frac{|\mathbf{M}_t^s|}{|\boldsymbol{\Sigma}_0|}\right)^{\frac{1}{2}} \frac{\gamma^{\frac{v}{2}}}{(H_t^s)^{(t-s+1+v)/2}} \frac{\Gamma\left(\frac{t-s+1+v}{2}\right)}{\Gamma(v/2)} \tag{3-9}$$

where $\Gamma(z)$ is the gamma function defined as $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$. The derivations of $P(s,t)$ can be found in APPENDIX G.

Based on Lemma 3-2, the posterior distribution of the process mean in the EBPLUV is a mixture of t non-standardized student's t distributions, which can be expressed as:

$$p(\theta_t|\mathbf{y}_{1:t}) = \sum_{s=1}^{t} w_t^s t\left(\mathbf{x}_t\boldsymbol{\mu}_t^s, \frac{H_t^s}{d_t^s}\mathbf{x}_t\mathbf{M}_t^s\mathbf{x}_t^T, d_t^s\right) \tag{3-10}$$

Following the same decision rule, in the EBPLUV we have:

$$P_t = 1 - \Pr(\theta_t \le M|\mathbf{y}_{1:t}) = 1 - \sum_{s=1}^{t} w_t^s F_{\mathrm{ST},d_t^s}\left(\frac{M - \mathbf{x}_t\boldsymbol{\mu}_t^s}{\sqrt{\frac{H_t^s}{d_t^s}\mathbf{x}_t\mathbf{M}_t^s\mathbf{x}_t^T}}\right) \tag{3-11}$$

where $F_{\mathrm{ST},d_t^s}(\cdot)$ is the CDF of the standard student's t distribution with degrees of freedom d_t^s.

In summary, our proposed EBPLUV method based on model (3-3) for detecting on-line whether the process mean has exceeded a critical threshold is given in the following Algorithm 3-2 as follows.

Algorithm 3-2. The EBPLUV for Detecting On-Line Whether θ_t has Crossed a Critical Threshold

1. At $t = 1$, initializing $P_1 = 0$ and $P(\tau_1 = 1|y_1) = 1$.
2. For $t = 2, \cdots, N$
   - For $s = 1, \cdots, t$

     Calculate the un-normalized w_t^s based on Eq. (3-5) and Eq. (3-9).

     End
   - Normalize w_t^s $(s = 1, \cdots, t)$ such that $\sum_{s=1}^{t} w_t^s = 1$.
   - Calculate the detection index P_t based on Eq. (3-11); if $P_t \geq \alpha$, the process mean is considered exceeding the critical threshold and the algorithm is terminated.

   End

3.2.4 On-Line Inference for Constant Noise Variance

In the previous subsection, the EBPLUV method is introduced for handling linear trends in the process where the noise variance is allowed to be changed to a very different value after an occasional jump. However, when the noise variance is actually constant over time (its value is still unknown), we will see that the EBPLUV method may not perform as well as expected for signals with multiple linear segments. As illustrated in Figure 3-1, we generate a simple piecewise linear signal with slope changing at time 200 and noises having a constant standard deviation $\sigma_t = \sigma = 0.01$. Result of the estimated

mean tracking of the EBPLUV method (black dashed line in the left plot) shows its slow response to the change of slope and a poor estimation of the mean for the time period immediately after the slope changes.

This is indeed a consequence of using a non-informative prior for the noise variance at the beginning of each new segment in the EBPLUV method. For the example in Figure 3-1, at a time t right after 200 (say, $t \in (200, 250)$), $P(\tau_t = 200|\mathbf{y}_{1:t})$ is much smaller than $P(\tau_t = 1|\mathbf{y}_{1:t})$ because of the non-informative prior for noise variance assumed right after the occasional jump at $t = 200$. Using a non-informative prior for $\tau_t = 200$ fails to utilize the information on σ^2 accumulated during $t < 200$. Therefore if the noise variance is known to be unchanged over time, we propose to revise the EBPLUV method based on the following idea: in order to fully utilize the information on noise variance accumulated in the previous segments, at each time the prior of noise variance for a new segment is obtained based on the posterior information from the previous segments. For a new segment $\mathbf{y}_{s:t}$, its prior of noise variance σ_s^2 can be obtained based on the posterior density $p(\sigma_{s-1}^2|\mathbf{y}_{1:s-1})$. Then, based on Eq. (3-5) and Eq. (3-9), to have a closed form solution for $P(s,t)$, which is needed for updating $w_t^s = P(\tau_t = s|\mathbf{y}_{1:t})$, the prior distribution of σ_s^2 for a new segment needs to follow an inverse Gamma distribution. However, if $p(\sigma_{s-1}^2|\mathbf{y}_{1:s-1})$ is used as the prior distribution, it generally does not follow an inverse Gamma distribution because $p(\sigma_{s-1}^2|\mathbf{y}_{1:s-1})$ depends on the previous change points. Therefore, in this chapter we propose a simple and effective approximation strategy. Denote v_s and γ_s as prior parameters of noise variance at time s, then in our proposed strategy the prior of noise variance σ_s^2 is approximated with the inverse Gamma distribution given by:

$$p(\sigma_{s-1}^2|\mathbf{y}_{1:s-1}) \approx IG\left(\frac{d_{s-1(v_{s^*})}^{s^*}}{2}, \frac{H_{s-1(\gamma_{s^*})}^{s^*}}{2}\right), s^* = \underset{s_0 \in \{1,2,\ldots,s-1\}}{\operatorname{argmax}} w_{s-1(v_{s_0},\gamma_{s_0})}^{s_0}, \quad (3\text{-}12)$$

where $d_{s-1(v_{s^*})}^{s^*}, H_{s-1(\gamma_{s^*})}^{s^*}, s^* \in \{1, 2, \ldots, s-1\}$ can be obtained by substituting v and γ in the last two equations of Eq. (3-8) with v_{s^*} and γ_{s^*}:

$$\begin{aligned} d_{s-1(v_{s^*})}^{s^*} &= (s-1) - s^* + v_{s^*} + 1 \\ H_{s-1(\gamma_{s^*})}^{s^*} &= \mathbf{y}_{s^*:(s-1)}\mathbf{y}_{s^*:(s-1)}^T + \gamma_{s^*} + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \left(\mathbf{N}_{s-1}^{s^*}\right)^T\mathbf{M}_{s-1}^{s^*}\mathbf{N}_{s-1}^{s^*} \end{aligned} \quad (3\text{-}13)$$

And $w_{s-1(v_{s_0},\gamma_{s_0})}^{s_0}, s_0 \in \{1, 2, \ldots, s-1\}$ in Eq. (3-12) can be obtained based on Eq. (3-5):

$$w_{s-1(v_{s_0},\gamma_{s_0})}^{s_0} \propto \begin{cases} \dfrac{P_{(v_{s_0},\gamma_{s_0})}(s_0, s-1)}{P_{(v_{s_0},\gamma_{s_0})}(s_0, s-2)} \cdot (1-p) \cdot w_{s-2(v_{s_0},\gamma_{s_0})}^{s_0} & \text{if } s_0 < s-1 \\ P_{(v_{s-1},\gamma_{s-1})}(s-1, s-1) \cdot p & \text{if } s_0 = s-1 \end{cases} \quad (3\text{-}14)$$

where $P_{(v_s,\gamma_s)}(s,t) = \pi^{-(t-s+1)/2}\left(\frac{|\mathbf{M}_t^s|}{|\boldsymbol{\Sigma}_0|}\right)^{\frac{1}{2}} \frac{\gamma_s^{\frac{v_s}{2}}}{\left(H_{t(\gamma_s)}^s\right)^{(t-s+1+v_s)/2}} \frac{\Gamma\left(\frac{t-s+1+v_s}{2}\right)}{\Gamma(v_s/2)}$, which is obtained from Eq. (3-9) by substituting v, γ, and H_t^s with v_s, γ_s, and $H_{t(\gamma_s)}^s$, respectively.

The idea of Eq. (3-12) is to approximate $p(\sigma_{s-1}^2|\mathbf{y}_{1:s-1})$ with the posterior distribution of σ_{s-1}^2 based on the data that are most likely from the last segment at time $s-1$ (represented by $\mathbf{y}_{s^*:s-1}$). By implementing this approximation, on the one hand, we can utilize the information on noise variance accumulated from previous segments; on the other hand, the resulting prior of noise variance σ_s^2 at each time s still follows an inverse Gamma distribution, which allows us to recursively update the prior of noise variance across different time points. That is, at time $s = 1$, we set the initial prior of noise variance as $v_1 = v$ and $\gamma_1 = \gamma$. Then, for $s \geq 2$, based on Eq. (3-12), prior of $\sigma_s^2 \sim IG\left(\frac{v_s}{2}, \frac{\gamma_s}{2}\right)$, where $v_s = d_{s-1(v_{s^*})}^{s^*}$ and $\gamma_s = H_{s-1(\gamma_{s^*})}^{s^*}$, $s^* \in \{1, 2, \ldots, s-1\}$.

Throughout the rest of this chapter, we refer to the revised EBPLUV method for the cases with constant and unknown variance as the EBPLUV-EV method. The EBPLUV-EV method is able to utilize the information on noise variance accumulated in the previous segments with similar computational cost as EBPLUV. Similar to the EBPLUV method, the posterior distribution of the process mean in the EBPLUV-EV method is still a mixture of t non-standardized student's t distributions:

$$p(\theta_t|\mathbf{y}_{1:t}) = \sum_{s=1}^{t} w^s_{t(v_s,\gamma_s)} t\left(\mathbf{x}_t\boldsymbol{\mu}^s_t, \frac{H^s_{t(\gamma_s)}}{d^s_{t(v_s)}}\mathbf{x}_t\mathbf{M}^s_t\mathbf{x}^T_t, d^s_{t(v_s)}\right) \tag{3-15}$$

The corresponding decision rule is similar to Eq. (3-11), by replacing d^s_t, H^s_t, w^s_t with $d^s_{t(v_s)}, H^s_{t(\gamma_s)}, w^s_{t(v_s,\gamma_s)}$, respectively.

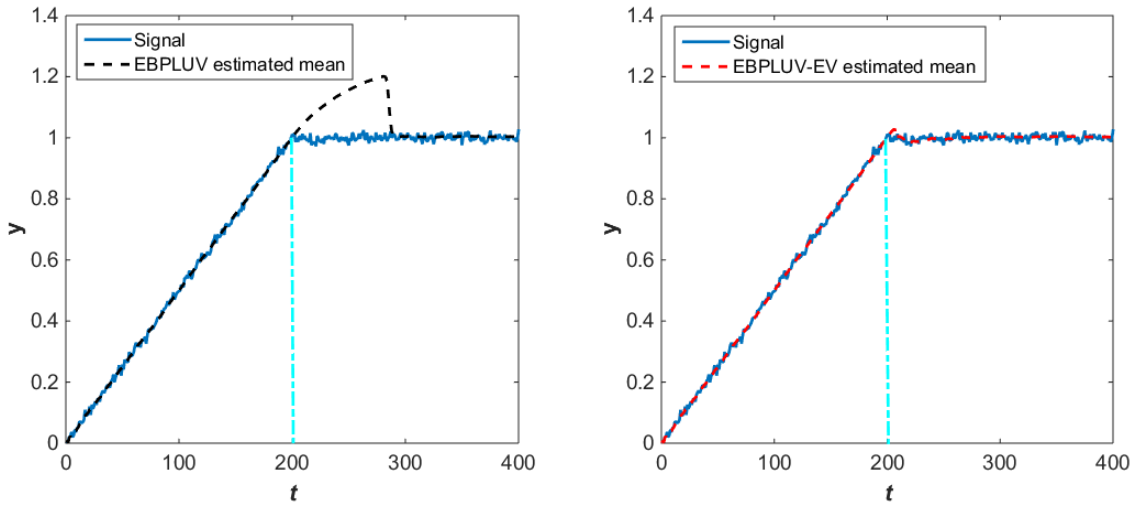


Figure 3-1. Estimated mean tracking performance of EBPLUV and EBPLUV-EV methods.

In summary, our proposed EBPLUV-EV method based on model (3-3) for detecting on-line whether the process mean has exceeded a critical threshold in signals with multiple linear trends and constant noise variance is given in Algorithm 3-3. As shown in Figure 3-1 (red dashed line in the right plot), the EBPLUV-EV method results

in a much faster response to the occasional jump at $t = 200$ and much more accurate estimation of the mean than EBPLUV.

Algorithm 3-3. The EBPLUV-EV for Detecting On-Line Whether θ_t has Crossed a Critical Threshold in Signals with Multiple Linear Trends and Constant Noise Variance

1. At $t = 1$, initializing $P_1 = 0$ and $w_1^1 = P(\tau_1 = 1|y_1) = 1$, set prior parameters $v_1 = v$ and $\gamma_1 = \gamma$. Obtain $d_{1(v_1)}^1$ and $H_{1(\gamma_1)}^1$ based on Eq. (3-13), and let $v_2 = d_{1(v_1)}^1, \gamma_2 = H_{1(\gamma_1)}^1$.

2. For $t = 2, \cdots, N$

   - For $s = 1, \cdots, t$

     Obtain $w_{t(v_s,\gamma_s)}^s$ based on Eq. (3-14).

     End

   - Normalize $w_{t(v_s,\gamma_s)}^s$ $(s = 1, \cdots, t)$ such that $\sum_{s=1}^t w_{t(v_s,\gamma_s)}^s = 1$.

   - Find $s^* = \underset{s\in\{1,2,\ldots,t\}}{\operatorname{argmax}}\, w_{t(v_s,\gamma_s)}^s$, and set $v_{t+1} = d_{t(v_{s^*})}^{s^*}$, $\gamma_{t+1} = H_{t(\gamma_{s^*})}^{s^*}$.

   - Calculate the detection index P_t; if $P_t \geq \alpha$, the process mean is considered exceeding the critical threshold and the algorithm is terminated.

   End

3.3 Approximated On-Line Inference of the Process Mean

The EB-based methods (EBKF, EBPLUV and EBPLUV-EV) presented in the previous section calculate the posterior probabilities of all possible τ_t at each time t.

Although the algorithm is much more efficient than the TH method, its complexity still increases quadratically with t. Therefore, for large t, the computational and memory cost may still become prohibitive. The particle filtering algorithm with Optimal Resampling in Algorithm 2-3 can be applied to our EB-based methods to make the complexity linear over time. Concretely, suppose at time t, the location of the latest occasional jumps τ_t has a set of n_t values $\tau_t^1, \tau_t^2, \cdots, \tau_t^{n_t}$ that are treated as particles, and their associated normalized posterior probabilities are $w_t^{\tau_t^1}, w_t^{\tau_t^2}, \ldots, w_t^{\tau_t^{n_t}}$ are considered as weights, in the context of particle filtering algorithm. By applying the PF algorithm with OR to our EB-based methods, at each time we would have at most n_d particles ($n_t = n_d + 1$).

The following Algorithm 3-4 presents a general framework for implementing the PF algorithm with OR for our original EB-based methods (EBKF, EBPLUV and EBPLUV-EV). Throughout the rest of this chapter, we refer to these approximated EB-based methods as AEBKF, AEBPLUV and AEBPLUV-EV, respectively.

Algorithm 3-4. A General Framework of PF Algorithm with OR for Our EB-Based Methods

1. At $t = 1$, initializing $n_t = 1$, $\tau_t^{n_t} = 1, P_1 = 0$ and $P\left(\tau_t^{n_t} = 1 \middle| y_1\right) = 1$. If using EBPLUV-EV method: obtain $d_{1(v_1)}^1$ and $H_{1(\gamma_1)}^1$ based on Eq. (3-13), and let $v_2 = d_{1(v_1)}^1, \gamma_2 = H_{1(\gamma_1)}^1$.
2. For $t = 2, \cdots, N$
   - Set $n_t = n_t + 1$ and $\tau_t^{n_t} = t$.
   - For $i = \tau_t^1, \tau_t^2, \ldots, \tau_t^{n_t}$

Calculate the un-normalized w_t^i when using EBKF or EBPLUV, or $w_{t(v_i,\gamma_i)}^i$ using Eq. (3-14) when applying EBPLUV-EV.

End

- Normalize the weights such that their sum is 1.
- If using EBPLUV-EV: find $s^* = \underset{s\in\{\tau_t^1,\tau_t^2,\ldots,\tau_t^{n_t}\}}{\operatorname{argmax}} w_{t(v_s,\gamma_s)}^s$, and let $v_{t+1} = d_{t(v_{s^*})}^{s^*}$, $\gamma_{t+1} = H_{t(\gamma_{s^*})}^{s^*}$.
- If $n_t = n_d + 1$, apply the OR algorithm in Algorithm 2-3 to sample n_d particles of τ_t and set $n_t = n_d$; otherwise skip to the next step.
- Calculate the detection index P_t; if $P_t \geq \alpha$, the process mean is considered exceeding the critical threshold and the algorithm is terminated.

End

3.4 Simulation Study

3.4.1 Data Generated from the Random Walk Plus Random Jump Model

In this section, simulations are carried out to evaluate and compare the performance of our proposed methods and TH method (Tsiamyrtzis and Hawkins, 2005) using the data generated from model (3-1). The goal of this subsection is two-folds. Firstly, we would like to evaluate our proposed methods and TH method from the perspective of robustness in estimation performance, by using both correct and misspecified model parameters for inference. Since the TH method is exactly designed for the data generated from model (3-1), we expect that the TH method performs well when true model parameters are used for inference. However, when model parameters

used for inference are incorrect, performance of TH method may be adversely affected. Note that the data generated from Eq. (3-1) do not satisfy the independence assumption of model (3-2) where parameters before and after occasional jumps are independent. Using the data generated from model (3-1) can also test how our model performs when the independence assumption of the data is not satisfied. The second goal of this subsection is to use simulations to demonstrate the computational advantages of our proposed methods over the TH method, which has been shown analytically in our previous discussions.

In the rest of this chapter, we consistently use the same simulation parameter settings, which are described as below. The maximum distinctive number of τ_t after resampling for approximated EB-based methods is $n_d = 10$. For EBPLUV, EBPLUV-EV and their approximated versions, we use relatively non-informative priors for the parameters with $\boldsymbol{\mu}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_0 = 1{\times}10^4\mathbf{I}$, and $v = \gamma = 4$. All simulations in this chapter are executed using MATLAB 2014b on a 3.40 GHz Intel processor.

