Copyright

by

Deyi Zhang

2017

The Dissertation Committee for Deyi Zhang Certifies that this is the approved version of the following dissertation:

UNCERTAINTY QUANTIFICATION AND ITS PROPERTIES FOR HIDDEN MARKOV MODELS WITH APPLICATION TO CONDITION BASED MAINTENANCE

Committee:

Dragan Djurdjanovic, Supervisor

J. Eric Bickel

Grani Hanasusanto

John Hasenbein

Stephen G. Walker

UNCERTAINTY QUANTIFICATION AND ITS PROPERTIES FOR HIDDEN MARKOV MODELS WITH APPLICATION TO CONDITION BASED MAINTENANCE

by

Deyi Zhang

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas at Austin

December 2017

Dedication

To my family.

Acknowledgements

I would like to show my deepest gratitude to the chair of the committee, Professor Dragan Djurdjanovic. I cannot thank enough for innumerous technical trainings from him, which expanded my versatility and built my prowess. Prof. Djudjanovic also taught me how to do research through harmonizing creation and elimination, simplicity and sophistication, and most importantly, meditation and execution. This dissertation cannot be shaped without his initiation in research topics, his generosity of providing the datasets, his assistance in the writing process, and his invaluable insights and guidance in developing the theoretical work.

I would also like to thank Dr. John Hasenbein for educating me in stochastic modeling and for many discussions with me that boosted maturity in this PhD research, to thank Dr. Eric Bickel for hosting the ORIE open house as a rewarding platform for exhibition of this research, and to thank Dr. Grani Hanasusanto for reviewing manuscript of the dissertation and provide feedbacks to improve the writing. I am also grateful for many pleasant communications with Prof. Stephen Walker that nourished my theoretical thinking. I would like further to acknowledge Dr. David Morton for his engaging and classy teaching of simulation methods that pillar the solutions from this research.

Then I want to thank all of my peers in the Limes lab and the students from ORIE program who have brought me changes in dimensions that I could ever have imagined. I am thankful to Michael Cholette, Alex Bleakie, Marcus Musselman, Merve Celen, and Yibo Jiao, the technical expertises of whom inspired me to overcome difficulties and excel in my own niche. I want to also thank Asad UI Haq, Keren Wang, Kent Zhang, Zicheng Cai, Yuyang Xie, and Mathew Graves for their reliable support in our teamwork and for being both professional and enthusiastic that created a productive lab

environment. Furthermore, Kaiwen Yang, Huidong Zhang, and Roberto Dailey kindly offered logistical conveniences in the period of finalizing this dissertation and I would like to acknowledge their help. Beyond that, I am grateful for my friends from ORIE program who has helped me, including Kun Zan, Jinho Lee, Zhufeng Gao, and Yufen Shao.

Finally and ultimately, I would like to thank my family, who supported me mentally, emotionally, and spiritually over the years, without ever claiming any of credits. Special thanks to my wife, Shihui Jia, who provided sublime consideration to compensate the turbulences that I met in both work and life during the entire course of seeking the PhD degree.

Uncertainty Quantification and Its Properties for Hidden Markov Models with Application to Condition Based Maintenance

Deyi Zhang, Ph.D.

The University of Texas at Austin, 2017

Supervisor: Dragan Djurdjanovic

Abstract

Condition-based maintenance (CBM) can be viewed as a transformation of data gathered from a piece of equipment into information about its condition, and further into decisions on what to do with the equipment. Hidden Markov model (HMM) is a useful framework to probabilistically model the condition of complex engineering systems with partial observability of the underlying states. Condition monitoring and prediction of such type of system requires accurate knowledge of HMM that describes the degradation of such a system with data collected from the sensors mounted on it, as well as understanding of the uncertainty of the HMMs identified from the available data.

To that end, this thesis proposes a novel HMM estimation scheme based on the principles of Bayes theorem. The newly proposed Bayesian estimation approach for estimating HMM parameters naturally yields information about model parametric uncertainties via posterior distributions of HMM parameters emanating from the estimation process. In addition, a novel condition monitoring scheme based on uncertain HMMs of the degradation process is proposed and demonstrated on a large dataset obtained from a semiconductor manufacturing facility. Portion of the data was used to build operating mode specific HMMs of machine degradation via the newly proposed Bayesian estimation process, while the remainder of the data was used for monitoring of machine condition using the uncertain degradation HMMs yielded by Bayesian estimation. Comparison with a traditional signature-based statistical monitoring method showed that the newly proposed approach effectively utilizes the fact that its parameters are uncertain themselves, leading to orders of magnitude fewer false alarms.

This methodology is further extended to address the practical issue that maintenance interventions are usually imperfect. We propose both a novel non-ergodic and non-homogeneous HMM that assumes imperfect maintenances and a novel process monitoring method capable of monitoring the hidden states considering model uncertainty. Significant improvement in both the log-likelihood of estimated HMM parameters and monitoring performance were observed, compared to those obtained using degradation HMMs that always assumed perfect maintenance.

Finally, behavior of the posterior distribution of parameters of unidirectional non-ergodic HMMs modeling in this thesis for degradation was theoretically analyzed in terms of their evolution as more data become available in the estimation process. The convergence problem is formulated as a Bernstein-von Mises theorem (BvMT), and under certain regularity conditions, the sequence of posterior distributions is proven to converge to a Gaussian distribution with variance matrix being the inverse of the Fisher information matrix. An example of a unidirectional HMM is presented for which the regularity conditions are verified, and illustrations of expected theoretical results are given using simulation. The understanding of such convergence of posterior distributions enables one to determine when Bayesian estimation of degradation HMMs is justified and converges toward true model parameters, as well as how much data one then needs to achieve desired accuracy of the resulting model. Understanding of these issues is of utmost important if HMMs are to be used for degradation modeling and monitoring.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background and Motivation

Rapid technological advances that occurred in the last couple of decades have resulted in increasingly complicated engineering systems that pose significant challenges in terms of their reliable and safe manufacturing, operation and use over their life cycles. The costs associated with the pursuit of these challenges are rapidly escalating. For example, in 1981, maintenance costs in the United States economy were estimated at $600 billion, a figure that doubled in the subsequent 20 years, with an estimated 30-50% of these costs wasted through ineffective maintenance and unexpected failures [1].

These costs are among the key driving factors towards research in condition-based maintenance (CBM), which can be viewed as a transformation of data gathered from a piece of equipment into information about its condition, and further into decisions on what to do with that equipment [2].

Different stages of this transformation are illustrated in Figure 1. Information about health of any piece of equipment is obtained from the readings of possibly multiple sensors mounted on that equipment. Often, situations exist where sensor readings are augmented with historical knowledge about equipment behavior, engineering model of phenomena occurring in the equipment, or human expertise. Based on these sources of information, features relevant to equipment health are extracted from sensor readings through various forms of sensory signal processing and feature extraction. These features form behavior models of equipment in different health states (normal behavior and different faulty behavior modes). Those models may be in various different forms, including a statistical form (distributions of sensory signatures under normal or various faulty conditions), dynamic model (differential equations describing various health states of the equipment), and others. Based on the models of normal and current equipment behavior, equipment health assessment can be accomplished by quantitatively expressing the proximity of the currently observed system behavior to the model describing its normal health state. This stage of CBM is also often referred to as the fault detection stage. Similarly, presence or absence of any fault can be diagnosed through proximity of the model of the currently observed equipment behavior to the behavior model corresponding to a specific fault. Finally, the temporal dynamics of signatures extracted from sensor readings can be captured and extrapolated to predict their behavior in the future and thus predict likelihoods of various behavior modes for the equipment.

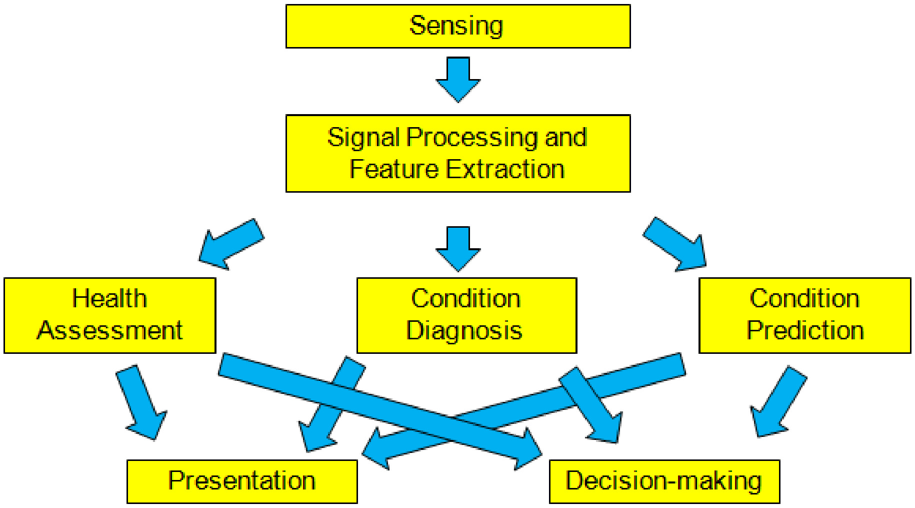


Figure 1: Concept of CBM as transformation of sensing data into information about equipment condition and further into maintenance and operational decisions.