3.4.1.1 Estimation Accuracy

We compare our EBKF method with the TH method. Both methods assume that the measurement noise variance σ^2 is known, which is fixed to be 1 in this study. The probability threshold of the detection index (P_t) is set as $\alpha = 0.7$. The TH method assumes the jump size δ is known; while the EBKF does not make this assumption and use a non-informative prior for δ with mean $\mu_J = 0$ and variance $\sigma_J^2 = 100$. The varying model parameters of the simulated data are as follows: $p = (0.1, 0.3), \sigma_{RW}^2 = (0.05, 1, 5)$, and $\delta = m\sigma_{RW}, m = (1, 3, 6)$ providing 18 different scenarios for the model parameters. The initial process mean θ_0 of the simulated data follows the normal

distribution with $\theta_0 \sim N(0, \sigma_{RW}^2)$. For each of the parameter scenarios, $S = 1000$ simulations are conducted. Inference of the process mean is based on the correct parameter specifications. The critical threshold M is defined as the 75th percentile of all simulated process mean values for each simulation.

The performance evaluations of different methods are based on comparing time n_1 when the true process mean exceeds M and time n_2 when the estimated process mean exceeds M, with 5 different cases which are defined by Tsiamyrtzis and Hawkins (2005):

(a). No signal (NS) cases: $n_1 = n_2 = 0$.

(b). Correct alarm (CA) cases: $n_1 = n_2 > 0$.

(c). Incorrect timing (IA) alarm cases: $n_1 > 0, n_2 > 0, n_1 \neq n_2$.

(d). False alarm (FA) cases: $n_1 = 0, n_2 > 0$.

(e). Missed alarm (MA) cases: $n_1 > 0, n_2 = 0$.

Table 3-1 gives the corresponding results using data length $N = 10$. Comparison of the values of n_1 and n_2 is done by recording the percentage of times that the above five cases are taking place out of 1000 simulations. As can be seen, performance of the EBKF method is only slightly worse than the TH method in terms of the overall detection accuracy (percentage of $n_1 = n_2$). The reason that EBKF is slightly worse than TH is that the EBKF method does not use any knowledge on the jump size δ, while the TH method uses all exact information of the data including δ.

Although the TH method performs quite well when using the correct parameters for inference, in practical situations, the exact knowledge of model parameters is rarely known. Therefore, using the correct parameter specifications cannot provide us the whole picture of the estimation performance of different methods. A more critical component in

the performance evaluation is the ability to achieve good detection accuracy with less sensitivity to the parameter misspecifications. In the following, we will carry out another simulation study to compare EBKF and TH methods using both correct and incorrect parameter specifications.

The parameter settings are exactly the same as the previous simulation study. However, in order to compare methods under both correct and incorrect parameter specifications, this time we define two types of parameter settings:

(a). True model parameter (TMP) setting: it is the true parameter setting used to generate all simulated signals.

(b). Inference model parameter (IMP) setting: it is the parameter setting used by a particular inference method. In our study, the TH method uses all 18 possible parameter settings to draw inference on the process mean, while the EBKF method uses only 6 parameter settings since its inference is based on only p and σ_{RW}^2 and does not assume known jump size δ.

The following Table 3-2 shows the simulation results of EBKF and TH methods with different IMP settings, using the TMP setting $p = 0.1, \sigma_{RW}^2 = 0.05, m = 1$. It can be seen that the TH method is more sensitive to the parameter misspecifications. When the IMP settings are more deviated from the TMP setting, detection performance of the TH method is more adversely affected. In contrast, EBKF is less sensitive to the parameter misspecifications and therefore it outperforms the TH method in terms of the overall detection accuracy. Moreover, from Table 3-2, for some parameter setting (e.g., $p = 0.3, \sigma_{RW}^2 = 0.05, m = 6$), the correct detection rate of TH can be as much as 37.5% lower than that of EBKF, while for all settings TH is no more than 9% lower than EBKF.

Table 3-1. Simulation results of TH and EBKF using the correct parameter specifications with $S = 1000, N = 10$

Model parameters			Methods	$n_1 = n_2$		$n_1 \neq n_2$		
p	σ^2_{RW}	m		**NS**	**CA**	**IA**	**FA**	**MA**
0.1	0.05	1	TH	0.552	0.01	0.195	0.013	0.23
			EBKF	0.547	0.018	0.185	0.018	0.232
0.1	0.05	3	TH	0.544	0.021	0.233	0.014	0.188
			EBKF	0.548	0.008	0.184	0.01	0.25
0.1	0.05	6	TH	0.533	0.055	0.269	0.018	0.125
			EBKF	0.538	0.029	0.227	0.013	0.193
0.1	1	1	TH	0.54	0.113	0.249	0.014	0.084
			EBKF	0.543	0.104	0.251	0.011	0.091
0.1	1	3	TH	0.528	0.197	0.194	0.02	0.061
			EBKF	0.537	0.168	0.217	0.011	0.067
0.1	1	6	TH	0.566	0.256	0.133	0.01	0.035
			EBKF	0.568	0.236	0.152	0.008	0.036
0.1	5	1	TH	0.544	0.252	0.156	0.006	0.042
			EBKF	0.543	0.251	0.155	0.007	0.044
0.1	5	3	TH	0.517	0.336	0.115	0.002	0.03
			EBKF	0.516	0.323	0.121	0.003	0.037
0.1	5	6	TH	0.554	0.347	0.079	0.007	0.013
			EBKF	0.555	0.331	0.095	0.006	0.013
0.3	0.05	1	TH	0.497	0.026	0.248	0.015	0.214
			EBKF	0.48	0.037	0.224	0.032	0.227
0.3	0.05	3	TH	0.439	0.056	0.334	0.022	0.149
			EBKF	0.433	0.06	0.286	0.028	0.193
0.3	0.05	6	TH	0.432	0.139	0.316	0.024	0.089
			EBKF	0.44	0.118	0.277	0.016	0.149
0.3	1	1	TH	0.532	0.164	0.234	0.009	0.061
			EBKF	0.53	0.157	0.24	0.011	0.062
0.3	1	3	TH	0.441	0.29	0.201	0.012	0.056
			EBKF	0.443	0.255	0.217	0.01	0.075
0.3	1	6	TH	0.44	0.444	0.083	0.006	0.027
			EBKF	0.442	0.4	0.118	0.004	0.036
0.3	5	1	TH	0.499	0.307	0.146	0.004	0.044
			EBKF	0.5	0.302	0.15	0.003	0.045
0.3	5	3	TH	0.463	0.4	0.102	0.01	0.025
			EBKF	0.464	0.374	0.116	0.009	0.037
0.3	5	6	TH	0.447	0.495	0.044	0.003	0.011
			EBKF	0.448	0.468	0.063	0.002	0.019
Overall			TH	0.504	0.217	0.185	0.012	0.082
			EBKF	0.504	0.202	0.182	0.012	0.1
			TH	0.721		0.279		
			EBKF	0.706		0.294		

Table 3-2. Simulation results of TH and EBKF using different IMP settings and the TMP setting of $p = 0.1, \sigma_{RW}^2 = 0.05, m = 1$

Inference parameters			Methods	$n_1 = n_2$		$n_1 \neq n_2$		
p	σ_W^{22}	m		**NS**	**CA**	**IA**	**FA**	**MA**
0.1	0.05	1	TH	0.52	0.003	0.188	0.022	0.267
			EBKF	0.522	0.005	0.173	0.02	0.28
0.1	0.05	3	TH	0.491	0.024	0.25	0.051	0.184
			EBKF	0.522	0.005	0.173	0.02	0.28
0.1	0.05	6	TH	0.459	0.041	0.263	0.083	0.154
			EBKF	0.522	0.005	0.173	0.02	0.28
0.1	1	1	TH	0.322	0.076	0.327	0.228	0.047
			EBKF	0.351	0.073	0.312	0.199	0.065
0.1	1	3	TH	0.31	0.079	0.329	0.24	0.042
			EBKF	0.351	0.073	0.312	0.199	0.065
0.1	1	6	TH	0.345	0.073	0.315	0.205	0.062
			EBKF	0.351	0.073	0.312	0.199	0.065
0.1	5	1	TH	0.238	0.073	0.34	0.313	0.036
			EBKF	0.257	0.076	0.33	0.294	0.043
0.1	5	3	TH	0.247	0.071	0.337	0.304	0.041
			EBKF	0.257	0.076	0.33	0.294	0.043
0.1	5	6	TH	0.259	0.073	0.333	0.292	0.043
			EBKF	0.257	0.076	0.33	0.294	0.043
0.3	0.05	1	TH	0.519	0.032	0.262	0.038	0.149
			EBKF	0.515	0.031	0.205	0.042	0.207
0.3	0.05	3	TH	0.389	0.062	0.327	0.168	0.054
			EBKF	0.515	0.031	0.205	0.042	0.207
0.3	0.05	6	TH	0.325	0.072	0.339	0.232	0.032
			EBKF	0.515	0.031	0.205	0.042	0.207
0.3	1	1	TH	0.293	0.076	0.33	0.258	0.043
			EBKF	0.359	0.071	0.308	0.192	0.07
0.3	1	3	TH	0.246	0.08	0.337	0.305	0.032
			EBKF	0.359	0.071	0.308	0.192	0.07
0.3	1	6	TH	0.347	0.064	0.314	0.204	0.071
			EBKF	0.359	0.071	0.308	0.192	0.07
0.3	5	1	TH	0.207	0.086	0.339	0.352	0.016
			EBKF	0.236	0.086	0.331	0.323	0.024
0.3	5	3	TH	0.218	0.081	0.34	0.341	0.02
			EBKF	0.236	0.086	0.331	0.323	0.024
0.3	5	6	TH	0.243	0.082	0.336	0.316	0.023
			EBKF	0.236	0.086	0.331	0.323	0.024
Overall			TH	0.332	0.064	0.311	0.220	0.073
			EBKF	0.373	0.057	0.277	0.178	0.115
			TH	0.396		0.604		
			EBKF	0.430		0.570		

3.4.1.2 Computational Efficiency

In the previous subsection, through simulations we have shown that one advantage of the EB-based methods over TH method lies in requiring less knowledge of process parameters to achieve good estimation accuracy of the process mean. In the following, we will demonstrate the advantage of our proposed methods from the computational perspective.

The following Figure 3-2 displays the computational costs of different methods on running 100 simulations ($S = 100$) of different data length N. Each data sample is generated from parameters $\sigma^2 = 1, p = 0.1, \sigma_{RW}^2 = 0.05, \delta = \sigma_{RW}$.

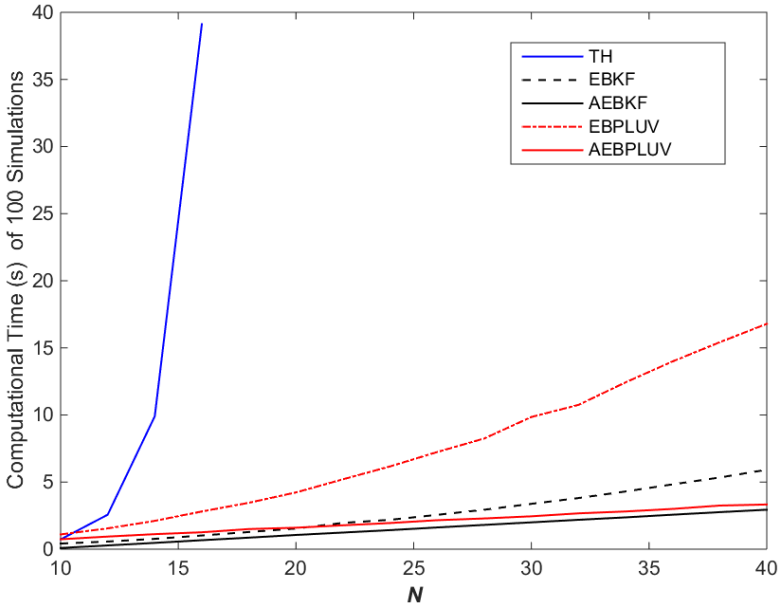


Figure 3-2. Computational costs of different methods of 100 simulations with varying data size N. Each signal is generated using $\sigma^2 = 1, p = 0.1, \sigma_{RW}^2 = 0.05, \delta = \sigma_{RW}$.

Results of Figure 3-2 are consistent with our previous discussions on the computational complexity. The TH method is the least efficient with exponentially increasing computational time. Both EB-based methods (EBPLUV and EBKF) exhibit quadratically increasing cost over time, with EBKF being superior because of using more

prior knowledge (variance of the random walk) and having smaller number of unknown parameters. The most efficient methods refer to the approximated EB-based methods (AEBPLUV and AEBKF) whose computational costs are linear over time. Here we do not show the corresponding results of EBPLUV-EV and AEBPLUV-EV since they have the same computation costs as EBPLUV and AEBPLUV, respectively.

3.4.2 Piecewise Linear Data

It is found in the literature that some degradation data can be modeled with a piecewise linear model (Chen and Tsui, 2013). In this subsection, we further evaluate and compare the performance of our methods and TH method using piecewise linear data, which are generated based on the following model:

$$\theta_t \sim \begin{cases} k_1 t, & 0 \le t \le 5 \\ k_2(t-5)+\theta_5, & t > 5 \end{cases}, k_2 = \Delta \cdot k_1 \ (\Delta > 1)$$

$$y_t = \theta_t + \varepsilon_t, \varepsilon_t \sim N(0, \sigma^2).$$

The process mean of the simulated data consists of two linear segments with different slopes k_1 and k_2, and $k_2 > k_1$. Since both EBPLUV and EBPLUV-EV methods are based on our piecewise linear model in Eq. (3-3), we will use them to compare with the TH method. The detection performance of different methods are still evaluated based on how 'close' the estimated failure time n_2 is to the true failure time n_1. We are going to use the following metrics for performance evaluation:

FAR = $\frac{\sum_{i=1}^{S} I\{n_{1i} - n_{2i} > 0\}}{S}$, MAR = $\frac{\sum_{i=1}^{S} I\{(n_{1i} > 0) \cap (n_{2i} = 0)\}}{S}$, and

DD-RMSE = $\sqrt{\frac{\sum_{\{i: n_{2i} \ge n_{1i} > 0, 1 \le i \le S\}} (n_{2i} - n_{1i})^2}{S}}$

where n_{1i} and n_{2i} are true failure time and estimated failure time of the i-th data sample, respectively, $n_{2i} = 0$ if no failure is detected by the end of the data sample, and

$I\{A\}$ is the indicator function with value 1 if A is true and 0 otherwise. FAR is the *false alarm rate* which measures the percentage of early detection while MAR is the *missed alarm rate* which is the rate that the algorithm does not give an alarm when the true failure occurs. The root mean squared error (RMSE) for delayed detections (DD-RMSE), which is defined on the closeness of n_1 and n_2 when $n_2 \geq n_1$, measures the detection delay when the failure is detected. They are all very common metrics for evaluating the performance of a statistical process control procedure.

In our simulations, σ^2 is set to be 1 and 10 and constant over time, $k_1 = 0.5$, and $\Delta = (4, 5, 6)$, which provides six different parameter scenarios. For each of the parameter scenario, $S = 1000$ samples with data length $N = 15$ are simulated. The critical threshold M is defined as $M = (\theta_{t=10} + \theta_{t=11})/2$, thus the true failure occurs at time 11 ($n_{1i} = 11, i = 1, 2, \dots, S$). For the inference parameters in TH method, we use $\sigma_{RW}^2 = q\sigma, q = (0, 1), \delta = m\sigma, m = (0.05, 1, 5)$ and only show the results for TH method using the parameter values leading to the best performance. Both our methods and TH method use $p = 0.1$.

Table 3-3 shows the corresponding results. Results using approximated versions of EBPLUV and EBPLUV-EV, namely AEBPLUV and AEBPLUV-EV, are also included. It can be seen that the TH method fails to detect the failure for each of the data sample (MAR=1). Comparing EBPLUV and EBPLUV-EV, EBPLUV-EV has significantly better detection performance when the process variance is small ($\sigma^2 = 1$), which is consistent with our observations in Section 3.2.4. When variance is large, the two methods have very similar performance. Note that compared to EBPLUV and EBPLUV-EV, their approximated versions yield very similar results with much less

computational load. Therefore in the rest of this chapter, we will use AEBPLUV and AEBPLUV-EV instead of their exact versions.

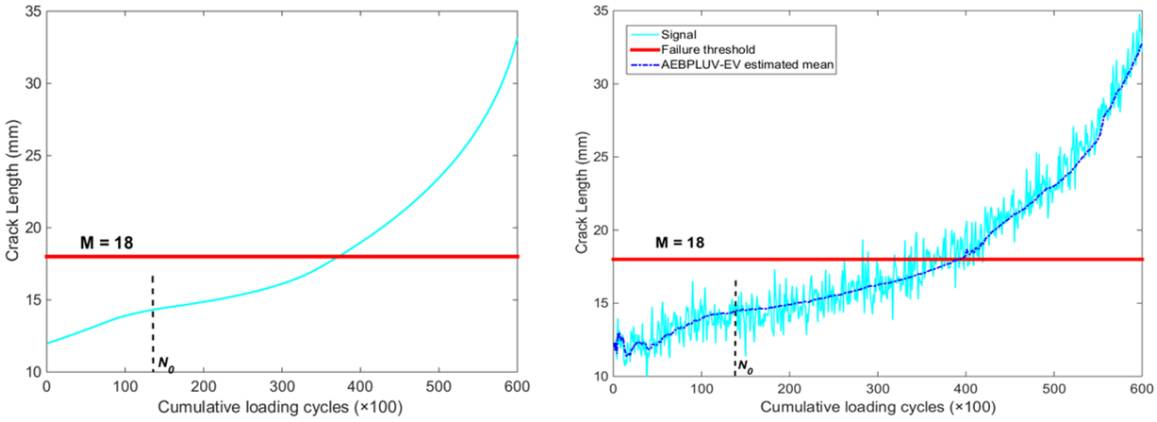
Table 3-3. Simulation results of different methods using data with two piecewise linear segments

True parameters		Inference parameters		Methods	**FAR**	**MAR**	**DD-RMSE**
σ^2	Δ	m	q				
1	4	0.05	0	TH	0	1	N/A
				EBPLUV	0	0	0.94
				AEBPLUV	0.01	0	0.92
				EBPLUV-EV	0.02	0	0.75
				AEBPLUV-EV	0.01	0	0.74
1	5	0.05	0	TH	0	1	N/A
				EBPLUV	0.01	0	0.87
				AEBPLUV	0.01	0	0.88
				EBPLUV-EV	0.02	0	0.57
				AEBPLUV-EV	0.02	0	0.58
1	6	0.05	0	TH	0	1	N/A
				EBPLUV	0	0	0.71
				AEBPLUV	0	0	0.70
				EBPLUV-EV	0.01	0	0.43
				AEBPLUV-EV	0.01	0	0.41
10	4	0.05	0	TH	0	1	N/A
				EBPLUV	0.13	0	1.30
				AEBPLUV	0.14	0	1.31
				EBPLUV-EV	0.14	0	1.28
				AEBPLUV-EV	0.14	0	1.29
10	5	0.05	0	TH	0	1	N/A
				EBPLUV	0.11	0	1.15
				AEBPLUV	0.12	0	1.15
				EBPLUV-EV	0.12	0	1.15
				AEBPLUV-EV	0.11	0	1.14
10	6	0.05	0	TH	0	1	N/A
				EBPLUV	0.09	0	1.03
				AEBPLUV	0.09	0	1.02
				EBPLUV-EV	0.09	0	1.02
				AEBPLUV-EV	0.08	0	1.02

3.4.3 Degradation Data for Fatigue Crack Propagation

In this subsection, we apply our proposed methods to a real type of degradation data called the *fatigue crack propagation*. Fatigue failure has been commonly seen in

practice, including some famous examples such as the collapse of the Silver Bridge (Lee, 1996), the fractures of the Liberty ships (Macdonald, 2011), etc. The fatigue failure is usually caused by the *crack propagation*, whose path can be considered as a degradation path. The crack propagation is subject to repeatedly fatigue loading, which is either cyclic with a constant amplitude or non-cyclic with amplitudes exceeding a threshold. The latter one is also called *overloads*. Si et al. (2016) studied the problem of *overload retardation*, which occurs in some materials where overloads decelerate rather than accelerate the crack propagation process. The analysis of the overload retardation effect is considered critical since ignoring such effect can lead to a biased estimation of product life.



(a). A crack length-cumulative loading cycles curve with no random noise.

(b). A crack length-cumulative loading cycles curve with noise variance $\sigma^2 = 1$ and its fitted trend using AEBPLUV-EV.

Figure 3-3. Illustrations of the crack length-cumulative loading cycles curve with and without noise.

Figure 3-3 shows a crack propagation path with an overload retardation effect, which is based on the experiment conducted in Si et al. (2016). In the experiment, the Compact Tension (CT) fatigue loads (Sih, 1973) mixed with a single load are added at

cycle N_0 on some testing specimens to obtain their crack propagation paths. Given the observed crack propagation paths, Si et al. (2016) proposed a physical-statistical model that is based on a modified Paris law model and a maximum likelihood estimation (MLE) approach to obtain the fitted crack propagation paths with no random errors, which will be further used to determine whether the testing material has an overload retardation effect. The crack propagation path in Figure 3-3(a) is one of the fitted crack propagation paths with no random noise that is obtained in Si et al. (2016).

We test our AEBPLUV and AEBPLUV-EV methods on this fitted crack propagation curve by adding simulated noises with three different variance levels: $\sigma^2 = (1, 3, 6)$. The failure threshold is $M = 18$ as defined in Si et al. (2016). As an illustration, Figure 3-3(b) shows the estimated mean trend using our AEBPLUV-EV method for the crack length curve in Figure 3-3(a) with noise variance $\sigma^2 = 1$. It can be seen that the estimated mean in Figure 3-3(b) using AEBPLUV-EV captures the main trends in the true mean curve in Figure 3-3(a), including the overload retardation phase starting at N_0. TH method is not included for comparison because of its exponentially increasing computational cost over time, which is obviously prohibitive for this data with 600 time points. The performance of the methods are again evaluated based on the metrics defined previously, namely FAR and DD-RMSE (MAR is not needed here since both methods have zero missed alarms). As can be seen in Table 3-4, both methods have the same performance on FAR. AEBPLUV-EV slightly outperforms AEBPLUV in terms of smaller detection delays.

Table 3-4. Simulation results of different methods using the fitted crack propagation path with different noise levels

σ^2	Methods	**FAR**	**DD-RMSE**
1	AEBPLUV	0.02	25.31
	AEBPLUV-EV	0.02	23.76
3	AEBPLUV	0	34.62
	AEBPLUV-EV	0	34.20
6	AEBPLUV	0.02	35.43
	AEBPLUV-EV	0.02	34.88

3.5 Summary

In this chapter we study the problem of detecting on-line whether the mean of a short-run process subject to random jumps and possibly random walks has exceeded a pre-specified threshold. We modify the model proposed by Tsiamyrtzis and Hawkins (2005) by assuming the process mean before and after occasional jumps are independent, which forms the basis of using the exact Bayesian method and its approximated version to carry out the posterior inference of the process mean. Our proposed model is also more flexible in that it can be extended for handling linear trends, which is also commonly seen in various short-run and degradation processes.

The contributions of our proposed method are three-folds. First and most important is its high computational efficiency. By utilizing the exact Bayesian method, the computational cost of updating the posterior distribution of the process mean is reduced to be quadratic over time, compared with exponential cost of the TH method. And using the approximated exact Bayesian with optimal resampling, we can further reduce the computational complexity to be linear over time, with little loss in estimation accuracy. Secondly, compared with the TH method, the proposed method is shown to require less knowledge on the process parameters such as the jump size and the

measurement variance. So it is much more robust to parameter misspecifications. Thirdly, for processes with a constant (but unknown) variance, we propose a new approximate method to utilize the variance information accumulated in previous segments while not increasing the computational costs.

The performance of our method is evaluated through both simulated signals and real data examples. Results demonstrate the robustness of our proposed methods in estimation performance when using different parameter specifications (both correct and incorrect) for inference, and significantly higher computational efficiency compared to the existing methods.

CHAPTER 4. WEIGHTED BAND DEPTH APPROACHES FOR FUNCTIONAL DATA ANALYSIS

Recently, functional data analysis has been a popular research topic in many areas. Among different techniques, the use of the concept of *depth* has shown its great usefulness in robust analysis of functional data, such as producing robust location estimators of sample curves, or detecting various outliers. In this chapter, we focus on using depth for detecting outliers, which are considered as coming from a different process compared to normal curves. We propose two new depth notions, the weighted band depth (WBD) and the localized weighted band depth (LWBD), for detecting various outliers based on the *band depth* proposed by López-Pintado and Romo (2009). A novel idea called the *shape distance* is introduced to make proposed depth approaches much more effective than the band depth method in detecting outliers that have different shapes with normal curves. Additionally, the proposed LWBD is also robust in detecting outliers with exceptionally large magnitude. Numerical studies based on simulated signals and application to a real data example are used to evaluate the performance of proposed depth methods and other existing depth notions. Our proposed depth approaches are shown to be much more effective and robust in detecting a variety of outliers.

The remainder of this chapter is organized as follows. In Section 4.1, we give a brief review on the original band depth. The proposed WBD and LWBD are presented in Section 4.2. Simulation study and application to a real example are contained in Sections 4.3 and 4.4, respectively. This chapter is concluded in Section 4.5.

4.1 Review of the Band Depth for Functional Data

In this section, we give a brief review on the BD, a depth approach recently proposed by López-Pintado and Romo (2009) that is shown to be more effective than existing functional depths in detecting shape outliers.

Let $f(t)$ be a stochastic function, $t \in \mathcal{T}$, belonging to the space $C(\mathcal{T})$ of real continuous functions on some compact interval $\mathcal{T} \in \mathbb{R}$. The band depth is defined through a graph-based approach. Suppose the graph of a function f is the subset of the plane $G(f) = \{(t,\ f(t)): t \in \mathcal{T}\}$, and $f_1(t), f_2(t), \ldots, f_n(t)$ are a collection of corresponding real functions. The band in $\mathbb{R}^2$ delimited by j $(2 \le j \le n)$ curves $f_{i_1}, f_{i_2}, \ldots, f_{i_j}$ is $B\left(f_{i_1}, f_{i_2}, \ldots, f_{i_j}\right) = \left\{(t, f(t)): t \in \mathcal{T}, \min_{r=1,\ldots,j} f_{i_r}(t) \le f(t) \le \max_{r=1,\ldots,j} f_{i_r}(t)\right\}$. The following Figure 4-1 displays the band given by three solid curves; the dotted curve is included in this band, while the dashed curve is not completely inside the band. For any function f in $f_1, \ldots, f_n$, its BD is defined as:

$$BD_{n,J}(f) = \sum_{j=2}^{J} BD_n^{(j)}(f) \tag{4-1}$$

$$BD_n^{(j)}(f) = \binom{n}{j}^{-1} \sum_{1 \le i_1 \ldots < i_j \le n} I\left\{G(f) \subseteq B\left(f_{i_1}, f_{i_2}, \ldots, f_{i_j}\right)\right\}, 2 \le j \le n \tag{4-2}$$

where $I\{A\}$ is the indicator function with value 1 if A is true and 0 otherwise. Therefore, the BD of the function f can be expressed as the proportion that f is completely contained in the bands $B\left(f_{i_1}, f_{i_2}, \ldots, f_{i_j}\right)$ delimited by j different curves $f_{i_1}, f_{i_2}, \ldots, f_{i_j}$. In López-Pintado and Romo (2009), $J = 3$ is recommended for use.

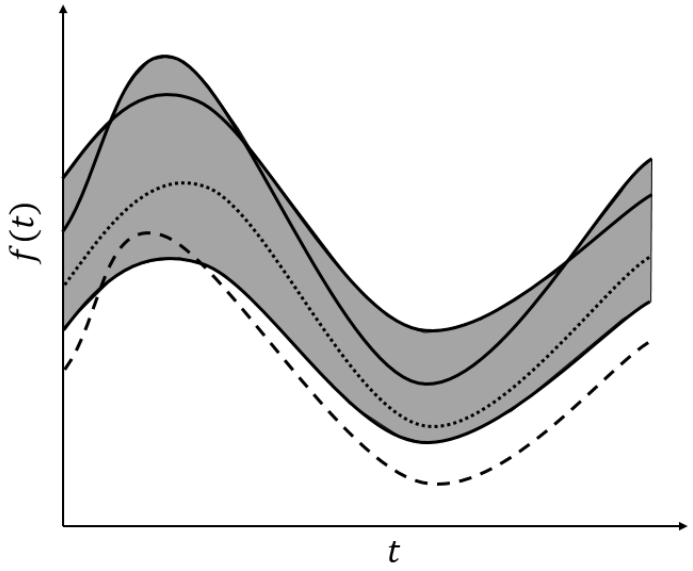


Figure 4-1. Band delimited by three solid curves.

The finite dimensional version of BD can be also constructed for the use of multivariate data. For any point **f** in $\mathbb{R}^d$**,** it can be regarded as a real function defined on the set of indexes $\{1, 2, \dots, d\}$ and expressed as $\mathbf{f} = \big(f(1), f(2), \dots, f(d)\big)$. Given the points $\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \dots, \mathbf{f}_{i_j}$ in $\mathbb{R}^d$, the corresponding band in parallel coordinates $B\left(\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \dots, \mathbf{f}_{i_j}\right) = \left\{(k, z): k \in \{1, 2, \dots, d\}, \min_{r=1,\dots,j} f_{i_r}(k) \le z \le \max_{r=1,\dots,j} f_{i_r}(k)\right\}$ becomes in Cartesian coordinates a d-dimensional interval with sides parallel to the axes delimited by the minimum and maximum of the coordinates of $\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \dots, \mathbf{f}_{i_j}$, expressed as:

$$Q\left(\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \dots, \mathbf{f}_{i_j}\right) = \left\{\mathbf{f} \in \mathbb{R}^d, \min_{r=1,\dots,j} f_{i_r}(k) \le f(k) \le \max_{r=1,\dots,j} f_{i_r}(k), k = 1, 2, \dots, d\right\} \tag{4-3}$$

The following Figure 4-2 displays the band delimited by three 2-dimensional points (1, 3), (4, 6) and (5, 4) in the plane in parallel and Cartesian coordinates, respectively; the point (2, 5) is inside the band, while the point (3, 7) is not included in

the band. Therefore, for any point $\mathbf{f}$ in $\mathbf{f}_1, \ldots, \mathbf{f}_n$, $BD_n^{(j)}(\mathbf{f})$ is the proportion of sets $Q\left(\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \ldots, \mathbf{f}_{i_j}\right)$ defined by j different points $\mathbf{f}_{i_1}, \mathbf{f}_{i_2}, \ldots, \mathbf{f}_{i_j}$ from the sample containing $\mathbf{f}$. The band depth of $\mathbf{f}$ can then be obtained through Eq. (4-1).

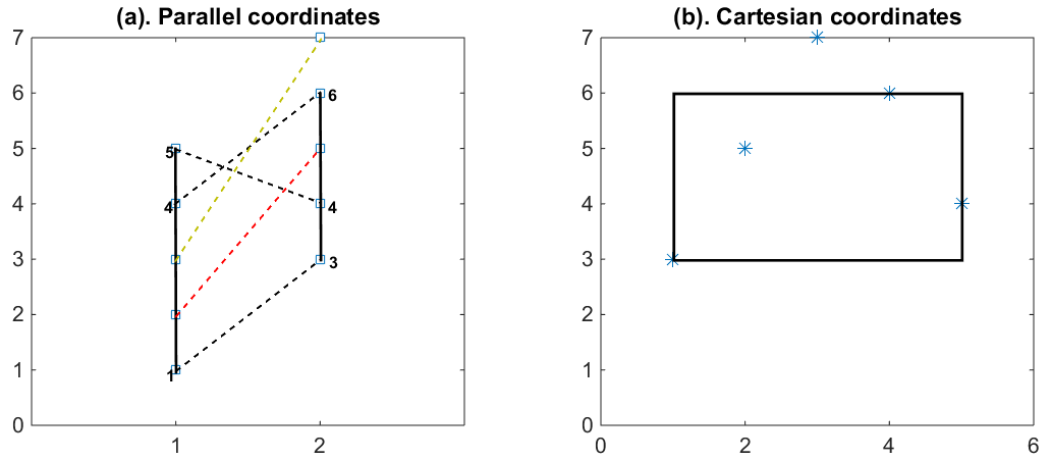


Figure 4-2. (a) The band (solid lines) delimited by points (1, 3), (4, 6), (5, 4) in parallel coordinates and (b) the same three points and interval (the rectangle) in Cartesian coordinates. Point (2, 5) is inside the band, point (3, 7) is outside the band.

By using the idea of band, the BD is able to capture useful shape information of each curve and as a result, it is more effective than existing functional depths in detecting shape outliers. However, as discussed previously, the major disadvantage of BD is that it is only sensitive for those shape outliers that are not fully contained in the band. For those shape outliers that are completely inside the band, BD is lack of strength in detecting them since it would treat these outliers exactly the same as those normal curves that are also fully inside the band. To overcome such drawback, in the following section we propose the weighted band depth (WBD) based on the idea of using *shape distance*.

4.2 The Weighted Band Depth (WBD) and Localized Weighted Band Depth (LWBD)

In this section, we first propose WBD based on the idea of using shape distance, which leads to greater strength of tackling different shape outliers. To make it also robust in detecting location outliers, we further introduce the localized WBD (LWBD). The finite dimensional version of the WBD and LWBD are also explored.

4.2.1 The Shape Distance in WBD and its Motivation

Our idea of using the shape distance is motivated by the example in Figure 1-1, where we would like to detect those shape outliers that are completely inside a band. When curves are fully contained in a band, rather than being treated all equally, their depths would be varied by their shapes. Specifically, the shape distance of each target curve to its enclosed band is obtained by 'squeezing' the band until it has minimum width to fully contain the target curve. The resulted band width can be considered as the shape distance of the target curve to the band. Figure 4-3 shows three target curves (f) with different shapes that are all fully inside a band delimited by two other curves f_1, f_2. As can be seen, as the shape of the target curve becomes more similar to the boundary curves of the band, the corresponding shape distance becomes smaller. Consequently, when the target curve is exactly the same as the boundary curves of the band, their shape distance becomes zero, as illustrated by the case (c).

Mathematically, suppose the target functional curve, f, is fully contained in a band delimited by two other curves f_1, f_2. Without loss of generality, assuming f_2 is above f, then the shape distance (SD) of f to the band is defined as:

$$SD\big(f, B(f_1, f_2)\big) = d(f_1, f_2) - \Big(\min_{t\in\mathcal{T}}\big(f_2(t) - f(t)\big) + \min_{t\in\mathcal{T}}\big(f(t) - f_1(t)\big)\Big) \tag{4-4}$$

The terms inside the parenthesis of Eq. (4-4) correspond to the 'squeezing' distance of the band. $d(f_1, f_2)$ corresponds to the distance measurement between curves f_1 and f_2. In this chapter, it is chosen to be the maximum norm (or infinity norm), which is $d(f_1, f_2) = \|f_1 - f_2\|_\infty = \max_{t\in\mathcal{T}}|f_1(t) - f_2(t)|$. Other distance measurements, like L_1-norm (absolute distance), L_2-norm (Euclidean distance), can be also used following the same approach.