Based on the quantitative information about current and/or predicted equipment health, maintenance and operational decisions that are optimal from the system level point of view can be made. In a manufacturing system, that entails maintenance and/or production decisions that optimize some operational metric of the system (productivity, quality, system availability etc.) [3], [4]. In practice this target decision point is defined by maximum profit or return on investment. Various aspects of this “data to information to decision” transformation have received significant attention, especially in the case of sophisticated, expensive and safety critical systems, such as manufacturing equipment, computer networks, automotive and aircraft engines, etc. A thorough survey of latest activities and achievements in CBM can be found in [5].

Traditional CBM approaches to extracting the information about the current or predicted condition of the monitored system rely on modeling the behavior of key signatures extracted from the available sensor readings and detecting or predicting when those signatures would exit areas in which they reside during normal system behavior. Such a paradigm implies an important assumption that sensory features directly indicate the condition of the monitored system. A problem occurs when such observables directly depicting system degradation cannot be ascertained. E.g, in a system whose condition is determined by a distributed phenomenon, such as plasma in various semiconductor manufacturing tools [6] or electrochemical field in a typical Li-ion battery [7], sensors provide information about the condition of a three dimensional field, but only at discrete points. Thus, even if more sensors are installed, a full picture about the state of the monitored system between the sensorized points can only be resolved using a highly detailed model of that field and its interactions with other surrounding subsystems. However, reliable and detailed multi-physics models of an entire plasma-based tool, such as a plasma-based etcher or plasma-based deposition tool, do not exist with sufficient fidelity or sufficient speed to be used for diagnostic purposes. In the case of batteries, highly detailed physics-based models became available only recently, but event these models are computationally so expensive that their application for monitoring and diagnostics purposes is still unfeasible [8].

The issue of needing to monitor systems with partial observability of their condition goes beyond just distributed systems. E.g., in cases when direct monitoring of a phenomenon requires sensors to operate in highly abrasive environments, such as in down-hole condition monitoring in oil and gas extraction, sensors are usually placed far away from the actual phenomenon and the relations between them and the actual condition they are monitoring are inherently indirect and nondeterministic. In addition, other complex systems, e.g. a diesel engine, may only have sensors for some critical variables, such as those related to performance and safety of the system. Unavailability of all of condition-related measurements results in partial characterization of the underlying condition of those systems.

Such inability to reliably and fully deduce the condition of a monitored system leads to situations in which two systems may exhibit very similar sensory signatures, but their conditions are sufficiently different that one may be operating normally, while the other one produces poor products or behaves abnormally. A “knee-jerk” approach to remedying such apparent lack of observability of the system condition is to add more sensors to it, but as mentioned before, such an approach may be futile in the case of a phenomenon for which a detailed and computationally tractable model does not exist.

Probabilistic modeling provides an alternative route to model the intuitive relations between the sensor information and machine condition. This can be done by assigning probabilities to obtain specific sensor signatures at different machine conditions, and by modeling the evolution of those conditions as a random process that is stochastically related to the sensor readings. Hidden Markov Model (HMM) [9] stands out as a useful probabilistic model that enables such approach to the CBM paradigm of complex engineering systems.

Identification of a HMM requires estimation of its parameters using the available sensory data. This estimation problem is commonly solved by finding the HMM parameters that maximize the likelihood of the sequence of observation used for identifying the HMM. The resulting Maximum Likelihood Estimator (MLE) has desirable asymptotic properties, including consistency [10] and asymptotic normality [11]. However the gradient-based search methods commonly used to obtain the MLE of HMM parameters, such as the Baum-Welch [9] or equivalently Expectation-Maximization (EM) [12] methods, do not guarantee the global optimality of the solution to this multi-modal optimization problem. Furthermore, MLE-based methods do not readily provide the information as to how close or how far the actual solution is from the results produced by estimation. On the other hand, understanding the uncertainty of HMM parameters is a very important point. E.g., it is crucial for understanding how far such models can be used for meaningful prediction of the condition of the monitored system and for subsequent maintenance decision making based on such models. Namely, model uncertainties accumulate as one attempts to predict probabilities of degradation states (hidden HMM states) further and further ahead and could quickly render those predictions useless.

In terms of anomaly detection, the traditional anomaly detection based on deterministic relation between sensor readings and condition of the system needs to take into account the stochasticity of the condition degradation model as well as the uncertainty in the estimation of that model. Involving risk analysis into those decisions by including considerations of uncertainties in the nominal model could lead to substantial savings of maintenance costs, since these refined decisions may induce significantly fewer false alarms than traditional signature based process control schemes [13] that neglect model uncertainty issues. Solutions to other decision-making problems, such as optimal scheduling of maintenance based on uncertain degradation models [14], should also benefit from the extra information of modeling uncertainty, though studies of these problems are outside of the scope of this PhD research.

Despite the obvious allure and applicability of HMM in CBM, understanding of the estimation uncertainty in HMM parameters for various form of HMMs is far from complete. For instance, consistency is a desirable property of any statistical estimation procedure [15], [16]. Consistency of HMM parameter estimation would imply the convergence of HMM estimators towards the true parameters that generated the observations (sensor readings) and the decrease of the estimation uncertainties towards zero. This property has not been established yet for the full variety of possible HMM forms. Most notably, it has not been shown yet for non-ergodic HMMs, which are inherently needed to describe degradation processes. On the other hand, understanding of the consistency will enable one to determine when the estimation procedure of HMM is justified and converges towards true model parameters, as well as how much data one needs to achieve the desired accuracy of the resulting estimates, if the estimation procedure indeed converges.

In this doctoral thesis, the aforementioned gaps will be addressed through a research on the problem of estimation of HMM parameters along with the understanding of parametric uncertainty of that estimation, as well as the use of such uncertain HMMs of system degradations for condition monitoring of complex engineering systems.

1.2 RESEARCH OBJECTIVES AND CHALLENGES

The main objective of this doctoral research is to develop a methodology for estimating parameters of non-homogeneous and/or non-ergodic HMM along with the parametric uncertainties, and to understand the statistical properties of the estimation procedure for HMM parameters.

The contribution of the work can be summarized as the following.

1. An HMM estimation scheme based on the Bayes theorem. This scheme provides a distribution of HMM parameters that depicts the parametric uncertainties in the HMM trained using a set of observation sequences.
2. Fault detection methods that can detect anomalous behavior based on an observation or an observation sequence using the uncertain HMM degradation models yielded by the novel estimation scheme.
3. A mathematical analysis and proof of asymptotic consistency and normality of Bayesian posterior distributions as the amount of observation sequences increases to infinity.

The challenges in achieving the contribution mentioned in this section involves the following. The first challenge lies in the development of a parameter estimation method that yields both point estimates and uncertainties of these estimates simultaneously for non-homogeneous and/or non-ergodic HMMs. Bayesian inference allows a formulation of such an estimation procedure, and sampling algorithm based on MCMC provides computational solution to the identification problem. The second challenge is to develop detection methods that take advantage of the newly available uncertainty information about HMM parameters to yield higher detection accuracy than traditional methods which ignore model uncertainty. The final challenge is to formulate and prove the asymptotic properties of the distributions of parameter estimates, which has not been addressed using infinite amount observation sequences when the assumption of ergodicity of HMM is absent.[1]

1.3 OUTLINE OF THE DISSERTATION

The rest of this doctoral dissertation is organized as follows. Chapter 2 presents a review of the literature on HMM estimation methods, HMM-based fault detection research, as well as on the explorations of existing asymptotic properties of Bayesian estimators of parameter for stationary and some non-stationary HMMs. Chapter 3 presents a novel Bayesian methodology for estimation of non-ergodic, non-homogeneous HMMs, as well as a new condition monitoring methodology based on degradation models described by HMMs identified using the aforementioned estimation procedure. The results from the application of the proposed methodologies on simulated datasets and real world datasets will be provided in the same chapter. In Chapter 4, a new type of HMM addressing the imperfect maintenance and a novel monitoring method that generates condition information based on such models are provided. Chapter 4 also offers results of applying the new HMM-based degradation modeling and monitoring methods that account for maintenance imperfections of a large fab data set, clearly indicating modeling and monitoring performance improvements over the corresponding methods that assume perfect maintenance operations. Chapter 5 gives the proof of asymptotic consistency and normality of posterior distribution for Bayesian estimation of HMMs under a set of commonly used regularity conditions but without assuming ergodicity of

[1] Successful solution to this problem would enable quantification of the trade-off between the amount of data used for parameter estimation and the uncertainty level of the estimates.

HMM. Finally, Chapter 6 details the achievements of this doctoral research and outline some suggested directions of future research.

Chapter 2: Review of Model Identification, Process Monitoring, and Estimation Properties for Hidden Markov Model

As we discussed in the previous chapter, condition-based maintenance (CBM) aims to facilitate maintenance operations based on the actual or predicted condition of the target system, as assessed from the available sensor readings. Key enabling factor for CBM is building of the model of the condition of the underlying system or process [1], [4], [19], [20]. First principle models of system degradation can be built where sufficient physics-based knowledge about systems exist. However these models are usually infeasible for degradation modeling of complex system for which such models are computationally too expensive to be determined or often cannot be determined at all [21]. Data-driven modeling is an alternative approach that utilizes sensory data to build various types of empirical models (statistical, dynamic, neural-network based) of the condition of the underlying system, enabling implementation of CBM for complex systems, such as diesel engine [22], biological system [23], semi-conductor manufacturing tools [23], etc. Though both physics-based and data-driven modeling approaches can complement each other in the so called hybrid models [24], recent availability of vast volumes of data in increasingly complicated manufacturing environment [26] provides unprecedented opportunities for data-driven models to play a more prominent role in CBM [27]-[29].

Recently, significant research attention has been dedicated to HMMs, which are essentially data driven models. As mentioned in Chapter 1, the proposed research concentrates on the use of HMMs for modeling and monitoring of degradation dynamics of complex engineering processes. Estimation of parameters of degradation HMM and the associated uncertainty are the prerequisites for the subsequent process monitoring of the target system. Therefore, it is necessary to review the existing literature on HMM identification and Section 2.1 provides an extensive review of the currently available identification methods for many types of hidden Markov models. Given the monitoring focus of the application side of this dissertation, Section 2.2 offers a review of HMM based monitoring research. Finally, given the interest in understanding the properties of the newly proposed method for Bayesian estimation of HMM parameters, Section 2.3 reviews theoretical work related to asymptotic behaviors of HMM parameter estimates.

2.1 HIDDEN MARKOV MODEL IDENTIFICATION

Parameter estimation in HMM is a difficult problem due to the hidden nature of the unobservable quantities and a large number of model parameters, which are needed for describing both the evolution of the underlying states and the probabilistic relationship between the states and the observables. This estimation problem is commonly solved by finding the HMM parameters that maximize the likelihood of the sequence of observations used for identifying the HMM. The resulting Maximum Likelihood Estimator (MLE) has desirable asymptotic properties, including consistency and asymptotic normality. However the gradient-based search methods commonly used to obtain the MLE of HMM parameters, such as the Baum-Welch [9] or Expectation-Maximization (EM) [11] methods, do not guarantee the global optimality of the solution to this multi-modal optimization problem. Despite this known issue of the EM algorithm, it has been accepted as the standard approach to identify HMM in a variety of disciplines [17] and many HMM applications in CBM, including monitoring of a gearbox [30], drilling tools [31], rotary machines [32], and bearings [33].

Despite the success of applying HMMs identified by EM to CBM of several systems, modern complex engineering processes pose further challenges that conventional EM-based identification of HMM parameters did not address. One challenge is quantifying uncertainty of estimation of HMM parameters and understanding when the estimation procedure converges. Another one is characterizing variability in degradation dynamics that could appear in complex engineering processes due to potentially strongly varying operating regimes.

Statistical confidence intervals (CI) [34] are commonly used to represent parameter uncertainty for probability models. A majority CI-based characterization of HMM parametric uncertainty that one can find in literature pertains to the MLE of HMM parameters. Based on the established asymptotic property of MLE [17] for HMM, approximated CI can be obtained by exact calculation of the variance-covariance matrix of the MLE [35], [36], or by approximation using numerical differentiation [38]. However, these methods could be justified to apply to more generalized HMM only if asymptotic normality of MLE for those HMMs can be established. Alternatively, bootstrapping [39] provides approximated CI via iterative resampling of observation symbols followed by reestimation of HMM parameters, which does not require asymptotic normality of MLE. However, it requires the run of a costly EM algorithm for each bootstrap iteration and leads to an overall very expensive computation, which may only be mitigated by using high performance computers [37]. Visser et al. [38] conducted benchmarking through simulation to compare several methods that produce CIs of HMM parameters and concluded that numerical differentiation provided too narrow CIs for covering the true parameters, whereas bootstrapping yield desired CIs that match with the actual confidence level. Zucchini and MacDonald [40] used bootstrapping for a wide range of natural science applications of HMM and observed the common computational issues of bootstrapping along with inaccurate estimates when true HMM parameters become degenerate. In sum, unless asymptotical theory can be established for a given HMM, bootstrapping is the standard choice for a likelihood-based approach to obtain CIs. However it suffers from very expensive computational cost, which limits its application to characterizing uncertainty of estimated HMM parameters, especially with large datasets frequently (usually) encountered in condition monitoring of complex systems.

Bayesian approach [41] distinctively integrates estimation of HMM parameters and derivation of uncertainty of those parameters. For the case of ergodic HMMs, Bayesian approach can always produce an estimate that strictly or asymptotically minimize estimation risk [42]. Such Bayesian estimates have desirable asymptotical properties as those produced by MLE for HMM [18] and are typically obtained via state-of-the-art Markov Chain Monte Carlo (MCMC) computation [43], whose convergence is guaranteed and can be robustly controlled [44]. Moreover, the output of MCMC can be immediately used to form a full (posterior) distribution of the estimated parameters to provide complete information about parameter uncertainty. From this distribution, Bayesian confidence intervals (or credible intervals) for the HMM parameters using its appropriate percentiles, yielding a result compatible with the CIs produced by MLE. Rydén [46] compared bootstrapping CI and Bayesian CI based on MCMC and concluded that both CIs have comparable performance, but bootstrapping was found to be much slower than MCMC. Martinez et al. [47] modeled epidemic in a population as hidden states of a HMM and use MCMC to identify parameter uncertainty of 2-state HMM with Gaussian emission density. They showed that Bayesian 95% CIs for two Gaussian means are non-overlapped which justifies their choices of using distinct Gaussian distributions. Chodera et al. [48] conducted a simulation study to show that Bayesian 95% CIs based on MCMC converge to the true parameters for a 3-state Gaussian HMM using samples of growing size for estimating HMM parameters. Nam et al. [49], [50] developed a novel sequential Monte Carlo (SMC) method to identify uncertain parameters of a HMM with multivariate normal distribution modeling emission and used the full distribution of HMM parameters to further characterize the uncertainty of change-points in the time series. Despite the tractability and recent innovative application of Bayesian posterior distributions of HMM parameters, the dynamics of the hidden states are all assumed to be regular (ergodic) and stationary. However, many of engineering processes concerned by CBM community that could be modeled by HMM exhibit irregular or non-stationary behavior, and therefore extensions of HMM need to be taken into account.

Engineering process data had primarily been considered as stationary in traditional CBM applications, either by assuming the data is independently and identically distributed (i.i.d) [13], [51] or by assuming a stationary time series model for the process data [52], [53]. However, modern engineering practices bring a plethora of sources that could generate non-stationary processes and process data. For example, system degradation occurs stochastically which could cause sudden abnormal measurements that, as was observed on automotive engines [54] or bearings [55] (these kinds of excursions typically happened during the start-up or shutdown of the machine [54]). Another source is the operator’s interaction with a machine that typically occurs in control of continuous processes, such as chemical process [56] or semiconductor manufacturing process [57]. In other words, process control especially through operator interventions can inherently introduce non-stationarities in the data. Furthermore, unpredictable operating environment such, as load variation, could directly result in dramatic fluctuation of signals, such as those from bit in oil drilling process [58] and from gearbox in ground excavation [59]. Although HMM can intuitively address the unobservability issue in the above mentioned applications, the various patterns of non-stationarity exhibited in these applications pose further challenges to process modeling and model identification using HMM. Therefore, adaptation of HMM is needed to deal with these sources of non-stationarity.

Most of the HMMs addressed in the literature discussed so far are stationary in a rigorous sense that the marginal distributions are time-invariant, because ergodicity conditions are imposed or presumed in this literature [17]. In contrast, a non-stationary HMM has been initially proposed by Sin and Kim [60] and applied in different areas, such as anomaly detection [61] and biology [62]. Motivated by a potential wide application domain of this model, Djuric and Chun [63] conducted a focused study to extend the MCMC designed by [43] for conventional HMM and estimate parameters (with uncertainty) of a non-stationary HMM proposed in [60]. However, this model is essentially a hidden semi-Markov model [64] under the alias of non-stationary HMM with several synonyms, as explained by Yu [65], and therefore it may be non-stationary, but it is strictly not a HMM. To overcome the ambiguity in this intuitive and useful term, we follow the definition of a non-stationary Markov chain [66] and define the non-stationary HMM (NSHMM) as a hidden Markov model (or equivalently, a HMM is non-stationary) where the underlying Markov chain does not have a stationary dynamics. The generality in HMM defined above can be shown to exist in many useful extensions of HMM. It may also facilitate design of new HMMs by correlating the non-stationarity in the data and the properties of NS-HMM.