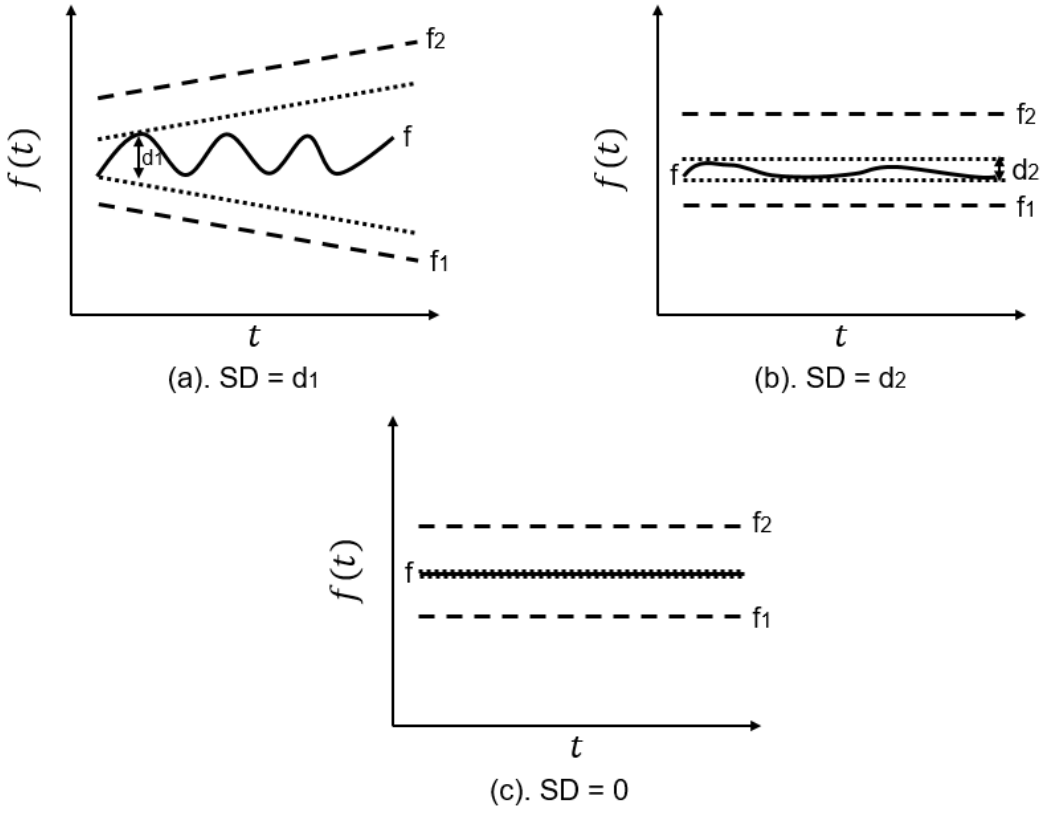
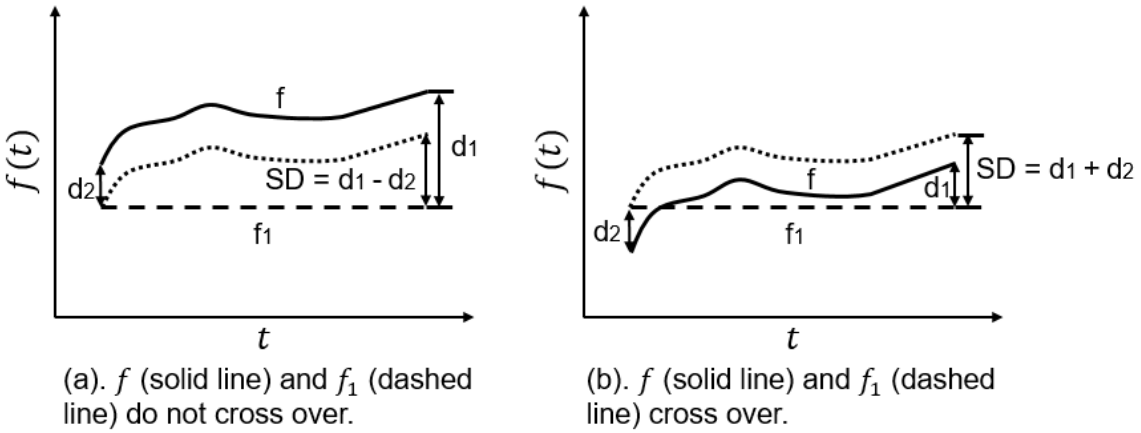


Figure 4-3. Illustration of the shape distance of the target curve f (solid line) to its enclosed band delimited by two dashed curves.

Our previous discussion focuses on the scenario where the target curve f is not the curve delimiting the band ($f \notin \{f_1, f_2\}$). When the target curve is just one of the curves that delimits the band ($f \in \{f_1, f_2\}$), the shape distance of f to the band can be also obtained based on Eq. (4-4). Specifically, suppose the band is determined by two functional curves – the target curve f and another curve f_1. By considering f overlaps f_2 and replacing f_2 with f in Eq. (4-4), in this case the shape distance of f to the band becomes:

$$SD\big(f, B(f, f_1)\big) = \max_{t \in \mathcal{T}} \big(f(t) - f_1(t)\big) - \min_{t \in \mathcal{T}} \big(f(t) - f_1(t)\big) \tag{4-5}$$

Eq. (4-5) suggests that, if f and f_1 do not cross over, the shape distance of f to the band simply becomes the difference between maximal and minimal distances of f and f_1, as illustrated in Figure 4-4(a). On the other hand, if f and f_1 cross with each other, the resulted shape distance is equivalent to the summation of two maximal distances, each corresponds to f being above f_1 and f being below f_1, which is illustrated in Figure 4-4(b).



(a). f (solid line) and f_1 (dashed line) do not cross over.

(b). f (solid line) and f_1 (dashed line) cross over.

Figure 4-4. Illustration of the shape distance of the target curve f to the band when it is one of the boundary curves of the band.

Based on the shape distance of the target curve f to different bands, the depth of f can be obtained. The shape distance essentially quantifies the shape similarity between the target curve and remaining curves in the data. Since smaller shape distance represents higher similarity of the target curve with the data, a larger depth should be assigned. Mathematically, for any functional curve f in $f_1, \dots, f_n$, the WBD of f when $J = 2$ is defined as:

$$WBD_n(f) = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} W\left(SD\left(f, B(f_{i_1}, f_{i_2})\right)\right) I\{G(f) \subseteq B(f_{i_1}, f_{i_2})\} \quad (4\text{-}6)$$

where $W\left(SD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$ is a non-increasing function of $SD\left(f, B(f_{i_1}, f_{i_2})\right)$, which can be obtained following Eq. (4-4) $\left(f \notin \{f_{i_1}, f_{i_2}\}\right)$ or Eq. (4-5) $\left(f \in \{f_{i_1}, f_{i_2}\}\right)$. Compared to the original band depth (Eq. (4-1) and Eq. (4-2)), the term $W\left(SD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$ is added in WBD to take into consideration different contributions of curves with different shapes. The reason for using $J = 2$ in WBD is that it involves reasonable computation and can already detect shape outliers very effectively, as will be shown in our simulation study.

In this chapter, the non-increasing function $W\left(SD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$ is chosen to be the truncated Gaussian kernel function, which is also used in HMD (Curvas et al., 2006) and KFSD (Sguera et al., 2014):

$$W\left(SD\left(f, B(f_{i_1}, f_{i_2})\right)\right) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{SD^2\left(f, B(f_{i_1}, f_{i_2})\right)}{2c_1^2}\right), SD\left(f, B(f_{i_1}, f_{i_2})\right) \ge 0, c_1 > 0 \quad (4\text{-}7)$$

where c_1 is the tuning parameter controlling how fast $W\left(SD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$ dilutes with the increase of the shape distance. Note that other types of non-increasing functions, such as the reciprocal function, can be also used for $W\left(SD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$.

Similar to the original band depth, the finite dimensional version of WBD can be also constructed for multivariate data. For any point $\mathbf{f}$ in $\mathbf{f}_1, \ldots, \mathbf{f}_n$, the WBD of $\mathbf{f}$ can be calculated as:

$$WBD_n(\mathbf{f}) = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} W\left(SD\left(\mathbf{f}, Q(\mathbf{f}_{i_1}, \mathbf{f}_{i_2})\right)\right) I\{\mathbf{f} \subseteq Q(\mathbf{f}_{i_1}, \mathbf{f}_{i_2})\} \quad (4\text{-}8)$$

where $Q(\mathbf{f}_{i_1}, \mathbf{f}_{i_2})$ is defined in Eq. (4-3) that represents the band delimited by points $\mathbf{f}_{i_1}, \mathbf{f}_{i_2}$ in Cartesian coordinates, $W\left(SD\left(\mathbf{f}, Q(\mathbf{f}_{i_1}, \mathbf{f}_{i_2})\right)\right)$ is the same non-increasing function defined in Eq. (4-7), and $SD\left(\mathbf{f}, Q(\mathbf{f}_{i_1}, \mathbf{f}_{i_2})\right)$ can be obtained similarly according to Eq. (4-4) or Eq. (4-5).

4.2.2 The Localized WBD

Although WBD can be highly effective in capturing the shape information by using the shape distance, it is specially designed for detecting shape outliers while ignoring exact location of each curve. Hence it may not be robust in detecting location outliers. As an illustration, considering the example in Figure 4-5, whose curves are obtained by first generating two groups of 'normal' data from a given process with the same distribution, respectively. Then an isolated curve who has the same shape with normal curves while being far away from both groups of normal curves is added. It is obvious that the isolated curve should be considered as the outlier (more precisely, the location outlier), whose depth should be the lowest. However, both BD and WBD treat

the isolated curve as the deepest curve. The reason is that, both BD and WBD focus on capturing the shape information of each curve, while do not pay attention to the exact location of each curve. Since the isolated curve is roughly in the geometric center of the data set and it is the curve that is contained in the largest number of bands, both BD and WBD assign it with the largest depth value, regardless of whether such curve being isolated from the rest of the normal data.

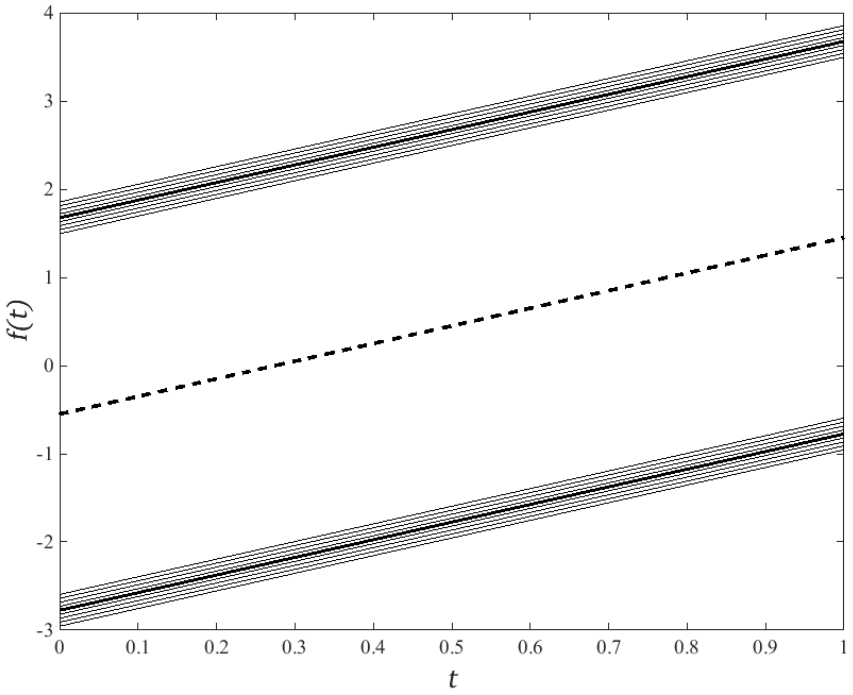


Figure 4-5. Illustration of detecting location outliers. The dashed curve is the outlier curve, while two bolded curves (in the geometric center of their corresponding groups of normal curves) are curves with largest depths obtained by LWBD (both c_1 and c_2 are chosen to be 50-th percentile of empirical distribution of $\left\{d\left(f_{i_1}, f_{i_2}\right) = \max_{t\in\mathcal{T}}\left|f_{i_1}(t) - f_{i_2}(t)\right|, 1 \le i_1 < i_2 \le n\right\}$).

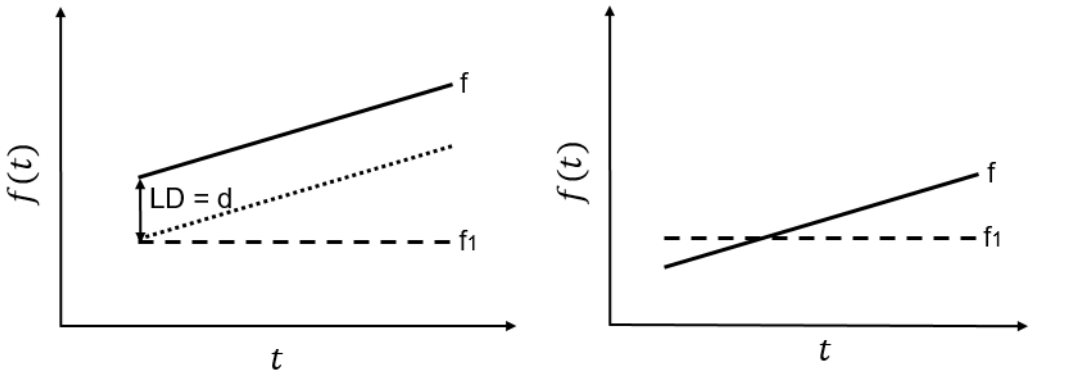
To make WBD also robust in detecting location outliers, we further propose the localized version of WBD (LWBD). The main idea in LWBD is that distance among

curves is decomposed into two parts – shape distance and location distance. Besides the shape distance, the location distance is used for capturing the relative location information between curves. And contributions of curves to the depth of target curve f decrease as their location distances to f increase, which essentially follows the localization idea in local-oriented functional depth approaches.

Specifically, considering a band determined by two functional curves – the target curve f and another curve f_1 (In LWBD, we still consider $J = 2$). The location distance (LD) of f to the band $B(f, f_1)$ is defined as:

$$LD\big(f, B(f, f_1)\big) = \begin{cases} \min\limits_{t \in \mathcal{T}} \big(f(t) - f_1(t)\big), & \text{if } f \text{ and } f_1 \text{ do not cross over} \\ 0, & \text{if } f \text{ and } f_1 \text{ cross over} \end{cases} \tag{4-9}$$

Eq. (4-9) suggests that, when f is one of the boundary curves of the band, its location distance to the band is the smallest distance when vertically moving f towards f_1 until they start crossing over for the first time. Therefore, when two curves already cross over, the corresponding location distance is simply just zero. The examples in Figure 4-6 illustrate the calculations of LD following our definition.



(a). f (solid line) and f_1 (dashed line) do not cross over, LD = d.

(b). f (solid line) and f_1 (dashed line) cross over, LD = 0.

Figure 4-6. Illustration of the location distance of the target curve f to the band when it is one of the boundary curves of the band.

On the other hand, when f is fully inside the band $B(f_1, f_2)$ delimited by two other curves f_1 and f_2; that is, $f \notin \{f_1, f_2\}$. Without loss of generality, assuming f_2 is above f, the location distance of f to the band $B(f_1, f_2)$ is defined as:

$$LD\big(f, B(f_1, f_2)\big) = \min_{t\in\mathcal{T}}\big(f_2(t) - f(t)\big) + \min_{t\in\mathcal{T}}\big(f(t) - f_1(t)\big) \tag{4-10}$$

Note that this is just the 'squeezing' distance in Eq. (4-4).

Therefore, for any functional curve f in $f_1, \dots, f_n$, the LWBD of f (when $J = 2$) is defined as

$$\begin{aligned} &LWBD_n(f) \\ &= \binom{n}{2}^{-1} \sum_{1\le i_1<i_2\le n} W\left(LD\left(f, B(f_{i_1}, f_{i_2})\right)\right) W\left(SD\left(f, B(f_{i_1}, f_{i_2})\right)\right) I\{G(f) \subseteq B(f_{i_1}, f_{i_2})\} \end{aligned} \tag{4-11}$$

where $W\left(LD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$ is a non-increasing function of $LD\left(f, B(f_{i_1}, f_{i_2})\right)$, which can be obtained following Eq. (4-9) $\left(f \in \{f_{i_1}, f_{i_2}\}\right)$ or Eq. (4-10) $\left(f \notin \{f_{i_1}, f_{i_2}\}\right)$. Compared to WBD, the localization term $W\left(LD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$ is added in LWBD to account for the effect of relative location distance between curves. In this chapter, the non-increasing function $W\left(LD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$ is chosen to follow the same form of Eq. (4-7):

$$\begin{aligned} &W\left(LD\left(f, B(f_{i_1}, f_{i_2})\right)\right) \\ &= \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{LD^2\left(f, B(f_{i_1}, f_{i_2})\right)}{2c_2^2}\right), LD\left(f, B(f_{i_1}, f_{i_2})\right) \ge 0, c_2 > 0 \end{aligned} \tag{4-12}$$

where c_2 is another tuning parameter controlling how fast $W\left(LD\left(f, B(f_{i_1}, f_{i_2})\right)\right)$ dilutes with the increase of the location distance.

The finite dimensional version of LWBD for multivariate data can be obtained similarly by adding the corresponding localization term to Eq. (4-8); that is, for any point $\mathbf{f}$ in $\mathbf{f}_1, \ldots, \mathbf{f}_n$, the LWBD of $\mathbf{f}$ can be calculated as:

$$LWBD_n(\mathbf{f}) = \binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le n} W\left(LD\left(\mathbf{f}, Q(\mathbf{f}_{i_1}, \mathbf{f}_{i_2})\right)\right) W\left(SD\left(\mathbf{f}, Q(\mathbf{f}_{i_1}, \mathbf{f}_{i_2})\right)\right) I\{\mathbf{f} \subseteq Q(\mathbf{f}_{i_1}, \mathbf{f}_{i_2})\} \quad (4\text{-}13)$$

Back to the illustrated example at the beginning of this section, by taking into account the relative location effect between curves, LWBD assigns the isolated curve with the lowest depth and thus successfully identifies the isolated curve as an outlier. Curves with largest depths obtained by LWBD are those located in the geometric center of two corresponding groups of normal curves, respectively, as shown in Figure 4-5. In fact, since LWBD considers both location and shape information of each curve, it can be robust to both location and shape outliers. In the following section, we would conduct simulation studies to demonstrate the usefulness of both WBD and LWBD.

4.3 Simulation Study

In this section, simulations are carried out to study the performance of proposed weighted band depth approaches in detecting a variety of outliers. We first introduce different models upon which the simulated data are generated. A metric is then developed to evaluate the detection performance, which is followed by detailed simulation results showing the detection performance of each depth approach.

4.3.1 Models for Generating Simulated Data

For each of simulated data set, it consists of n sample curves $f_1, \dots, f_n$ that are generated from a mixture F_{mix} of two stochastic processes, one for normal curves F_{nor} and one for outliers F_{out}. That is,

$$F_{mix} = \begin{cases} F_{nor}, & \text{with probability } 1-\alpha \\ F_{out}, & \text{with probability } \alpha \end{cases} \tag{4-14}$$

where α is the outlier probability (the probability that a curve being an outlier equals to α). All simulated curves are unlabeled, and our goal is to determine for each curve whether it is an outlier or not.

In our simulation study, we consider both shape outliers and location outliers. The shape outliers we test are irregular curves in a set of smooth curves (normal curves). When the irregularity of the shape outlier is relatively weak, they would have higher chance to be enclosed by other bands, and hence being more difficult to be detected. Note that this is exactly our motivation of proposing WBD. In addition, we also test different types of location outliers to study the robustness of our localized WBD in detecting them compared with other popular local-oriented functional depth approaches (HMD and KFSD).

For shape outliers, we consider six models based on the mixture of two stochastic processes in Eq. (4-14). The first three mixture models are SHMM1, SHMM2 and SHMM3, which share the same F_{nor}, with curves generated by:

$$f_{nor}(t) = 4t + \varepsilon(t), t \in [0,1], \tag{4-15}$$

where $\varepsilon(t)$ is a zero mean Gaussian process with covariance function: $E\big(\varepsilon(s), \varepsilon(t)\big) = 0.25\exp(-(s-t)^2)$.

SHMM1, SHMM2 and SHMM3 differ in their F_{out} components. Their corresponding outlier curves are generated by adding to Eq. (4-15) different levels of noise:

$$f_{out}(t) = f_{nor}(t) + \varepsilon'(t), \varepsilon'(t) \sim N(0, \sigma^2),$$

which results in various levels of irregularity of outlier curves. Similar models are also considered in Sguera et al. (2016). Note that smaller noise variance levels may lead to outliers having higher chance to be enclosed in other bands, making them much more difficult to be detected. In our simulation, we choose $\sigma = 0.2, 0.25, 0.3$ for SHMM1, SHMM2 and SHMM3, respectively.

Another three mixture models for generating shape outliers are SHMM4, SHMM5 and SHMM6, which also share the same F_{nor}, with curves generated by:

$$f_{nor}(t) = 30(1-t)t^2 + \varepsilon(t), t \in [0, 1] \quad (4\text{-}16)$$

The outliers in SHMM4, SHMM5 and SHMM6 are generated by adding to Eq. (4-16) different levels of irregular functions:

$$f_{out}(t) = f_{nor}(t) + m|sin(20\pi t)|, t \in [0, 1],$$

with $m = 0.8, 1, 1.2$ for SHMM4, SHMM5 and SHMM6, respectively. Similarly, when values of m are relatively small, the resulted outliers would have more chance to be fully inside other bands. Figure 4-7 displays the simulated data set with at least one outlier for each of the mixture model in SHMM1 – 6.

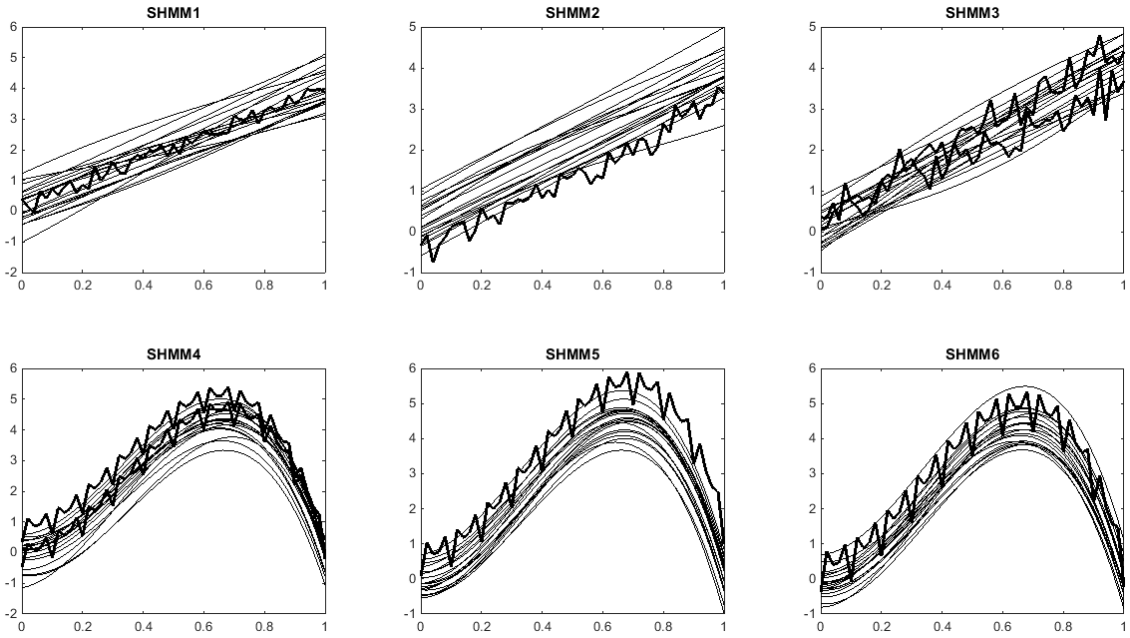


Figure 4-7. Examples of simulated data sets generated by SHMM1 – 6. *Bolded curves* are outlier curves while others are normal curves.

The last two mixture models we consider for generating shape outliers are SHMM7 and SHMM8. SHMM7 shares the same F_{nor} with curves generated by Eq. (4-15), while its outlier curves are generated by:

$$f_{out}(t) = 8t - 2 + \varepsilon(t),$$

which indicates the outliers are different from normal curves in their slopes and intercepts.

For SHMM8, its normal curves are generated by:

$$f_{nor}(t) = a_1 \sin(t) + a_2 \cos(t), t \in [0, 2\pi],$$

where a_1 and a_2 are both continuous random numbers generated from a uniform distribution on $[0.05, 0.15]$. The outlier curves in SHMM8 are generated by:

$$f_{out}(t) = a_1 \sin(t) + \exp\left(\frac{0.69t}{2\pi}\right) a_3 \cos(t),$$

where a_3 is a continuous random number generated from a uniform distribution on $[0.1, 0.15]$. Figure 4-8 displays the simulated data set with at least one outlier for mixture models in SHMM7 and SHMM8.

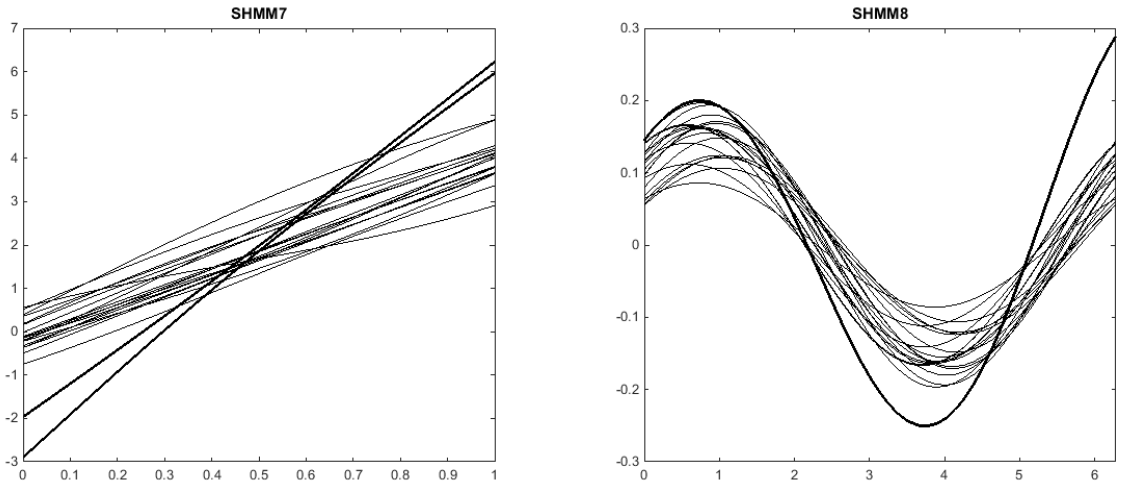


Figure 4-8. Examples of simulated data sets generated by SHMM7 and SHMM8. *Bolded curves* are outlier curves while others are normal curves.

For location outliers, we consider four different models LOMM1 – 4 based on the mixture of two stochastic processes in Eq. (4-14). Specifically, LOMM1 – 4 share the same F_{nor} with curves generated by Eq. (4-15), while varying in their F_{out} components. Specifically, for LOMM1, the outlier curves are generated by:

$$f_{out}(t) = 4\exp(t) + \varepsilon(t),$$

where the outliers are normal in the initial time domain, but become exponentially outlying afterwards. This mixture model is also considered in Sguera et al. (2016).

For LOMM2, the outlier curves are generated by:

$$f_{out}(t) = f_{nor}(t) + K,$$

where K is a constant contaminated location of normal curves.

For LOMM3, the outlier curves are generated by:

$$f_{out}(t) = \begin{cases} f_{nor}(t), & t \in [0, T) \\ f_{nor}(t) + K, t \in [T, 1] \end{cases},$$

where T is a random number generated from a uniform distribution on $[0, 1]$. Different from LOMM2, the resulted outliers are partial contaminations of normal curves. They are the same as normal curves before time T, while increasing by the contamination constant K after time T.

For LOMM4, the outlier curves are generated by:

$$f_{out}(t) = \begin{cases} f_{nor}(t), & t \notin [T, T + l] \\ f_{nor}(t) + K, t \in [T, T + l] \end{cases}, l = \frac{2}{50},$$

where T is a random number generated from a uniform distribution on $[0, 1 - l]$. The resulted outliers are contaminations of peaks of normal curves. LOMM2 – 4 are similar to mixture models considered in López-Pintado and Romo (2009), with different covariance functions of $\varepsilon(t)$.

Figure 4-9 illustrates the simulated data set with at least one outlier for each of the mixture model in LOMM1 – 4. As can be seen, strictly speaking, outliers in LOMM1, LOMM3 and LOMM4 are actually a mixture of location and shape outliers. Since these outliers having different shapes are mainly caused by their largely shifted locations, in this chapter we simply consider them as location outliers.

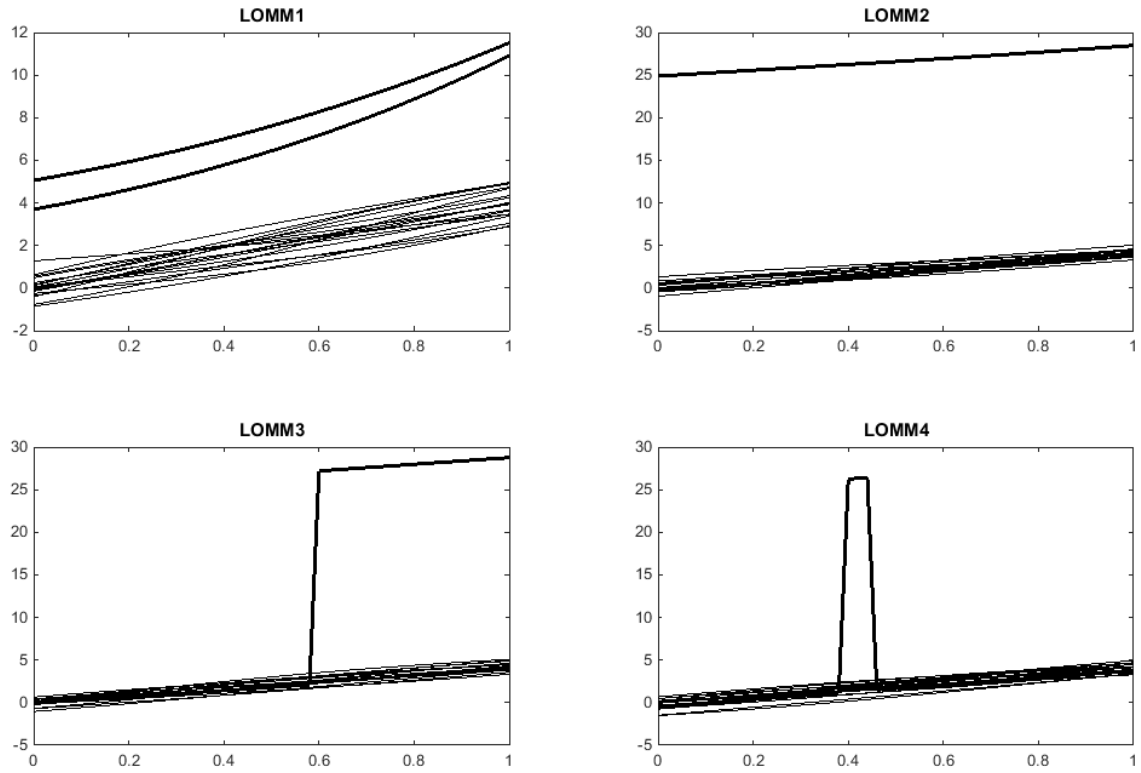


Figure 4-9. Examples of simulated data sets generated by LOMM1 – 4 ($K = 25$ for LOMM2, LOMM3 and LOMM4). *Bolded curves* are outlier curves while others are normal curves.

4.3.2 Performance Evaluation

Let N be the number of simulated data sets for each mixture model, and $n_{out,i}, i = 1, \ldots, N$ be the number of outliers in the i-th data set (in this chapter, we only consider $n_{out,i} > 0$; that is, if $n_{out,i} = 0$, the data set would be re-generated until $n_{out,i} > 0$). For each data set and calculated functional depths of curves in the data set, it is desirable that the $n_{out,i}$ lowest depths are assigned to the $n_{out,i}$ generated outliers. Hence, to evaluate the detection performance of each method, the outlier detection accuracy (ODA) for these N data sets is defined as:

$$ODA = \frac{1}{N}\sum_{i=1}^{N}\frac{n(A_i \cap B_i)}{n_{out,i}}, \tag{4-17}$$

where $A_i = \{f^{(1)}, f^{(2)}, \dots f^{(n_{out,i})}\}$ is the set of curves in the i-th data set having $n_{out,i}$ smallest depths obtained by each depth approach, B_i is the set of all outlier curves generated in the i-th data set, and $n(A_i \cap B_i)$ is the number of common elements in A_i and B_i. It is obvious that the higher the ODA is, the better the detection performance of the method is.

4.3.3 Simulation Results

Our simulation setup is as follows: for each of the mixture model introduced above, we consider $N = 100$ data sets, each consists of $n = 50$ curves with outlier probability $\alpha = 0.05$, contamination constant $K = 25$ for LOMM2, LOMM3 and LOMM4. And each curve is generated using a finite set of 51 equally spaced points in the corresponding time domain. Hence finite dimensional versions of each method are used to calculate the depth of each curve. The kernel function used in HMD is $\kappa(f_{i_1}, f_{i_2}) = \frac{2}{\sqrt{2\pi}}\exp\left(-\frac{d(f_{i_1}, f_{i_2})}{2\kappa_1^2}\right), \kappa_1 > 0$ (Febrero et al., 2008) and the kernel function used in KFSD is $\kappa(f_{i_1}, f_{i_2}) = \exp\left(-\frac{d(f_{i_1}, f_{i_2})}{\kappa_2^2}\right), \kappa_2 > 0$ (Sguera et al., 2016), where the distance measurements $d(f_{i_1}, f_{i_2})$ are both chosen to be the maximum norm, which is consistent with our methods. The tuning parameters $c_1, c_2, \kappa_1, \kappa_2$ in kernel functions of WBD, LWBD, HMD and KFSD are all chosen to be the p-th ($0 < p \leq 100$) percentile of the empirical distribution of $\{d(f_{i_1}, f_{i_2}), 1 \leq i_1 < i_2 \leq n\}$, where $d(f_{i_1}, f_{i_2})$ is also chosen to be the maximum norm. We consider 14 different values of p, and each different value

corresponds to one of 14 different percentiles of corresponding empirical distributions for each data set.