When it comes to non-ergodic HMMs [9][2], they have been a popular choice to model machine degradation due to their applicability to both anomaly detection [55] and forecasting remaining useful life [67]. Identification of non-ergodic HMM, as pointed by

[2] Non-ergodic HMM is non-stationary under generic conditions (see section 2.3 for further discussion).

multiple studies [9], [17], [55], requires multiple observations sequences. To estimate HMM parameters in such situation one can simply modify re-estimation steps using multiple sequences as a segmented long sequence (see [9], section V-B), or one can derive an ensemble estimate of HMM parameters from separate estimates by applying EM on each sequence [68]. However, inability to obtain uncertainty information persists in these extensions of EM algorithm, and to the best of author's knowledge, no study has addressed the uncertainty of non-ergodic HMM parameter estimation (as one type of NS-HMM) using likelihood approach or Bayesian approach, although some peripheral studies exist. Bibbona and Ditlevsen [70] developed asymptotic theory of MLE for parameters of a non-ergodic diffusion process and use the theory to derive confidence intervals of the MLE. Jarsa et al. [70] developed both MCMC and SMC to obtain Bayesian posterior distribution of parameters of a non-ergodic Markov chain with missing observations. The methodology to characterize uncertainty in these two recent studies is consistent with what is discussed in this chapter.

Non-homogeneous HMMs have time-varying transition matrices for the underlying Markov chain as well as potentially time-varying observation symbols and emission matrices. They are a very flexible type of non-stationary HMM that has diverse applications in economics [72], [73], biomedical science [74], [75], and geoscience [76], [77]. Extraneous information (covariates) is typically utilized to drive the time-changing transition matrix, although it is possible to build seasonal variations within the transition matrix using Fourier series without using extra data [78]. To deal with additional parameters that describe non-homogeneous HMMs, extensions of EM [79] and MCMC [80] both exist for identifying their parameters. Cholette and Djurdjanovic [81] modified the EM algorithm to identify a non-homogeneous and non-ergodic HMM. They showed an improvement in accuracy to found by their modified EM, but did not address uncertainty of the MLE estimates. Bartolucci and Farcomeni [82] derived exact formula for information matrix needed for standard errors of MLE of non-homogenous HMM. They also showed that CIs using their formula have comparable performance as bootstrapping CI, but yield dramatic savings in computations for moderate-size dataset. However, they ignored establishing asymptotic normality of MLE for non-homogeneous HMM needed for justifying their CI, and it could result in undesirable performance of their CIs on larger datasets. On the other hand, Meligkotsidou and Dellaportas [83] developed a MCMC procedure that resolves both the parameter uncertainty and structural uncertainty (number of hidden states) for a non-homogeneous, but ergodic HMM. They also showed a significant improvement in predictive performance using non-homogenous HMM than conventional homogenous HMM.

Despite this strong recent research interest in methodologies enabling identification of non-homogenous HMM and non-ergodic along with parameter uncertainty that is crucial for CBM applications, the review of the available literature reveals a gap between the need to use non-stationary non-ergodic HMMs to model complex engineering processes and limitations of practical methods and theory for HMM parameter identification and quantification of the associated uncertainties. The goal of the method pursued in this doctoral research is to address this gap and provide uncertainty estimation for a non-homogeneous and non-ergodic HMM parameter estimation. Specific details are given in Chapter 3.

2.2 FAULT DIAGNOSIS WITH HIDDEN MARKOV MODEL

Once an anomaly is detected, root causes of such anomalous behavior are identified via the process of fault diagnosis (FD)[3] in which the information contained in the features extracted from sensor readings is mapped to the space of machine faults[4]. The most widely used HMM-based FD approach matches HMM observation sequence indicating the actual system performance with one of the known faulty behavior models. Such matching is a selection of the HMM, among a collection of HMMs modeling different fault modes, that yields the largest likelihood of the given observation sequence. Despite the fact that multiple HMMs have to be trained, this method allows very efficient on-line diagnosis that only requires inexpensive calculation of the likelihoods of the observation sequence based on each HMM (known as the evaluation problem in [9]). Hence, it has been extensively applied to monitoring of dynamic systems, including rotating machinery (rotor [32], [89], drilling machine [90], [91], cutting tool [92], stamping machine [93], rotating shaft [94], rotating rig [55]), chemical plants (tank reactor [95]-[97], melting furnace [98], [99], multi-system plant [100], [101]), nuclear plant [102], as well as electronic system [103]. However, the major issue of this approach lies at its fault space that only consists of a fixed and often small number of faults known at the period of training [5]. Depending on the complexity of the monitored system, such fault space is an over-simplification that limits its diagnostic coverage of all possible abnormal behaviors during system operation.

[3] According to [84], "a fault is an unpermitted deviation of at least one characteristic property (feature) of the system from the acceptable, usual, standard condition." Degradation in this dissertation is defined as a type of constantly happening (if no intervention of operation) and accumulative fault that causes gradual deterioration of system performance. On the other hand, fault mode (or fault pattern, or simply faults if the context is clear) is a type of distinguishable fault whose cause can be clearly specified.

[4] Since there is a lack of consensus about the definition of FD ([27], [84]-[88]), the one that we adopt here is based on [5], which is widely accepted in CBM community.

[5] The number is usually less than 10 in the reviewed literature.

In previous approach, the most plausible HMM is chosen for explaining the actual data out of multiple HMMs. One can alternatively use a single HMM to model the system behavior traversing the faults, with the actual condition being modeled using the hidden states, and the most likely trajectory of the faults being determined by the most likely trajectory of hidden states. This approach allows physical interpretation of the hidden states [29], [104] as fault modes, and uncertainty statement of diagnosis [105], i.e., conditional probabilities of the occurrence of each known fault given the current observation sequence [6]. However, it has been difficult to estimate the transition probability matrix (TPM) without an appropriate training set where the true system (hidden) state changes from one fault to another. Bunks et al. [29] encountered this difficulty for FD on a helicopter gearbox and suggested that physics-based model of the gearbox could be used to obtain the TPM. Smyth [105] overcame this difficulty in monitoring of an antenna by introducing prior reliability knowledge of that system to estimate transition probabilities between two known faults and the normal condition.

A major advantage of the single-HMM approach over the multiple-HMMs approach is that the former offers an elegant framework to expand the fault space, by adding novel faults detected in the test period to the known faults in the training period. This is because the single HMM is used as a generative model, which is well known to be suited for on-line adaptation [106], [107]. Smyth [107] (following his earlier work [105]) introduced an unknown fault in the test data and showed that a flexible-state HMM can detect the existence of such new fault, while a fixed-state HMM constantly misclassified it as one of the known faults. Recently, Lee et al. [109] developed a scheme to update the HMM states and the associated HMM parameters as long as a novel fault is detected on-

[6] This method is similar to particle filtering method [106] for state space models where the underlying states are continuous.

line. This method was shown to outperform a fixed-state HMM significantly in detection rate. Unfortunately, these extensions of the single-HMM approach still assume that the fault space is discrete and finite, and therefore the issue of limited coverage of all abnormalities mentioned earlier cannot be resolved using neither the multiple-HMMs nor the single-HMM approach. Thus, even if an appropriate training data are available [104], [55] and HMM parameters can be estimated purely using the training data (as it should), e.g., using the Baum-Welch algorithm, the faults remain fixed and finite for the studies reviewed so far.

Modern engineering systems are highly complex that often consist of interacting dynamic systems. For such applications, the traditional approach for realizing HMM based diagnostic functionality[7] becomes excessively cumbersome because of the need to train the condition monitoring processes to recognize a large number of faults or faults of various severities, some of which often cannot be anticipated in advance. Even for the cases one is able to anticipate in advance, many faults manifest themselves very differently under different control inputs and environmental conditions, which makes training of diagnostic units for all possible conditions and all possible faults infeasible. Finally, such systems consist of numerous subsystems, each of which could contain significant non-linearities, with multiple control and environmental inputs, as well as inputs from other subsystems. This situation permits anomalies in one system to cascade and incite anomalous behavior of other systems connected to it, which effectively masks the real source of the anomaly.

Considering the contiguous variability of faults and multiple (or even convoluted) sources for the faults, the fault space for complex system should be continuous and multi-

[7] Diagnostics based on fault-specific HMMs.

dimensional. Fault Detection and Isolation (FDI) enables fault diagnosis under this realistic assumption and has been recently used for monitoring based on physics-based models [110]. It is a relatively new concept based on data-driven models for modeling dynamics of complex systems [111], [112] and its implementation using HMM has been initiated only recently [113]. FDI is a two-step procedure including fault detection[8] stage, i.e., deciding whether a fault has happened, followed by fault isolation stage in which the source of the fault is localized.

Anomaly detection with HMM can use limit checking based on probabilistic residuals that measure the discrepancy between new HMM observations and the nominal HMM trained under normal condition. Fox et al. [113] and Brown et al. [114] demonstrated the efficacy of this approach to detect faults when HMM is used for modeling a robotic system and an electric power plant, respectively. More recently, Cholette and Djurdjanovic [81] used this approach to monitor a semiconductor manufacturing tool whose condition was modeled by regime specific degradation HMMs.

A more robust approach based on hypothesis testing [117] allows detection of changes in the behavior model itself, i.e., the changes in the parameters of the HMM over time [9]. LeGland and Mevel [119] developed a non-local and a local Generalized Likelihood Ratio Test (GLRT) to decide whether the TPM shifts from it nominal value given each new observation sequence of the same length. While the non-local test has a desirable theoretical property that both false alarm and missed detection rates converge to

[8] Again, due the ambiguity of this term as that in fault diagnosis, we equate fault detection as anomaly detection [115] or novelty detection [116]. Following the literatures, Fault Diagnosis (FD) and anomaly detection (AD) terms are used interchangeably here.

[9] There is another approach by defining the residual as the goodness of fit given the data and the HMM [118]. However, it has majorly been used for model selection rather than fault detection. In model section, the same sample is fitted to multiple models. In fault detection, samples of possibly different sizes are fitted to the same model, and therefore the goodness of fit cannot be used if it cannot be compared between samples of different sizes. On the other hand, the likelihood slope can be viewed as a rate, which is free of the different size issue.

zero upon infinite observations, the local test is much simpler for implementation. Chen and Willett [120] developed a Sequential Likelihood Ratio Test (SLRT) to detect transient signals by detecting a change from one HMM (modeling the normal signal) to another HMM (modeling the transient signal). They demonstrated in simulation the superiority of this method over the conventional SLRT based on i.i.d models using various detection performance measures, including a receiver operating characteristic (ROC) curve. However, their experiments assumed that the HMM that models the transient signal is known, and therefore this method is not amenable to detect unanticipated dynamics, which, as mentioned before, is of paramount importance for fault detection in complex systems. Despite the fact that the methods mentioned above can robustly recognize the existence of anomalies using the concept of HMMs, the uncertainty of HMMs themselves has not been addressed in any of the fault detection method reviewed so far.

Section 3.3 addresses this gap by introducing an anomaly detection method that is aware of the uncertainty of the degradation HMM. It accomplishes this task by explicitly incorporating the distribution of estimated degradation HMM parameters obtained from the identification procedure discussed in Section 3.2.

2.3 CONVERGENCE PROPERTIES OF HMM PARAMETER ESTIMATION

With increasing amount of data generated from a true HMM, it is desired to see the Bayesian posterior distributions for HMM parameters increasingly concentrating[10] around the true HMM parameters. Bernstein-von Mises Theorem (BvMT) formulated in [123] (or equivalently asymptotic normality of posterior distributions [124]) proves such

[10] Studies exist for the concentration and its rate of posterior distribution for *nonparametric* HMMs, such as [125], whereas we focus on parametric HMMs.

convergence, under very general assumption on the prior distribution and the underlying HMM, demonstrating that the sequence of posterior distributions of HMM parameters tends toward a sequence of multi-variable Gaussian distributions with their centers converging to the true HMM parameters and with the covariance matrices converging to a zero matrix.

Consistency and asymptotic normality of MLEs are typical essential properties for establishing a BvMT, and these properties have been well-studied for stationary HMMs, For HMM with discrete observations, Baum and Petrie [125] and Petrie [127] proved the consistency and asymptotic normality for MLE. For HMMs with general emission distributions, Leroux [10] proved the consistency, and Bickel et al. [11] proved the asymptotic normality of MLE of HMM parameters. Rydén [128] proved both consistency and asymptotic normality of the MLE of HMM paramaters by imposing the assumptions that the sequence of observations used for estimation can be segmented into sequences of equal length. Finally, for HMM with generalizable state space, i.e., HMM whose hidden states could be discrete or continuous, under a measure-theoretic framework, Cappe et al. [17] proved both consistency and asymptotic normality of MLE. Based on the previous properties of MLE, Gunst and Shcherbakova [18] recently proved the BvMT for stationary HMM. Nevertheless, the stationarity in HMM has been a critical assumption that exists in the aforementioned work[11]. Such stationarity assumption also requires that HMM starts with a predetermined initial distribution of the hidden states.

For asymptotically stationary HMMs that have arbitrary initial distributions and converge to their stationary distributions, properties of MLE have also been explored and understood, but the BvMT has not been proven yet. More specifically, LeGland and

[11] It enables their necessary technical treatments, e.g., approximation based on treating a finite observation sequence as a sequence with infinite past observations.

Mevel [129] proved the consistency and asymptotic normality of MLE [130] for such type of HMM. Furthermore, Mevel and Finesso [131] proved the convergence rate of MLE, and they then extended many known properties about MLE to the case when HMM is even misspecified [132]. However, the above results are still limited in that both stationary HMM and asymptotically stationary HMM require ergodicity of the transition probability matrix, meaning that a stationary distribution of the hidden states exist and is unique.

As we mentioned in Section 2.1, many HMMs of practical interest are non-stationary, and the studies on properties of MLE for non-stationary HMMs are rare. Because description of non-stationary dynamics in HMM usually requires additional HMM parameters or relaxation in the domain of parameters for stationary HMM, accommodation is needed in the formulation of asymptotic properties of both MLE and posterior distribution for non-stationary HMM parameters. For a type of non-ergodic HMM (known as partially HMM in [133] [12]), Bordes and Vandekerkhove [133] provided a proof of consistency and asymptotical normality of MLE when assuming the data is generated from multiple observation sequences and the number of sequences of observations tends to infinity. Ailliot and Pene [134] recently proved the consistency of MLE for a non-homogeneous HMM under several regularity conditions on the covariate process that influences the time-varying transition probability matrix in this non-homogeneous HMM. Such work was extended in [132] to address the asymptotic properties of MLE for mis-specified non-homogeneous HMMs.

Considering the lack of understanding of limiting properties of Bayesian posterior distribution for non-stationary HMM parameters, we provide a BvMT formulated for

[12] Such non-ergodic HMM contains an absorbing state that can correspond to a failure state in degradation modeling.

unidirectional non-ergodic HMM and the proof of that theorem is enclosed in Chapter 5 of this doctoral dissertation.

Chapter 3: Identification for Hidden Markov Model and Monitoring Methodology using HMM [13]

3.1 Hidden Markov Model

Hidden Markov model is a doubly embedded stochastic process $\{X_t, Y_t\}_{t=0}^{\infty}$ in whose foundation is an unobservable Markov chain X_t, which drives an observable process Y_t for which at each time t, the observable variable Y_t is probabilistically related to the hidden state X_t. Assuming that the set of possible states for the hidden process X_t is $S = \{s_1, s_2, \dots, s_N\}$ and the set of possible observable symbols is $O = \{o_1, o_2, \dots, o_M\}$, the HMM can be described by a parameter triplet [14] $\boldsymbol{\theta} = (\boldsymbol{\nu}, \mathbf{P}, \mathbf{Q})$, composed of the initial state distribution $\boldsymbol{\nu} \in [0,1]^N$, state transition probability matrix $\mathbf{P} \in [0,1]^{N \times N}$ and emission probability matrix $\mathbf{Q} \in [0,1]^{M \times N}$.

In many applications, physics of the process modeled using the HMM can lead to specific patterns in the state transition matrix. For example, if the states $S = \{1,2,3\}$ represent condition of a monitored system, with state 1 denoting excellent condition, state 2 denoting OK condition and state 3 representing the bad condition, the state transition matrix P is constrained to be an upper triangular matrix, or $p_{ij} = 0, \forall i > j$, since without a maintenance operation, degradation state of the system can only deteriorate. Such "left-

[13] This chapter is based on [152]: Deyi Zhang, Andrew D. Bailey III, and Dragan Djurdjanovic, "Bayesian identification of hidden Markov models and their use for condition-based monitoring," *IEEE Transaction. on Reliability*, vol. 65, no. 3, pp. 1471-1482, June 2016. The contribution to this collaborated work from the first author includes: (1) proposing the identification method and the monitoring method; and (2) conducting data analysis on simulated datasets and a dataset collected from a PECVD tool by the third author along with his students.

[14] Note that this elaboration focuses on the HMMs with discrete states and observations, though other type of states and observation symbols can be considered, e.g. continuous state-dependent emission distributions, such as Gaussian distributions, can be conceptualized and parameterized, leading to a vector of state dependent means and variances substituting the emission matrix Q in the parameter triplet $\boldsymbol{\theta}$.

to-right" HMM structure [9] will indeed be utilized for degradation modeling in this chapter.