Table 4-1. Simulation results of ODA using WBD, LWBD, HMD, KFSD and BD for mixture models SHMM1 – 8 producing shape outliers (For HMD and KFSD, only results of the best average ODA are shown. See APPENDIX H for their complete results)

	p	SHMM1	SHMM2	SHMM3	SHMM4	SHMM5	SHMM6	SHMM7	SHMM8	Average
WBD	0.1	95.72%	97.17%	98.17%	74.84%	64.88%	72.93%	71.31%	80.61%	81.95%
	0.5	95.72%	97.17%	98.17%	75.73%	66.17%	74.88%	72.38%	84.77%	83.12%
	1	96.22%	97.17%	98.17%	77.85%	67.08%	74.99%	74.23%	84.77%	83.81%
	5	94.38%	97.72%	99.50%	87.45%	84.25%	86.25%	86.38%	86.62%	90.32%
	10	**89.26%**	**95.98%**	**99.25%**	**89.50%**	**91.35%**	**91.43%**	**93.11%**	**91.62%**	**92.69%**
	20	78.26%	88.87%	97.13%	83.80%	94.10%	95.98%	97.68%	98.04%	91.73%
	30	64.78%	84.02%	95.47%	78.60%	94.10%	96.65%	99.67%	99.75%	89.13%
	40	58.96%	79.99%	93.13%	72.68%	92.35%	96.18%	99.83%	99.88%	86.63%
	50	57.53%	77.66%	89.35%	69.43%	87.68%	94.18%	100.00%	99.88%	84.46%
	60	53.41%	72.97%	88.12%	67.76%	85.53%	93.35%	100.00%	99.88%	82.63%
	70	51.05%	69.39%	87.21%	65.26%	83.87%	92.35%	100.00%	99.88%	81.12%
	80	47.31%	65.36%	86.96%	64.93%	82.12%	91.73%	100.00%	99.88%	79.79%
	90	46.86%	61.46%	86.81%	61.81%	79.08%	90.12%	100.00%	99.01%	78.15%
	100	41.35%	59.62%	81.73%	57.83%	72.25%	88.04%	100.00%	96.83%	74.71%
	Overall	**69.34%**	**81.75%**	**92.80%**	**73.39%**	**81.77%**	**88.51%**	**92.47%**	**94.39%**	**84.30%**
LWBD	0.1	90.26%	96.22%	99.75%	64.48%	78.66%	72.25%	75.70%	85.03%	82.79%
	0.5	91.74%	97.22%	99.75%	65.47%	79.48%	73.33%	77.27%	88.53%	84.10%
	1	90.61%	97.72%	99.75%	67.83%	79.87%	74.26%	78.67%	88.53%	84.66%
	5	89.43%	96.13%	99.75%	76.15%	86.93%	84.23%	88.15%	90.73%	88.94%
	10	**81.41%**	**93.77%**	**99.42%**	**79.64%**	**92.91%**	**91.85%**	**92.96%**	**94.20%**	**90.77%**
	20	67.56%	85.27%	95.75%	76.93%	94.93%	96.98%	99.09%	97.47%	89.25%
	30	58.44%	79.58%	95.33%	74.73%	92.85%	97.45%	100.00%	99.55%	87.24%
	40	54.66%	78.02%	91.83%	70.13%	91.10%	96.07%	100.00%	100.00%	85.23%
	50	50.45%	70.55%	90.25%	63.14%	86.97%	94.82%	100.00%	100.00%	82.02%
	60	48.41%	65.31%	87.33%	58.79%	83.46%	92.74%	100.00%	100.00%	79.51%
	70	46.91%	62.11%	85.83%	57.92%	82.34%	92.74%	100.00%	99.50%	78.42%
	80	45.91%	60.16%	82.27%	55.80%	81.59%	92.54%	100.00%	99.50%	77.22%
	90	43.88%	57.58%	80.40%	53.43%	80.63%	91.01%	100.00%	99.30%	75.78%
	100	42.29%	55.25%	77.53%	50.15%	78.46%	88.64%	99.83%	98.52%	73.83%
	Overall	**64.43%**	**78.21%**	**91.78%**	**65.33%**	**85.01%**	**88.49%**	**93.69%**	**95.78%**	**82.84%**
BD		13.91%	18.17%	15.85%	14.80%	19.47%	17.99%	24.73%	21.13%	18.26%
Best HMD		62.79%	76.16%	86.98%	60.43%	60.67%	73.62%	76.70%	85.59%	72.87%
Best KFSD		58.66%	74.74%	85.90%	62.26%	61.62%	75.90%	77.96%	91.72%	73.60%

Table 4-1 presents simulation results of the outlier detection accuracy with comparison of WBD, LWBD, BD, HMD and KFSD (see APPENDIX H for complete results of HMD and KFSD) for the eight mixture models that produce shape outliers. For the first six models SHMM1 – 6, since their levels of irregularity are relatively small, the resulted shape outliers would have high chance to be enclosed by other bands, which

results in poor detection performance of BD. However, as can be seen, the detection accuracy of both WBD and LWBD significantly outperform BD, HMD and KFSD for six models SHMM1 – 6, regardless of values of tuning parameters. For mixture models SHMM7 and SHMM8, WBD, LWBD, HMD and KFSD have similar detection accuracy, while all outperform BD. In comparing WBD and LWBD, WBD has slightly better detection accuracy than LWBD. This demonstrates that the idea of shape distance in WBD and LWBD can be particularly effective in detecting shape outliers confounded with those normal curves that are also fully inside the bands.

Table 4-2. Simulation results of ODA using WBD, LWBD, HMD, KFSD and BD for mixture models LOMM1 – 4 producing location outliers (For HMD and KFSD, only results of the best average ODA are shown. See APPENDIX I for their complete results)

	p	LOMM1	LOMM2	LOMM3	LOMM4	Average
WBD	0.1	65.88%	5.63%	98.93%	99.35%	67.45%
	0.5	67.32%	5.38%	98.93%	99.55%	67.80%
	1	71.00%	5.38%	98.93%	99.88%	68.80%
	5	80.05%	5.72%	100.00%	99.88%	71.41%
	10	86.45%	4.58%	100.00%	99.88%	72.73%
	20	90.18%	5.48%	100.00%	100.00%	73.92%
	30	95.10%	5.65%	100.00%	100.00%	75.19%
	40	99.02%	5.65%	100.00%	100.00%	76.17%
	50	**99.60%**	**5.65%**	**100.00%**	**100.00%**	**76.31%**
	60	99.60%	5.45%	100.00%	100.00%	76.26%
	70	99.60%	5.28%	100.00%	100.00%	76.22%
	80	98.10%	5.78%	100.00%	100.00%	75.97%
	90	82.92%	10.20%	100.00%	100.00%	73.28%
	100	77.23%	14.20%	99.83%	100.00%	72.82%
	Overall	**86.58%**	**6.43%**	**99.76%**	**99.89%**	**73.17%**
LWBD	0.1	65.98%	82.11%	100.00%	99.80%	86.97%
	0.5	66.52%	87.17%	100.00%	100.00%	88.42%
	1	67.85%	89.07%	100.00%	100.00%	89.23%
	5	77.38%	95.94%	100.00%	100.00%	93.33%
	10	87.55%	99.12%	100.00%	100.00%	96.67%
	20	97.23%	99.80%	100.00%	100.00%	99.26%
	30	99.50%	99.80%	100.00%	100.00%	99.83%
	40	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	50	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	60	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	70	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	80	99.86%	100.00%	100.00%	100.00%	99.96%
	90	99.71%	100.00%	100.00%	100.00%	99.93%
	100	97.28%	100.00%	100.00%	100.00%	99.32%
	Overall	**89.92%**	**96.64%**	**100.00%**	**99.99%**	**96.64%**
BD		27.73%	18.12%	12.94%	12.89%	17.92%
Best HMD		100.00%	100.00%	100.00%	100.00%	100.00%
Best KFSD		100.00%	100.00%	100.00%	100.00%	100.00%

For the four mixture models producing location outliers, Table 4-2 gives corresponding simulation results of ODA using WBD, LWBD, BD, HMD and KFSD (see APPENDIX I for complete results of HMD and KFSD). The detection accuracy of three local-oriented depth approaches (LWBD, HMD and KFSD) all outperform WBD and BD. In several cases, they even achieve all perfect detection accuracy. WBD achieves high detection accuracy across six models except for LOMM2, which produces location outliers that have exactly the same shape with normal curves. Therefore WBD has difficulty in detecting them. The low detection accuracy of BD results from having many depth ties – depths of outliers are equal to those of normal curves located near or on the geometric boundary of all curves in the data set. Hence BD lacks strength in distinguishing between them.

4.4 Application

In this section, we study the performance of proposed weighted band depth approaches applied to a real data example from the food industry. Each observation in the data set corresponds to the fat content of a finely chopped meat sample based on the near-infrared absorbance spectrum, with a 100-channel spectrum (that is, 100 equally spaced points) of absorbance in the wavelength range 850 – 1050 nm. This data set separates different meat samples into two levels of fat content – high fat content (fat content greater than 20%, totally 77 observations) and low fat content (fat content less than 20%, totally 138 observations), as shown in Figure 4-10. For more details of this data set, please refer to Ferraty and Vieu (2003) and Rossi and Villa (2006).

It can be observed from Figure 4-10 that the major difference of these two levels of fat content is in the shape of the functional curves. More precisely, they are different in

their curvature (i.e., the second derivative): observations with high fat content have sometimes two local maxima rather than one in observations of low fat content. However, such difference is relatively small, as we can see the first local maxima in observations with high fat content is sometimes rather subtle, which makes it much difficult to be distinguished with observations having low fat content. Because of this, most of the existing methods classify these two levels of fat content by taking the second derivative of the data (Ferraty and Vieu, 2002; Ferraty and Vieu, 2003; Rossi and Villa, 2006), which introduces additional computational cost. Since the shape distance in our proposed depth approaches are particularly sensitive for distinguishing curves with different shapes, we would like to apply our methods directly to the data without taking any derivative.

For the purpose of the outlier detection, we treat observations with low fat content as normal curves while observations having high fat content are outliers. The outlier detection performance of the proposed methods is evaluated based on calculating the ODA defined in Eq. (4-17), with $N = 100$ data sets. Each data set consists of $n = 50$ curves, of which 45 curves (normal curves) are randomly selected from totally 138 observations with low fat content and 5 curves (outliers) are randomly selected from totally 77 observations having high fat content. BD, HMD and KFSD are also implemented as comparison.

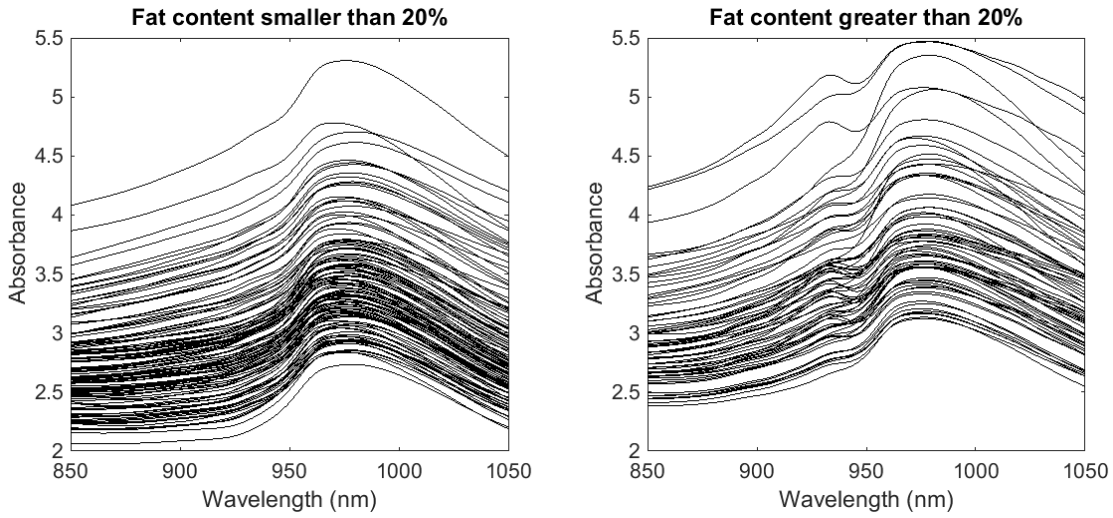


Figure 4-10. The near-infrared absorbance spectrum for both high and low fat content of meat samples.

Table 4-3 gives results of the best ODA using different depth approaches for the fat data (complete results with different values of tuning parameters for WBD, LWBD, HMD and KFSD can be found in APPENDIX J). As can be seen, the proposed depth methods achieve much better detection performance than the other depth approaches, which demonstrate the usefulness of our proposed methods in differentiating curves with various shapes.

Table 4-3. Best ODA of different depth approaches

Method	WBD	LWBD	BD	HMD	KFSD
Best ODA	50.20%	40.60%	9.20%	30.20%	30.40%

4.5 Summary

In this chapter we focus on detecting outliers in functional data by the use of functional depth. Inspired by the band depth (BD) proposed in recent literature, we propose two new depth approaches for functional data – the weighted band depth (WBD) and its localized version (LWBD). By introducing the idea of shape distance, which is

highly sensitive to shape differences among curves, both WBD and LWBD have substantially improved strength in detecting various shape outliers compared to BD. The LWBD is also robust in detecting location outliers by using the localization idea that takes into account the relative location information among curves.

The detection performance of our proposed depth approaches are evaluated through both simulated data and a real data example. Results demonstrate advantages of WBD and LWBD over BD in detecting both shape and location outliers. LWBD also shows its robustness in detecting location outliers and is less sensitive to the tuning parameter compared to other popular local-oriented functional depth approaches.

CHAPTER 5. CONCLUSION AND FUTURE WORK

In this thesis, we study change and outlier detection problems and their applications for longitudinal and functional data. Two types of change detection problems are considered. The first type of problems concerns about on-line change detection within a single longitudinal trend. In particularly, we discuss two engineering applications. The first application focuses on the on-line steady state detection, where we would like to identify the transition point between the transient period and steady state period in real time. In Chapter 2, we develop a new on-line steady state detection algorithm under the Bayesian framework based on a multiple change-point state space model and the sequential Monte Carlo methods. The main contribution of this work is by using the Rao-Blackwellization technique, variance of Monte Carlo estimation is significantly reduced and the computational efficiency is significantly improved compared to the standard SMC algorithm. A resampling method called the Optimal Resampling is also used to make more efficient use of the particle information and further improve the robustness of the steady state detection.

In Chapter 3, we consider another engineering application of the on-line change detection, which is on-line detection of changes of process mean in the short-run process. Inspired by the model proposed by Tsiamyrtzis and Hawkins (2005), we propose two new models that are more suitable and flexible for practical situations. There are three main contributions of this work. Firstly, compared to the exponential computational cost in Tsiamyrtzis and Hawkins (2005), our work leads to significantly reduced computation burden. Secondly, our method is shown to be more robust to parameter misspecifications by requiring less knowledge on the process parameters (such as the jump size and

measurement noise variance). Thirdly, a new efficient approximation method is also proposed for processes with constant and unknown variance, while the computational cost is unchanged.

The other type of change detection problems is studied in Chapter 4, where we focus on outlier detection for a set of longitudinal or functional data. Inspired by the band depth in López-Pintado and Romo (2009), we propose two new depth approaches for functional data – the weighted band depth and the localized weighted band depth. Compared to existing depth approaches for functional data, our proposed depth notions are shown to be more robust in detecting a variety type of outliers (both shape and location outliers) by using our novel idea called the shape distance.

For the on-line change detection problems, note that we only consider the problem with univariate observations. A potential future work is to develop methods for multivariate data. Regarding the outlier detection for functional data, based on our current work, in the future more applications of WBD and LWBD can be explored, such as making robust inference (the trimmed mean) or performing classification for functional data. We can also consider a more general definition of the band in WBD and LWBD with more than two boundary curves. Additionally, study on theoretical properties of the proposed functional depths can be also a potential future work.

APPENDIX A. PROOF OF LEMMA 2-1 IN SECTION 2.2.2

The joint posterior distribution of $\boldsymbol{\beta}_t$ and σ_t^2 is:

$$p(\boldsymbol{\beta}_t, \sigma_t^2|\tau_t, \mathbf{y}_{\tau_t:t}) = p(\boldsymbol{\beta}_t|\sigma_t^2, \tau_t, \mathbf{y}_{\tau_t:t})p(\sigma_t^2|\tau_t, \mathbf{y}_{\tau_t:t})$$

$$\propto (\sigma_t^2)^{-1} exp\left\{-\frac{1}{2\sigma_t^2}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})^T \boldsymbol{\Sigma}_{\tau_t,t}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})\right\}(\sigma_t^2)^{-(a_{\tau_t,t}+1)} exp\left\{-\frac{b_{\tau_t,t}}{\sigma_t^2}\right\}$$

$$= (\sigma_t^2)^{-(a_{\tau_t,t}+1+1)} exp\left\{-\frac{1}{\sigma_t^2}\left(b_{\tau_t,t} + \frac{1}{2}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})^T \boldsymbol{\Sigma}_{\tau_t,t}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})\right)\right\}$$

Let $A = b_{\tau_t,t} + \frac{1}{2}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})^T \boldsymbol{\Sigma}_{\tau_t,t}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})$

The marginal posterior distribution of $\boldsymbol{\beta}_t$ can be obtained after integrating out σ_t^2:

$$p(\boldsymbol{\beta}_t|\tau_t, \mathbf{y}_{\tau_t:t}) = \int p(\boldsymbol{\beta}_t, \sigma_t^2|\tau_t, \mathbf{y}_{\tau_t:t})d\sigma_t^2 \propto \int (\sigma_t^2)^{-(a_{\tau_t,t}+1+1)}\, exp\left\{-\frac{1}{\sigma_t^2}A\right\}d\sigma_t^2$$

$$= \int \frac{A^{a_{\tau_t,t}+1}}{\Gamma(a_{\tau_t,t}+1)}(\sigma_t^2)^{-(a_{\tau_t,t}+1+1)}\, exp\left\{-\frac{1}{\sigma_t^2}A\right\}d\sigma_t^2 \frac{\Gamma(a_{\tau_t,t}+1)}{A^{a_{\tau_t,t}+1}} \propto A^{-(a_{\tau_t,t}+1)}$$

$$= \left(b_{\tau_t,t} + \frac{1}{2}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})^T \boldsymbol{\Sigma}_{\tau_t,t}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})\right)^{-(a_{\tau_t,t}+1)}$$

$$\propto \left(1 + \frac{1}{2a_{\tau_t,t}} \frac{(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})^T \boldsymbol{\Sigma}_{\tau_t,t}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_{\tau_t,t})}{\frac{b_{\tau_t,t}}{a_{\tau_t,t}}}\right)^{-\left(\frac{2a_{\tau_t,t}+2}{2}\right)}$$

Therefore $(\boldsymbol{\beta}_t|\tau_t, \mathbf{y}_{\tau_t:t}) \sim t_2\left(\boldsymbol{\mu}_{\tau_t,t}, \frac{b_{\tau_t,t}}{a_{\tau_t,t}}\boldsymbol{\Sigma}_{\tau_t,t}, 2a_{\tau_t,t}\right)$.

APPENDIX B. PROOF OF LEMMA 2-2 IN SECTION 2.2.2

(a). If $\tau_t = \tau_{t-1}$, we are making posterior inferences based on the observations $\mathbf{y}_{\tau_{t-1}:t-1}$.