Recently, the standard HMM construct described above has been extended to regime-specific HMMs by incorporating time-varying dynamics and observation models, in order to account for variability in degradation models due to potentially variable operating regimes of the monitored system [81]. Suppose the operating regimes over time are denoted by a sequence $z_t, t = 0,1,2, \dots$, with each z_t having a known value from the set of possible operating regimes $R = \{r_1, r_2, \dots, r_L\}$. For each regime r, let us allow different HMM dynamics and observation probabilities by introducing a regime-specific HMM concept, which, assuming N hidden states $\{s_1, s_2, \dots, s_N\}$, can be described by parameters

$$\boldsymbol{\theta}^{(R)} = \left(\boldsymbol{\nu}, \mathbf{P}^{(r_1)}, \mathbf{Q}^{(r_1)}, \mathbf{P}^{(r_2)}, \mathbf{Q}^{(r_2)}, \dots, \mathbf{P}^{(r_L)}, \mathbf{Q}^{(r_L)}\right),$$

with initial state probability vector

$$\boldsymbol{\nu} = [\nu_1 \quad \nu_2 \quad \cdots \quad \nu_N]^T; \nu_i = \Pr(X_0 = s_i)\,, i = 1,2, \dots, N$$

regime-specific state transition matrices $\mathbf{P}^{(r)}, r \in \{r_1, r_2, \dots, r_L\}$ where

$$\mathbf{P}^{(r)} = \left[p_{i,j}^{(r)}\right]_{i,j=1,2,\dots,N}, \; p_{i,j}^{(r)} = \Pr\left(X_{t+1} = s_j \middle| X_t = s_i\right) \text{if } z_t = r$$

regime-specific emission probability matrices $\boldsymbol{Q}^{(r)}, r \in \{r_1, r_2, \dots, r_L\}$ satisfying

$$\mathbf{Q}^{(r)} = \left[q_{i,j}^{(r)}\right]_{\substack{i=1,2,\dots,N \\ j=1,2,\dots,M}}, q_{i,j}^{(r)} = \Pr\left(Y_t = o_j \middle| X_t = s_i\right) \text{if } z_t = r$$

and the hidden states process X_t progressing according to probabilities[15]

$$\begin{bmatrix} \Pr(X_t = s_1) \\ \Pr(X_t = s_2) \\ \vdots \\ \Pr(X_t = s_N) \end{bmatrix} = \boldsymbol{\nu}\left(\prod_{i=0}^{t} \mathbf{P}^{(z_i)}\right) \qquad (1)$$

One should note that in the context of regime-specific HMMs being used for degradation modeling, Eq. (1) formalizes the well-known notion of continuity of degradation, stipulating that the last state of degradation after one operating regime becomes the initial state of degradation for the next operating regime.

The HMM parameters $\boldsymbol{\theta}$ need to be identified from the available sensor readings (realizations of observable variables), which is one of the classical HMM problems – model identification. In the next section, we will introduce a novel Bayesian estimation based approach to identification of parameters of regime-specific HMMs.

3.2 HMM Identification Problem and Bayesian Estimation

Following the traditional approaches to identification of HMM parameters, such as the well-known Baum-Welch procedure [9] one can pursue a likelihood based estimation of parameters for regime-specific HMMs, seeking model parameters for which the sequence of observations based on which the model is identified is the most likely. More formally, given a sequence of observables $\boldsymbol{y}_t = (y_1, y_2, \dots, y_t)$ and relevant operating regimes $\boldsymbol{z}_t = (z_1, z_2, \dots, z_t)$, with each $z_i \in R = \{1,2, \dots, r\}$ being known at any

[15] These probabilities assert the probability of degradation states in future time t assuming current time is the starting time. If past observations are available, the probability of current state can be estimated by forward algorithm [9] and can be similarly used for assessing predicative degradation condition of the modeled system.

given time $i = 1,2, \dots t$, Maximum Likelihood Estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{t}}^{[\boldsymbol{MLE}]}$ of parameters $\boldsymbol{\theta}^{(R)}$ of regime-specific HMMs can be pursued by solving the following optimization problem

$$\widehat{\boldsymbol{\theta}}_t^{[MLE]} = \max_{\boldsymbol{\theta}^{(R)} \in \Omega^{(R)}} \Pr(\boldsymbol{y}_t | \boldsymbol{\theta}^{(R)}), \tag{2}$$

where

$$\Omega^{(R)} = \left\{ \boldsymbol{\theta}^{(R)} \in \mathbb{R}_{+}^{N+L\cdot(N^2+N\cdot M)} : \sum_{i=1}^{N} \nu_i = 1, \sum_{j=1}^{N} p_{i,j}^{(r)} = 1, \sum_{j=1}^{M} q_{i,j}^{(r)} = 1, \forall 1 \leq i \leq N \right\}.$$

The multi-modal nature of the likelihood function $\Pr\left(\boldsymbol{y}_t \middle| \boldsymbol{\theta}^{(\boldsymbol{R})}\right)$ in (2) poses substantial difficulty in solving the optimization problem [9]. Gradient-based methods, such as the Baum-Welch procedure, potentially get trapped in local optima, which is a well-known inherent drawback of gradient-based searches. Recently, Cholette and Djurdjanovic addressed this concern by using a genetic algorithm to modify initial points for the gradient-based optimization of the problem [81]. Unfortunately, although this modified algorithm greatly improves the likelihood over the purely gradient-based search, optimality of the resulting solution still cannot be guaranteed. In addition, gradient-based methods (or metaheuristically modified gradient based methods, such as the one in [81]) do not give information about the distribution of estimation errors (distribution of how close or far the estimate is from the true model parameters $\boldsymbol{\theta}_{\boldsymbol{0}}^{(\boldsymbol{R})}$). Understanding this uncertainty of the solution to (2) is highly important if, for example, such a solution is to be used for condition monitoring and prediction for a system whose degradation is modeled using regime specific HMMs. Model uncertainty accumulates rapidly as one tries to predict system condition further and further ahead using an uncertain degradation model. Hence, pursuit of a methodology to estimate parameters of the regime-specific HMMs, along with the information how near or far that estimate is from the true model parameters, is a highly useful and impacting goal that we will pursue now.

In the foundation of the Bayesian estimation is the well-known Bayes theorem, which transforms and updates whatever prior knowledge about the underlying random variable or process, using observations obtained from that variable or process. More formally, let $\Omega^{(R)}$ denote the domain of possible values of model parameters $\boldsymbol{\theta}^{(\boldsymbol{R})}$. If we represent whatever prior knowledge about model parameters we have using a prior distribution $\pi(\cdot)$ defined on $\Omega^{(R)}$, the observation sequence Y_t emitted by the regime specific HMM and their corresponding regimes Z_t can be incorporated to update the information about model parameters using posterior distribution

$$\pi_{\mathrm{t}}\big(\boldsymbol{\theta}^{(R)}\big|\boldsymbol{y}_t\big) = \frac{\Pr(y_t|\boldsymbol{\theta}^{(R)})\pi(\boldsymbol{\theta}^{(R)})}{\int_{\Omega^{(R)}}\Pr(y_t|\boldsymbol{\theta}^{(R)})\pi(\boldsymbol{\theta}^{(R)})d\boldsymbol{\theta}^{(R)}}. \qquad (3)$$

In the Bayesian framework, $\pi_{\mathrm{t}}\big(\boldsymbol{\theta}^{(R)}\big|\boldsymbol{y}_t\big)$ can be maximized similarly as $\Pr(y_t|\boldsymbol{\theta}^{(R)})$ in leading to the so called Maximum A Posteriori (MAP) estimate [136]. However MAP has the same issue of multimodality of the objective function as MLE [136]. Alternatively, the Bayesian estimate (BE) $\widetilde{\boldsymbol{\theta}}_t^{(R)}$ based on statistical decision theory [42] can be obtained by solving the following optimization problem

$$\widetilde{\boldsymbol{\theta}}_t^{(R)} := \mathrm{argmin}_{\widetilde{\boldsymbol{\theta}}^{(R)}\in\Omega^{(R)}} \int_{\Omega^{(R)}} \mathcal{L}\big(\widetilde{\boldsymbol{\theta}}^{(R)},\boldsymbol{\theta}^{(R)}\big)\pi_t\big(\boldsymbol{\theta}^{(R)}\big|\boldsymbol{y}_t\big)d\boldsymbol{\theta}^{(R)}, \qquad (4)$$

where $\mathcal{L}: \Omega^{(R)} \times \Omega^{(R)} \to \mathbb{R}$ is a ‘loss function’ that grows with the distance away from the true model parameters $\boldsymbol{\theta}_0^{(R)}$. For a commonly used quadratic loss[16] $\mathcal{L}\big(\widetilde{\boldsymbol{\theta}}^{(R)},\boldsymbol{\theta}^{(R)}\big) =$

[16] Quadratic loss is more mathematically tractable. For other convex loss function such as 0-1 loss, explicit solution to (4) is in general not available. However (4) can be solved numerically using stochastic optimization techniques such as sample average approximation (4) that requires a sample from Π_t, which can be obtained by the Gibbs sampler described in this section.

$\sum_{\tilde{\theta}\in\tilde{\boldsymbol{\theta}}^{(R)},\theta\in\boldsymbol{\theta}^{(R)}}\left|\tilde{\theta}-\theta\right|^{2}$, it can be shown immediately that solution to (4) is in the form of the posterior mean [42]

$$\tilde{\boldsymbol{\theta}}_{t}^{(R)} = \int_{\Omega^{(R)}} \boldsymbol{\theta}^{(R)}\, \pi_t\left(\boldsymbol{\theta}^{(R)}\middle|\boldsymbol{y}_t\right) d\boldsymbol{\theta}^{(R)},$$

or in terms of parameters of regime specific HMMs studied in this chapter,

$$\begin{aligned}
\tilde{v}_{i,t} &= \int_{\Omega^{(R)}} v_i\, \pi_t\left(\boldsymbol{\theta}^{(R)}\middle|\boldsymbol{y}_t\right) d\boldsymbol{\theta}^{(R)}, \forall 1 \le i \le N,\\
\tilde{p}_{i,j,t}^{(r)} &= \int_{\Omega^{(R)}} p_{i,j}^{r}\, \pi_t\left(\boldsymbol{\theta}^{(R)}\middle|\boldsymbol{y}_t\right) d\boldsymbol{\theta}^{(R)}, \forall 1 \le i,j \le N, r \in R,\\
\tilde{q}_{i,j,t}^{(r)} &= \int_{\Omega^{(R)}} q_{i,j}^{r}\, \pi_t\left(\boldsymbol{\theta}^{(R)}\middle|\boldsymbol{y}_t\right) d\boldsymbol{\theta}^{(R)}, \forall 1 \le i \le N, 1 \le j \le M, r \in R.
\end{aligned} \tag{5}$$

Considering the difficulty of numerically evaluating the high-dimensional integrals (5), we adapt a Gibbs sampling procedure from [137] to produce a random sample $\left\{\boldsymbol{\theta}_t^{(R),k}\right\}_{k=1}^{\delta}$ from the distribution $\pi_t\left(\boldsymbol{\theta}^{(R)}\middle|\boldsymbol{y}_t\right)$. This sample can be used not only for approximating the solution to (4), but also for obtaining credible intervals (a measure of uncertainty) for the model parameters via some approximation of the distribution $\pi_t\left(\boldsymbol{\theta}^{(R)}\middle|\boldsymbol{y}_t\right)$ from the sample $\left\{\boldsymbol{\theta}_t^{(R),k}\right\}_{k=1}^{\delta}$. The following assumptions are made to initiate the sampling procedure.

1. Following [17], it is assumed that $\boldsymbol{\Theta}^{(R)}$ has independent prior distributions in each of its component
$$\pi\left(\boldsymbol{\theta}^{(R)}\right) = \pi(\boldsymbol{\nu}) \prod_{r\in R} \pi\left(\boldsymbol{P}^{(r)}\right)\pi\left(\boldsymbol{Q}^{(r)}\right).$$
2. Following [17], initial probabilities, along with transition probabilities and emission probabilities conditioned at each state at all regimes are independently Dirichlet distributed[17] as

[17] The setting of Dirichlet distribution as priors allows derivation of closed form (see Lemma 13.1.6 on [17]) of the conditional distributions $\pi\left(v\middle|\boldsymbol{x}_t^{k-1},\boldsymbol{y}_t\right)$, $\pi\left(P^{(r)}\middle|\boldsymbol{x}_t^{k-1},\boldsymbol{y}_t\right)$ and $\pi\left(Q^{(r)}\middle|\boldsymbol{x}_t^{k-1},\boldsymbol{y}_t\right)$ needed in the Gibbs sampling (Fig. 2), and therefore the HMM parameters can be efficiently sampled. If the Dirichlet distribution is not flexible enough to represent prior information, hierarchical Dirichlet prior [139] may be used instead at the expense of more computation.

$$\mathcal{V} \sim Dirichlet(\alpha_{\mathcal{V}}),$$
$$\mathcal{P}_{i,\cdot}^{(r)} \sim Dirichlet(\alpha_{\mathcal{P}}), 1 \leq i \leq N, r \in R,$$
$$\mathcal{Q}_{i,\cdot}^{(r)} \sim Dirichlet(\alpha_{\mathcal{Q}}), 1 \leq i \leq N, r \in R,$$

where $\alpha_{\mathcal{V}}, \alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}$ are the corresponding concentration parameters. Following [137], for a non-informative case, all elements of $\alpha_{\mathcal{V}}, \alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}$ can be set to 0.5[18] whereas other values these distribution parameters can reflect prior knowledge of the model parameters, which may accelerate learning of the model.

3. The first sample point $\boldsymbol{\theta}_t^{(R),1}$ can be directly obtained from the prior distribution of the model parameters. The Gibbs sampler then successively obtains more samples of the model parameters using the estimate of hidden states by conjugate Bayesian computation [41] and updates of the hidden states given the estimate of model parameters by Forward-Backward procedure [17].
4. After δ samples are obtained from this procedure, point estimate of model parameters can be obtained as the sample mean, i.e. $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{t}}^{(R),\delta} \coloneqq \frac{1}{\delta}\sum_{k=1}^{\delta} \boldsymbol{\theta}_t^{(R),k}$**.** In addition, credible intervals of those parameters can be obtained from some approximation of the distribution $\pi_t\left(\boldsymbol{\theta}^{(R)} \middle| \boldsymbol{y}_t\right)$ from the sample points $\left\{\boldsymbol{\theta}_t^{(R),k}\right\}_{k=1}^{\delta}$ (or just by simply obtaining appropriate empirical quantiles from the sample). Decision on how many samples are needed (how big should δ be) can be made ad hoc, which is what we did in this chapter, or perhaps more formally, using the width of the confidence interval on the posterior mean, following the central limit theorem reported in [43] (procedure can be terminated when the width of the confidence interval falls sufficiently).

[18] These concentration parameters could be estimated using likelihood approach [139] if detailed prior information is available.