Since $(\boldsymbol{\beta}_{t-1}|\sigma_{t-1}^2, \tau_t = \tau_{t-1}, \mathbf{y}_{\tau_{t-1}:t-1}) \sim N(\boldsymbol{\mu}_{\tau_t,t-1}, \sigma_{t-1}^2 \boldsymbol{\Sigma}_{\tau_t,t-1})$, we have

$$(\mathbf{X}_t\boldsymbol{\beta}_{t-1}|\sigma_{t-1}^2, \tau_t = \tau_{t-1}, \mathbf{y}_{\tau_{t-1}:t-1}) \sim N(\mathbf{X}_t\boldsymbol{\mu}_{\tau_t,t-1}, \sigma_{t-1}^2 \mathbf{X}_t\boldsymbol{\Sigma}_{\tau_t,t-1}\mathbf{X}_t^T)$$

And the conditional distribution of y_t would be:

$$(y_t|\sigma_{t-1}^2, \tau_t = \tau_{t-1}, \mathbf{y}_{\tau_{t-1}:t-1}) \sim N\left(\mathbf{X}_t\boldsymbol{\mu}_{\tau_t,t-1}, \sigma_{t-1}^2\left(1 + \mathbf{X}_t\boldsymbol{\Sigma}_{\tau_t,t-1}\mathbf{X}_t^T\right)\right)$$

The joint posterior distribution of y_t, σ_{t-1}^2 is:

$$p\left(y_t, \sigma_{t-1}^2|\tau_t = \tau_{t-1}, \mathbf{y}_{\tau_{t-1}:t-1}\right)$$

$$= p(y_t|\sigma_{t-1}^2, \tau_t = \tau_{t-1}, \mathbf{y}_{\tau_{t-1}:t-1}) p\left(\sigma_{t-1}^2|\tau_t = \tau_{t-1}, \mathbf{y}_{\tau_{t-1}:t-1}\right)$$

$$\propto (\sigma_{t-1}^2)^{-1/2} exp\left\{-\frac{\left(y_t - \mathbf{X}_t\boldsymbol{\mu}_{\tau_t,t-1}\right)^2}{2\sigma_{t-1}^2(1 + \mathbf{X}_t\boldsymbol{\Sigma}_{\tau_t,t-1}\mathbf{X}_t^T)}\right\} (\sigma_{t-1}^2)^{-(a_{\tau_t,t-1}+1)} exp\left\{-\frac{b_{\tau_t,t-1}}{\sigma_{t-1}^2}\right\}$$

$$= (\sigma_{t-1}^2)^{-\left(a_{\tau_t,t-1}+\frac{1}{2}+1\right)} exp\left\{-\frac{1}{\sigma_{t-1}^2}\left(b_{\tau_t,t-1} + \frac{\left(y_t - \mathbf{X}_t\boldsymbol{\mu}_{\tau_t,t-1}\right)^2}{2(1 + \mathbf{X}_t\boldsymbol{\Sigma}_{\tau_t,t-1}\mathbf{X}_t^T)}\right)\right\}$$

Let $B = b_{\tau_t,t-1} + \frac{\left(y_t - \mathbf{X}_t\boldsymbol{\mu}_{\tau_t,t-1}\right)^2}{2(1+\mathbf{X}_t\boldsymbol{\Sigma}_{\tau_t,t-1}\mathbf{X}_t^T)}$

The posterior predictive distribution of y_t can be obtained after integrating out σ_{t-1}^2:

$$p\left(y_t|\tau_t = \tau_{t-1}, \mathbf{y}_{\tau_{t-1}:t-1}\right) = \int p\left(y_t, \sigma_{t-1}^2|\tau_t = \tau_{t-1}, \mathbf{y}_{\tau_{t-1}:t-1}\right) d\sigma_{t-1}^2$$

$$\propto \int (\sigma_{t-1}^2)^{-\left(a_{\tau_t,t-1}+\frac{1}{2}+1\right)} exp\left\{-\frac{1}{\sigma_{t-1}^2}B\right\} d\sigma_{t-1}^2$$

$$= \int \frac{B^{a_{\tau_t,t-1}+\frac{1}{2}}}{\Gamma\left(a_{\tau_t,t-1} + \frac{1}{2}\right)} (\sigma_{t-1}^2)^{-\left(a_{\tau_t,t-1}+\frac{1}{2}+1\right)} exp\left\{-\frac{1}{\sigma_{t-1}^2}B\right\} d\sigma_{t-1}^2 \frac{\Gamma\left(a_{\tau_t,t-1} + \frac{1}{2}\right)}{B^{a_{\tau_t,t-1}+\frac{1}{2}}}$$

$$\propto B^{-\left(a_{\tau_t,t-1}+\frac{1}{2}\right)} = \left(b_{\tau_t,t-1} + \frac{\left(y_t - \mathbf{X}_t\boldsymbol{\mu}_{\tau_t,t-1}\right)^2}{2(1 + \mathbf{X}_t\boldsymbol{\Sigma}_{\tau_t,t-1}\mathbf{X}_t^T)}\right)^{-\left(a_{\tau_t,t-1}+\frac{1}{2}\right)}$$

$$\propto \left(1+\frac{1}{2a_{\tau_t,t-1}}\frac{\left(y_t-\mathbf{X}_t\boldsymbol{\mu}_{\tau_t,t-1}\right)^2}{\frac{b_{\tau_t,t-1}}{a_{\tau_t,t-1}}\left(1+\mathbf{X}_t\boldsymbol{\Sigma}_{\tau_t,t-1}\mathbf{X}_t^T\right)}\right)^{-\left(\frac{2a_{\tau_t,t-1}+1}{2}\right)}$$

Therefore $\left(y_t|\tau_t=\tau_{t-1},\mathbf{y}_{\tau_{t-1}:t-1}\right)\sim t_1\left(\mathbf{X}_t\boldsymbol{\mu}_{\tau_t,t-1},\frac{b_{\tau_t,t-1}}{a_{\tau_t,t-1}}\left(1+\mathbf{X}_t\boldsymbol{\Sigma}_{\tau_t,t-1}\mathbf{X}_t^T\right),2a_{\tau_t,t-1}\right)$

(b). If $\tau_t=t$, y_t is independent of $\mathbf{y}_{0:t-1}$ and we are making posterior inferences based on the prior information. That is to say, $p\left(y_t|\tau_t=t,\mathbf{y}_{\tau_t:t-1}\right)=p(y_t)$. Therefore, we have:

$\left(y_t|\tau_t=t,\mathbf{y}_{\tau_t:t-1}\right)\sim t_1\left(\mathbf{X}_t\boldsymbol{\mu}_0,\frac{b_0}{a_0}\left(1+\mathbf{X}_t\boldsymbol{\Sigma}_0\mathbf{X}_t^T\right),2a_0\right)$.

APPENDIX C. THE STRATIFIED SAMPLING ALGORITHM IN SECTION 2.2.3

Let $w_t^{(i)}$, $i = 1,2,\dots,n_d + 1 - q$ be the weights of remaining particles to be resampled, a total of $n_d - q$ particles need to be resampled. With the unique solution c, we apply the stratified sampling algorithm of Carpenter et al. (1999):

- Initialize: simulate s as the realization of a uniform random variable on $[0, 1/c]$, and set $i = 1$.
- While $i \leq n_d + 1 - q$ do:

  $s = s - w_t^{(i)}$.

  If $s < 0$ do

    Resample particle i, assign it with weight $1/c$;

    $s = s + 1/c$.

  End

  $i = i + 1$.

End

APPENDIX D. NORMAL APPROXIMATION OF RBSMC DETECTION RESULTS IN SECTION 2.3.4

RBSMC detection results for AR(0) noise with both exact calculations and normal approximations for the pdf of student's t distribution:

Signal		σ	WSDE		FAR	
			Exact	Approximation	Exact	Approximation
Linear	h=1, T_0=200	0.06	52.6	52.6	0	0
		0.10	55.7	55.0	0	0
		0.14	59.6	59.6	0	0
	h=1, T_0=300	0.06	46.7	45.9	0	0
		0.10	49.9	49.5	0	0
		0.14	53.5	54.5	0.01	0.01
Quadratic	h=1, T_0=200	0.06	17.6	17.4	0.01	0.01
		0.10	22.8	22.8	0.03	0.02
		0.14	26.3	27.5	0.06	0.06
	h=1, T_0=300	0.06	20.0	21.3	0.89	0.93
		0.10	24.4	23.7	0.81	0.77
		0.14	29.9	28.7	0.71	0.75
Exponential	h=1, T_0=200	0.06	23.3	24.5	0.01	0.01
		0.10	28.1	28.9	0.07	0.06
		0.14	31.7	32.5	0.12	0.11
	h=1, T_0=300	0.06	37.6	38.9	0.93	0.92
		0.10	45.4	44.0	0.85	0.87
		0.14	51.1	52.1	0.85	0.86
Oscillating	h=1, T_0=200	0.06	23.1	23.4	0	0
		0.10	29.3	28.6	0	0
		0.14	38.7	40.0	0	0
	h=1, T_0=300	0.06	9.4	9.8	0.39	0.39
		0.10	19.5	20.8	0.21	0.19
		0.14	32.6	31.4	0.07	0.05
Overall			37.3	37.3	0.25	0.25

APPENDIX E. PROOF OF LEMMA 3-1 IN SECTION 3.2.2

(1) If $s < t$, using posterior distributions from line segment $\mathbf{y}_{s:t-1}$ as prior, based on the KF recursive procedures we can easily obtain:

$$\mu_{\theta_{s,t}} = \frac{\sigma^2}{\sigma^2_{\theta_{s,t-1}} + \sigma^2_{RW} + \sigma^2}\mu_{\theta_{s,t-1}} + \frac{\sigma^2_{\theta_{s,t-1}} + \sigma^2_{RW}}{\sigma^2_{\theta_{s,t-1}} + \sigma^2_{RW} + \sigma^2} y_t$$

$$\sigma^2_{\theta_{s,t}} = \frac{\sigma^2_{\theta_{s,t-1}} + \sigma^2_{RW}}{\sigma^2_{\theta_{s,t-1}} + \sigma^2_{RW} + \sigma^2}\sigma^2$$

Since $\frac{P(s,t)}{P(s,t-1)} = p(y_t|\mathbf{y}_{s:t-1})$. With KF, $(\theta_{t-1}|\mathbf{y}_{s:t-1}) \sim N\left(\mu_{\theta_{s,t-1}}, \sigma^2_{\theta_{s,t-1}}\right)$, and following model (3-2), $(\theta_t|\theta_{t-1}) \sim N(\theta_{t-1}, \sigma^2_{RW})$, it is easy to have $(\theta_t|\mathbf{y}_{s:t-1}) \sim N\left(\mu_{\theta_{s,t-1}}, \sigma^2_{\theta_{s,t-1}} + \sigma^2_{RW}\right)$. Since $(y_t|\theta_t) \sim N(\theta_t, \sigma^2)$, we have $(y_t|\mathbf{y}_{s:t-1}) \sim N\left(\mu_{\theta_{s,t-1}}, \sigma^2_{\theta_{s,t-1}} + \sigma^2 + \sigma^2_{RW}\right)$.

Therefore we have $\frac{P(s,t)}{P(s,t-1)} = p(y_t|\mathbf{y}_{s:t-1}) \sim N\left(\mu_{\theta_{s,t-1}}, \sigma^2_{\theta_{s,t-1}} + \sigma^2 + \sigma^2_{RW}\right)$.

(2) If $s = t$, following the same logic, we can obtain:

$\mu_{\theta_{t,t}} = \frac{\sigma^2}{\sigma^2_J + \sigma^2_{RW} + \sigma^2}\mu_J + \frac{\sigma^2_J + \sigma^2_{RW}}{\sigma^2_J + \sigma^2_{RW} + \sigma^2} y_t$, $\sigma^2_{\theta_{t,t}} = \frac{\sigma^2_J + \sigma^2_{RW}}{\sigma^2_J + \sigma^2_{RW} + \sigma^2}\sigma^2$, and

$P(t,t) = p(y_t) \sim N\left(\mu_J, \sigma^2_{RW} + \sigma^2 + \sigma^2_J\right)$.

APPENDIX F. PROOF OF LEMMA 3-2 IN SECTION 3.2.3

(1) The proof follows the same procedure in Chapter 9 (Pages 244-246) of O'Hagan (1994).

(2) Since the process mean $\theta_t = \mathbf{x}_t\boldsymbol{\beta}_t$. Based on the results from (1), we have:

$$(\theta_t|\sigma_t^2, \tau_t = s, \boldsymbol{y}_{s:t}) \sim N(\mathbf{x}_t\boldsymbol{\mu}_t^s, \sigma_t^2\mathbf{x}_t\mathbf{M}_t^s\mathbf{x}_t^T)$$

The joint posterior distribution of θ_t and σ_t^2 is:

$$p(\theta_t, \sigma_t^2|\tau_t = s, \mathbf{y}_{s:t}) = p(\theta_t|\sigma_t^2, \tau_t = s, \mathbf{y}_{s:t})p(\sigma_t^2|\tau_t = s, \mathbf{y}_{s:t})$$

$$\propto (\sigma_t^2)^{-\frac{1}{2}} exp\left\{-\frac{(\theta_t - \mathbf{x}_t\boldsymbol{\mu}_t^s)^2}{2\sigma_t^2\mathbf{x}_t\mathbf{M}_t^s\mathbf{x}_t^T}\right\}(\sigma_t^2)^{-\left(\frac{d_t^s}{2}+1\right)} exp\left\{-\frac{H_t^s}{2\sigma_t^2}\right\}$$

$$= (\sigma_t^2)^{-\left(\frac{d_t^s}{2}+\frac{1}{2}+1\right)} exp\left\{-\frac{1}{\sigma_t^2}\left(\frac{H_t^s}{2} + \frac{(\theta_t - \mathbf{x}_t\boldsymbol{\mu}_t^s)^2}{2\mathbf{x}_t\mathbf{M}_t^s\mathbf{x}_t^T}\right)\right\}$$

Let $A = \frac{H_t^s}{2} + \frac{(\theta_t - \mathbf{x}_t\boldsymbol{\mu}_t^s)^2}{2\mathbf{x}_t\mathbf{M}_t^s\mathbf{x}_t^T}$, the marginal posterior distribution of θ_t can be obtained after integrating out σ_t^2:

$$p(\theta_t|\tau_t = s, \mathbf{y}_{s:t}) = \int p(\theta_t, \sigma_t^2|\tau_t = s, \mathbf{y}_{s:t})d\sigma_t^2 \propto \int (\sigma_t^2)^{-\left(\frac{d_t^s}{2}+\frac{1}{2}+1\right)} exp\left\{-\frac{A}{\sigma_t^2}\right\}d\sigma_t^2$$

$$= \int \frac{A^{\frac{d_t^s}{2}+\frac{1}{2}}}{\Gamma\left(\frac{d_t^s}{2}+\frac{1}{2}\right)}(\sigma_t^2)^{-\left(\frac{d_t^s}{2}+\frac{1}{2}+1\right)} exp\left\{-\frac{A}{\sigma_t^2}\right\}d\sigma_t^2 \frac{\Gamma\left(\frac{d_t^s}{2}+\frac{1}{2}\right)}{A^{\frac{d_t^s}{2}+\frac{1}{2}}} \propto \left(1 + \frac{1}{d_t^s}\frac{(\theta_t - \mathbf{x}_t\boldsymbol{\mu}_t^s)^2}{\frac{H_t^s}{d_t^s}\mathbf{x}_t\mathbf{M}_t^s\mathbf{x}_t^T}\right)^{-\left(\frac{d_t^s+1}{2}\right)}$$

Therefore $(\theta_t|\tau_t = s, \mathbf{y}_{s:t}) \sim t\left(\mathbf{x}_t\boldsymbol{\mu}_t^s, \frac{H_t^s}{d_t^s}\mathbf{x}_t\mathbf{M}_t^s\mathbf{x}_t^T, d_t^s\right)$.

APPENDIX G. CALCULATION OF $P(s,t)$ IN SECTION 3.2.3

$$P(s,t) = \int p(\mathbf{y}_{s:t}|\sigma_t^2, \boldsymbol{\beta}_t) p(\sigma_t^2) p(\boldsymbol{\beta}_t|\sigma_t^2) d\sigma_t^2 d\boldsymbol{\beta}_t$$

$$= \int \left[\frac{\left(\frac{\gamma}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} (\sigma_t^2)^{-\frac{v}{2}-1} e^{-\frac{\gamma}{2\sigma_t^2}} \right] \left\{ (2\pi)^{-1} |\sigma_t^2 \boldsymbol{\Sigma}_0|^{-\frac{1}{2}} exp\left[-\frac{(\boldsymbol{\beta}_t - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta}_t - \boldsymbol{\mu}_0)}{2\sigma_t^2} \right] \right\}$$

$$\left\{ (2\pi)^{-\frac{t-s+1}{2}} (\sigma_t^2)^{-\frac{t-s+1}{2}} exp\left[-\frac{\|\mathbf{y}_{s:t} - (\mathbf{X}_t^s \boldsymbol{\beta}_t)^T\|^2}{2\sigma_t^2} \right] \right\} d\sigma_t^2 d\boldsymbol{\beta}_t$$

$$= \int \frac{\left(\frac{\gamma}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} (2\pi)^{-\frac{t-s+3}{2}} (\sigma_t^2)^{-\frac{t-s+v+4}{2}} |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} exp\left\{ -\frac{H_t^s}{2\sigma_t^2} \right\}$$

$$exp\left\{ -\frac{(\boldsymbol{\beta}_t - \boldsymbol{\mu}_t^s)^T ((\mathbf{X}_t^s)^T \mathbf{X}_t^s + \boldsymbol{\Sigma}_0^{-1})(\boldsymbol{\beta}_t - \boldsymbol{\mu}_t^s)}{2\sigma_t^2} \right\} d\sigma_t^2 d\boldsymbol{\beta}_t$$

$$= \int \frac{\left(\frac{\gamma}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} (2\pi)^{-\frac{t-s+1}{2}} (\sigma_t^2)^{-\frac{t-s+v+3}{2}} |\boldsymbol{\Sigma}_0|^{-\frac{1}{2}} exp\left\{ -\frac{H_t^s}{2\sigma_t^2} \right\} |(\mathbf{X}_t^s)^T \mathbf{X}_t^s + \boldsymbol{\Sigma}_0^{-1}|^{-\frac{1}{2}} d\sigma_t^2$$

$$= \pi^{-(t-s+1)/2} \left(\frac{|\mathbf{M}_t^s|}{|\boldsymbol{\Sigma}_0|} \right)^{\frac{1}{2}} \frac{\gamma^{\frac{v}{2}}}{(H_t^s)^{(t-s+1+v)/2}} \frac{\Gamma\left(\frac{t-s+1+v}{2}\right)}{\Gamma(v/2)}$$

APPENDIX H. COMPLETE ODA RESULTS OF HMD AND KFSD FOR SHMM1 – 8 IN SECTION 4.3.3

	p	SHMM1	SHMM2	SHMM3	SHMM4	SHMM5	SHMM6	SHMM7	SHMM8	Average
HMD	0.1	64.04%	69.92%	78.98%	60.87%	55.92%	65.17%	62.04%	55.59%	64.07%
	0.5	**62.79%**	**76.16%**	**86.98%**	**60.43%**	**60.67%**	**73.62%**	**76.70%**	**85.59%**	**72.87%**
	1	60.95%	75.16%	86.82%	60.51%	60.37%	74.07%	77.37%	86.54%	72.72%
	5	47.40%	67.57%	80.47%	60.81%	64.88%	76.10%	82.67%	98.17%	72.26%
	10	37.48%	56.33%	69.90%	55.69%	64.50%	75.70%	92.62%	99.88%	69.01%
	20	29.43%	45.22%	61.55%	47.77%	58.10%	73.72%	97.99%	100.00%	64.22%
	30	26.05%	41.12%	57.04%	46.10%	55.02%	68.39%	99.19%	100.00%	61.61%
	40	24.39%	37.42%	54.70%	44.15%	54.40%	65.27%	98.69%	100.00%	59.88%
	50	24.14%	35.05%	49.47%	41.65%	53.15%	63.44%	98.69%	100.00%	58.20%
	60	22.88%	33.86%	47.14%	41.95%	51.90%	63.10%	98.69%	100.00%	57.44%
	70	22.21%	33.99%	44.50%	41.12%	51.90%	62.60%	98.44%	100.00%	56.85%
	80	21.38%	32.96%	42.85%	39.12%	50.65%	61.54%	98.44%	100.00%	55.87%
	90	20.88%	32.38%	39.76%	39.25%	49.90%	60.17%	98.24%	100.00%	55.07%
	100	20.63%	31.59%	38.76%	38.72%	48.15%	58.50%	98.24%	100.00%	54.32%
	Overall	**34.62%**	**47.77%**	**59.92%**	**48.44%**	**55.68%**	**67.24%**	**91.29%**	**94.70%**	**62.46%**
KFSD	0.1	51.05%	57.82%	63.13%	51.17%	45.70%	51.34%	45.79%	30.12%	49.51%
	0.5	63.29%	75.66%	79.48%	60.29%	58.33%	68.54%	71.79%	74.79%	69.02%
	1	63.79%	77.36%	87.48%	60.79%	60.42%	73.29%	75.70%	86.12%	73.12%
	5	**58.66%**	**74.74%**	**85.90%**	**62.26%**	**61.62%**	**75.90%**	**77.96%**	**91.72%**	**73.60%**
	10	49.18%	69.51%	84.90%	62.55%	63.35%	77.32%	80.25%	97.02%	73.01%
	20	41.88%	63.92%	74.36%	57.64%	66.68%	76.59%	90.25%	99.50%	71.35%
	30	37.48%	55.29%	69.85%	54.19%	62.97%	74.99%	94.01%	99.88%	68.58%
	40	33.08%	50.87%	65.14%	51.35%	59.27%	73.30%	96.00%	99.88%	66.11%
	50	30.40%	47.55%	62.05%	47.99%	56.85%	70.37%	97.37%	99.88%	64.06%
	60	28.15%	43.78%	59.90%	46.00%	53.78%	68.87%	98.46%	99.88%	62.35%
	70	26.45%	42.37%	58.32%	44.89%	52.95%	64.85%	98.37%	99.88%	61.01%
	80	24.45%	39.41%	54.95%	43.89%	49.72%	60.90%	97.06%	99.42%	58.72%
	90	23.50%	36.52%	50.73%	41.19%	47.38%	58.08%	95.21%	97.82%	56.30%
	100	20.46%	33.54%	38.60%	37.96%	42.73%	49.50%	84.76%	94.96%	50.32%
	Overall	**39.42%**	**54.88%**	**66.77%**	**51.58%**	**55.84%**	**67.42%**	**85.93%**	**90.77%**	**64.08%**

APPENDIX I. COMPLETE ODA RESULTS OF HMD AND KFSD FOR LOMM1 – 4 IN SECTION 4.3.3

	p	LOMM1	LOMM2	LOMM3	LOMM4	Average
HMD	0.1	63.55%	88.33%	99.86%	100.00%	87.94%
	0.5	72.13%	93.73%	100.00%	100.00%	91.47%
	1	72.67%	95.67%	100.00%	100.00%	92.08%
	5	81.65%	98.57%	100.00%	100.00%	95.05%
	10	90.18%	99.02%	100.00%	100.00%	97.30%
	20	97.35%	99.58%	100.00%	100.00%	99.23%
	30	99.75%	100.00%	100.00%	100.00%	99.94%
	40	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	50	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	60	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	70	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	80	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	90	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	100	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	Overall	**91.23%**	**98.21%**	**99.99%**	**100.00%**	**97.36%**
KFSD	0.1	45.35%	80.35%	99.86%	99.68%	81.31%
	0.5	64.80%	82.40%	99.86%	99.80%	86.71%
	1	67.63%	84.68%	99.86%	100.00%	88.04%
	5	73.95%	91.70%	100.00%	100.00%	91.41%
	10	79.63%	95.10%	100.00%	100.00%	93.68%
	20	84.38%	97.70%	100.00%	100.00%	95.52%
	30	91.43%	98.57%	100.00%	100.00%	97.50%
	40	94.95%	99.02%	100.00%	100.00%	98.49%
	50	97.98%	99.02%	100.00%	100.00%	99.25%
	60	99.02%	99.02%	100.00%	100.00%	99.51%
	70	99.55%	99.58%	100.00%	100.00%	99.78%
	80	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**
	90	99.22%	100.00%	100.00%	100.00%	99.80%
	100	96.38%	100.00%	100.00%	100.00%	99.10%
	Overall	**85.31%**	**94.80%**	**99.97%**	**99.96%**	**95.01%**

APPENDIX J. COMPLETE ODA RESULTS OF WBD, LWBD, HMD AND KFSD FOR OUTLIER DETECTION IN SECTION 4.4

p	WBD	LWBD	HMD	KFSD
0.1	28.20%	5.60%	6.20%	5.60%
0.5	49.80%	**40.60%**	**30.20%**	29.80%
1	**50.20%**	40.00%	29.60%	**30.40%**
5	47.20%	34.40%	21.00%	28.40%
10	45.00%	25.60%	18.20%	24.60%
20	43.00%	18.00%	15.80%	18.60%
30	36.60%	17.40%	15.40%	16.60%
40	23.80%	17.00%	14.80%	15.80%
50	19.60%	12.60%	14.20%	15.60%
60	17.20%	11.40%	13.80%	15.40%
70	16.40%	9.20%	13.80%	14.80%
80	15.60%	8.40%	13.80%	13.80%
90	15.00%	8.40%	13.80%	12.60%
100	13.80%	8.40%	13.80%	12.00%

REFERENCES

Apley, D. W. (2012), 'Posterior Distribution Charts: A Bayesian Approach for Graphically Exploring a Process Mean', Technometrics, 54, 296-310.

Arulampalam, M.S., Maskell, S., Gordon, N. and Clapp, T. (2002), 'A tutorial on particle filters for online non-linear/non-Gaussion Bayesian tracking', IEEE Transactions on Signal Processing, 50, 174-188.

Bagajewicz, M.J. and Jiang, Q. (2000), 'Comparison of steady state and integral dynamic data reconciliation', Computers and Chemical Engineering, 24, 2367-2383.

Bethea, R.M. and Rhinehart, R.R. (1991), 'Applied engineering statistics': Marcel Dekker.

Bhat, S.A. and Saraf, D.N. (2004), 'Steady-state identification, gross error detection, and data reconcilation for industrial process units', Industrial & Engineering Chemistry Research, 43, 4323-4336.

Cao, S. and Rhinehart, R.R. (1995), 'An efficient method for on-line identification of steady state', Journal of Process Control, 5, 363-374.

Carpenter, J., Clifford, P. and Fearnhead, P. (1999), 'An improved particle filter for non-linear problems', IEE Proceedings Radar Sonar and Navigation, 146, 2-7.

Casella, G. and Robert, C.P. (1996), 'Rao-Blackwellization of sampling schemes', Biometrika, 83, 81-94.

Chakraborty, A. and Chaudhuri, P. (2014), 'On data depth in infinite dimensional spaces', Annals of the Institute of Statistical Mathematics, 66, 303-324.

Chaudhuri, P. (1996), 'On a Geometric Notion of Quantiles for Multivariate Data', Journal of the American Statistical Association, 91, 862-872.

Chen, N. and Tsui, K.L. (2013), 'Condition Monitoring and Remaining Useful Life Prediction Using Degradation Signals: Revisited', IIE Transactions, 45, 939-952.

Chopin, N. (2007), 'Dynamic detection of change points in long time series', Annals of the Institute of Statistical Mathematics, 59, 349-366.

Crow, E.L., Davis, F.A. and Maxfiled, M.W. (1960), ' Statistics Manual: with examples taken from ordnance development', New York: Dover Publications.

Crowder, S.V. and Eshleman, L. (2001), 'Small sample properties of an adaptive filter applied to low volume SPC', Journal of Quality Technology, 33, 29-46.

Cuevas, A., Febrero, M. and Fraiman, R. (2006), 'On the use of bootstrap for estimating functions with functional data', Computational Statistics and Data Analysis, 51, 1063-1074.

Cuevas, A., Febrero, M. and Fraiman, R. (2007), 'Robust estimation and classification for functional data via projection-based depth notions', Computational Statistics, 22, 481-496.

Del Castillo, E. and Montgomery, D.C. (1992), 'Average run length properties of Q-charts for short-run statistical control: enhancements and alternative method', Publication 92-30, Quality and Reliability Group, Arizona State University.

Doucet, A., de Freitas, N. and Gordon, N. (2001), 'Sequential Monte Carlo methods in practice', Berlin Heidelberg New York, Springer.

Doucet, A., Godsill, S. and Andrieu, C. (2000), 'On sequential Monte Carlo sampling methods for Bayesian filtering', Statistics and Computing, 10, 197-208.

Fearnhead, P. and Clifford, P. (2003), 'On-line inference for hidden Markov models via particle filters', Journal of the Royal Statistical Society, Series B, 65, 887-899.

Fearnhead, P. and Liu, Z. (2007), 'Online Inference for Multiple Changepoint Problems', Journal of the Royal Statistical Society, Series B, 69, 589-605.

Febrero, M., Galeano, P. and González-Manteiga, W. (2008), 'Outlier detection in functional data by depth measures, with application to identify abnormal NO*x* levels', Environmetrics, 19, 331-345.

Ferraty, F. and Vieu, P. (2002), 'The functional nonparametric model and application to spectrometric data', Computational Statistics, 17, 545-564.

Ferraty, F. and Vieu, P. (2003), 'Curves discrimination: a nonparametric functional approach', Computational Statistics and Data Analysis, 44, 161-173.

Ferraty, F. and Vieu, P. (2006), Nonparametric functional data analysis: theory and practice. Springer, New York.

Fishman, G.S. (2001), 'Discrete-event simulation, modelling, programming, and analysis', Springer-Verlag, New York.

Fraiman, R. and Muniz, G. (2001), 'Trimmed means for functional data', Test, 10, 419-440.

Gallagher, M.A., Bauer, K.W. and Maybeck, P.S. (1996), 'Initial data truncation for univariate output of discrete-event simulations using the Kalman Filter', Management Science, 42.

Gebraeel, N., Lawley, M., Li, R. and Ryan, J. (2005), 'Residual-life distributions from component degradation signals: A Bayesian approach', IIE Transactions, 37, 543-557.

Hoad, K., Robinson, S. and Davies, R. (2010), 'Automating warm-up length estimation', Journal of the Operational Research Society, 61, 1389-1403.

Holly, W., Cook, R. and Crowe, C. (1989), 'Reconciliation of mass flow rate measurements in a chemical extraction plant', The Canadian Journal of Chemical Engineering, 67, 595-601.

Jiang, T., Chen, B., He, X. and Stuart, P. (2003), 'Application of steady-state detection method based on wavelet transform', Computers and Chemical Engineering, 27, 569-578.

Kalman, R.E. (1960), 'A new approach to linear filtering and prediction problems', Journal of Basic Engineering, 82, 35-45.

Kelton, W.D. and Law, A.M. (1983), 'A new approach for dealing with the startup problem in discrete event simulation', Naval Research Logistics Quarterly, 30, 641-658.

Kim, M., Yoon, S.H., Domanski, P.A. and Payne, W.V. (2008), 'Design of a steady-state detector for fault detection and diagnosis of a residential air conditioner', International Journal of Refrigeration, 31, 790-799.

Korbel, M., Bellec, S., Jiang, T. and Stuart, P. (2014), 'Steady state identification for on-line data reconcilation based on wavelet transform and filtering', Computers and Chemical Engineering, 63, 206-218.

Koshevoy, G. and Mosler, K. (1997), 'Zonoid Trimming for Multivariate Distributions', Annals of Statistics, 25, 1998-2017.

Lee, S.B. (1996), 'Fatigue failure of welded vertical members of a steel truss bridge', Engineering Failure Analysis, 3, 103-108.

Li, B. and Moor, D.B. (1999), 'A corrected normal approximation for the Student's *t* distribution', Computational Statistics and Data Analysis, 29, 213-216.

Liu, J.S. (1998), Monte Carlo strategies in scientific computing: springer.

Liu, J.S. and Chen, R. (1998), 'Sequential Monte Carlo Methods for Dynamic Systems', Journal of the American Statistical Association, 93, 1032-1044.

Liu, R. (1990), 'On a notion of data depth based on random simplices', Annals of Statistics, 18, 405-414.

López-Pintado, S. and Romo, J. (2009), 'On the concept of depth for functional data', Journal of the American Statistical Association, 104, 718-734.

López-Pintado, S. and Romo, J. (2011), 'A half-region depth for functional data', Computational Statistics and Data Analysis, 55, 1679-1695.

Lu, C.J. and Meeker, W.Q. (1993), 'Using degradation measures to estimate a time-to-failure distribution', Technometrics, 35, 161-174.

Macdonald, K. (2011), 'Fracture and fatigue of welded joints and structures', Woodhead Publishing, Oxford, UK.

Mahuli, S.K., Rhinehart, R.R. and Riggs, J.B. (1992), 'Experimental demonstration of nonlinear model-based in-line control of pH', Journal of Process Control, 2, 145-153.

Mhamdi, A., Geffers, W., Flehmig, F. and Marquardt, W. (2010), 'On-line optimization of MSF desalination plants', Desalination and Water Resources: Thermal Desalination Processes, 1, 136-162.

Narasimhan, S., Kao, C.S. and Mah, R. (1987), 'Detecting changes of steady states using the mathematical theory of evidence', AIChE Journal, 33, 1930-1932.

Narasimhan, S., Mah, R.S.H, Tamhane, A.C., Woodward, J.W. and Hale, J.C. (1986), 'A composite statistical test for detecting changes of steady states', AIChE Journal, 32, 1409-1418.

O'Hagan, A. (1994). Bayesian Inference, Vol. 2B of Kendall´s Advanced Theory of Statistics, Arnold, London.

Oja, H. (1983), 'Descriptive statistics for multivariate distributions', Statistics and Probability Letters, 1, 327-332.

Page, E.S. (1961), 'Cumulative sum charts', Technometrics, 3, 1-9.

Quesenberry, C.P. (1990), 'SPC variables Q-Charts for short runs', Quality Concepts' 90 Conference Proceedings, 223-236.

Quesenberry, C.P. (1991), 'SPC Q-Charts for start-up processes and short or long runs', Journal of Quality Technology, 23, 213-224.

Ramsay, J. O. and Silverman, B. W. (2005), Functional Data Analysis (2nd ed.), Springer Verlag, New York.

Robert, C.P. and Casella, G. (2004), 'Monte Carlo statistical methods', Springer.

Robinson, S. (2007), 'A statistical process control approach to selecting a warm-up period for a discrete-event simulation', European Journal of Operational Research, 176, 332-346.

Rossi, F. and Villa, N. (2006), 'Support vector machine for functional data classification', Neurocomputing, 69, 730-742.

Roux, G.A.C.L., Santoro, B.F., Sotelo, F.F., Teissier, M. and Joulia, X. (2008), 'Improving steady-state identification', 18th European Symposium on Computer Aided Process Engineering, 25, 459-464.

SAE International. (2013), 'Comprehensive life test for 12 V automotive storage batteries', SAE Standards J2801.

Savitzky, A. and Golay, M.J.E. (1964), 'Smoothing and differentiation of data by simplified least squares procedures', Analytical Chemistry, 36, 1627-1639.

Sguera, C., Galeano, P. and Lillo, R. (2014), 'Spatial depth-based classification for functional data', Test, 23, 725-750.

Sguera, C., Galeano, P. and Lillo, R. (2016), 'Functional outlier detection by a local depth with application to NO_x levels', Stochastic Environmental Research and Risk Assessment, 30, 1115-1130.

Shewhart, W.A. (1931), 'Economic control of quality of manufactured product', New York: D. Van Nostrand Company, pp. 501.

Si, W., Yang, Q. and Wu, X. (2016), 'A physical-statistical model of overload retardation for crack propagation and application in reliability estimation', IIE Transactions, 48, 347-358.

Sih, G.C. (1973), 'Handbook of stress-intensity factors', Institute of Fracture and Solid Mechanics, Lehigh University, Bethlehem, PA.

Son, J., Zhang, Y., Sankavaram, C. and Zhou, S. (2015), 'RUL Prediction for Individual Units Based on Condition Monitoring Signals With a Change Point', IEEE Transactions on Reliability, 64, 182-196.

Tsiamyrtzis, P. and Hawkins, D. M. (2005), 'A Bayesian scheme to detect changes in the mean of a short run process', Technometrics, 47, 446-456.

Tukey, J.W. (1975), 'Mathematics and the picturing of data', In Proceedings of the International Congress of Mathematicians, 2, 523-531.

Wei, X., Zhu, B. and Xu, W. (2009), 'Internal resistance identification in vehicle power lithium-ion battery and application in lifetime evaluation', Proceedings of International Conference on Measuring Technology and Mechatronics Automation, 388-392.

Womack, J.P., Jones, D.T. and Roos, D. (1990), 'The machine that changed the world', Harper Perennial Press, New York.

Woodward, P.W. and Naylor, J.C. (1993), 'An application to Bayesian methods in SPC', The Statistician, 42, 461-469.

Wu, J., Chen, Y., Zhou, S., and Li, X. (2015), 'Online Steady-state Detection for Process Control Using Multiple Change-point Models and Particle Filters', IEEE Transactions on Automation Science and Engineering, in-press.

Wu, J., Zhou, S. and Li, X. (2013), 'Acoustic emission monitoring for ultraonic cavitation based dispersion process', Journal of Manufacturing Science and Engineering, 135, 031015.

Zuo, Y. (2003), 'Projection based data depth functions and associated medians', Annals of Statistics, 31, 1460-1490.

Zuo, Y. and Serfling, R. (2000), 'General notions of statistical depth function', Annals of Statistics, 28, 461-482.