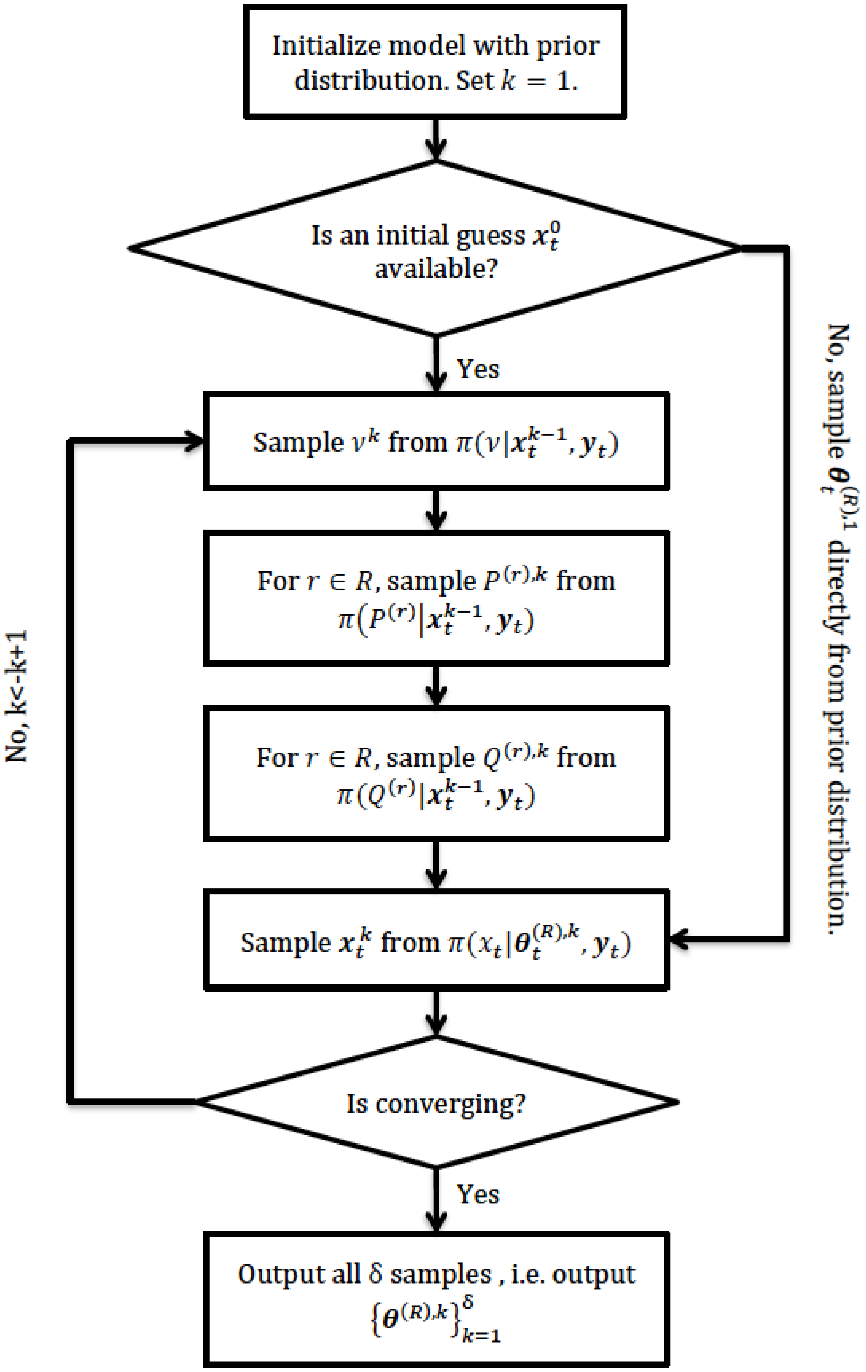

Initialize model with prior distribution. Set $k = 1$.
Is an initial guess $\boldsymbol{x}_t^0$ available?
Yes
No, sample $\boldsymbol{\theta}_t^{(R),1}$ directly from prior distribution.
Sample ν^k from $\pi(\nu|\boldsymbol{x}_t^{k-1}, \boldsymbol{y}_t)$
For $r \in R$, sample $P^{(r),k}$ from $\pi(P^{(r)}|\boldsymbol{x}_t^{k-1}, \boldsymbol{y}_t)$
For $r \in R$, sample $Q^{(r),k}$ from $\pi(Q^{(r)}|\boldsymbol{x}_t^{k-1}, \boldsymbol{y}_t)$
Sample $\boldsymbol{x}_t^k$ from $\pi(x_t|\boldsymbol{\theta}_t^{(R),k}, \boldsymbol{y}_t)$
Is converging?
No, k<-k+1
Yes
Output all δ samples , i.e. output $\{\boldsymbol{\theta}^{(R),k}\}_{k=1}^{\delta}$

Figure 2: Gibbs sampling procedure for Bayesian estimation of HMM parameters.

A caution needs to be mentioned when such a sample of model parameters is used to characterize model uncertainty. Since the dynamics of observation process Y_t generated by a HMM with parameters $\boldsymbol{\theta}^{(R)}$ remains identical under arbitrary permutation of labels of the hidden states and the associated HMM parameters, there could be mis-ordered components within the sample of HMM parameters yielded by the Gibbs sampler. For ergodic HMMs, calibration of the order of the labels can be achieved by various methods, as suggested in [140], [141]. In the case of non-ergodic HMMs emphasized in this chapter, following [17], we propose to apply asymmetric priors on transition probabilities, namely, for each row of $\mathbf{P}$ and for each row of $\mathbf{Q}$, so that labels for the states cannot be switched.

The Bayesian approach described above has desirable convergence properties for the ergodic and homogeneous HMM [18]. Namely, for such HMMs, as the length of the observation sequence $\boldsymbol{y}_t$ approaches infinity (as t approaches infinity), the posterior distribution π_t of Bayesian estimates of HMM parameters obtained following the Gibbs sampling procedure illustrated in Figure 2 tends towards a multivariate normal distribution centered at true value of HMM parameters, with variance-covariance matrix shrinking toward zero at the rate of $\sqrt{t}$. In other words, for ergodic and homogeneous HMMs, the Bayesian estimation procedure shown in Figure 2 yields estimates of HMM parameters that converge in distribution to the true values of those parameters. As we will see in Section 3.5, this desirable property of the distribution of estimated parameters shrinking around the true parameter values as more data is used to estimate the model will be empirically demonstrated for non-ergodic and non-homogenous HMMs used in this chapter for modeling and monitoring of degradation process, though firm theoretical proofs require future studies.

3.3 PROCESS MONITORING BASED ON UNCERTAIN HMM

Following Fox et al. [113] and Brown et al. [114], Chollete and Djurdjanovic [81] proposed to monitor the slope of log-likelihoods of observation sequences, given the regime-specific degradation HMM defined by parameters $\boldsymbol{\theta}^{(R)}$. For an observation sequence $\boldsymbol{y}_T = \left[\boldsymbol{y}_{\tau_0}^T, \boldsymbol{y}_{\tau_1}^T, \ldots, \boldsymbol{y}_{\tau_H}^T\right]^T$, where $1 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_H = T$ indicate time instances where the regimes change, log likelihood slope λ_h for the time interval (τ_{h-1}, τ_h) is defined as

$$\lambda_h = \frac{\log \Pr\left(\boldsymbol{y}_{\tau_h} \middle| \boldsymbol{\theta}^{(R)}\right) - \log \Pr\left(\boldsymbol{y}_{\tau_{h-1}+1} \middle| \boldsymbol{\theta}^{(R)}\right)}{\tau_h - \tau_{h-1} - 1}, 1 \leq h \leq H, \tag{6}$$

Since the Bayesian HMM estimation procedure described in this chapter yields a distribution of model parameters, rather than a point estimate of those parameters [81], one should monitor the process degradation using the entire distribution of log-likelihoods of slopes Λ_h, rather than a single slope estimate (6). Namely, the estimate of the degradation model parameters $\boldsymbol{\Theta}^{(R)}$ is a distribution and thus, for a given observation sequence $\boldsymbol{y}_T = \left[\boldsymbol{y}_{\tau_0}^T, \boldsymbol{y}_{\tau_1}^T, \ldots, \boldsymbol{y}_{\tau_H}^T\right]^T$, the distribution Λ_h of log likelihood slopes for the time interval (τ_{h-1}, τ_h) becomes

$$\Lambda_h := \frac{\log \Pr\left(\boldsymbol{y}_{\tau_h} \middle| \boldsymbol{\theta}^{(R)}\right) - \log \Pr\left(\boldsymbol{y}_{\tau_{h-1}+1} \middle| \boldsymbol{\theta}^{(R)}\right)}{\tau_h - \tau_{h-1} - 1}, 1 \leq h \leq H, \tag{7}$$

and corresponds to the distribution of individual slopes (7), as model parameters $\boldsymbol{\theta}^{(R)}$ follow the distribution $\boldsymbol{\Theta}^{(R)}$ obtained through the Bayesian estimation.

Following Chollete and Djurdjanovic [81], distributions Λ_h can be normalized as follows to account for regime changing

$$\tilde{\Lambda}_h := \frac{\Lambda_h - \mu^{(r_h)}}{\sigma^{(r_h)}}, r_h \in R \tag{8}$$

where $\mu^{(r_h)}$ and $\sigma^{(r_h)}$ are the sample mean and sample standard deviation of the distribution of log-likelihood slopes for regime r_h relevant during the time interval[19] (τ_{h-1}, τ_h), as observed in the training data.

Once the normalized distributions $\tilde{\Lambda}_h$ of log likelihood slopes corresponding to nominal behavior are available, one can quantitatively evaluate how much a given observation $\boldsymbol{y}_T$ corresponds to that normal behavior. In order to do that, it is necessary to evaluate a similarity/dissimilarity measure of probability distributions of log-likelihood slopes observed on the dataset corresponding to the normal system behavior and that corresponding to a newly arrived observation sequence. For that purpose, one can use the Kolmogrov-Smirnov (KS) distance [142], a widely used tool in statistics, measuring the maximum difference between two cumulative distribution functions F_0 and F_1. It is a non-negative quantify defined as

$$KS(F_0, F_1) = \sup_x |F_0(x) - F_1(x)|. \tag{9}$$

and equaling 0 if and only if distributions F_0 and F_1 are identical

Over time, one can monitor the KS distance of the normalized distributions of log likelihood slopes of newly arrived observation sequences away from those observed on the training data, i.e. on the data corresponding to the normal system behavior. In this chapter, we approximate slope distributions using simple histograms [143], i.e. non-parametrically, without imposing specific assumptions on their forms [20]. For monitoring and fault detection purposes, we propose to define a KS distance-based Confidence Index (CI) as

[19] Note that during the time interval (τ_{h-1}, τ_h), only one regime of operation $r_h \in R$ takes place (and hence only one of the regime-specific HMMs is relevant).

[20] Please note that KS distances of empirical distributions of F_0 and F_1 based on two samples can be computed exactly, and it will converge to $KS(F_0, F_1)$ when sample sizes grow to infinity [142].

$$CI_{test} = \min_{1 \le i \le training\_size} KS(\tilde{\Lambda}_{training_i}, \tilde{\Lambda}_{test}), \tag{10}$$

where the term *training_size* is the number of training sequences. The CI can then be tracked using techniques from traditional statistical process control [13], raising alarms upon detection of CI values that violate some threshold.

3.4 Results and Discussion

We will first use a simulated dataset to illustrate the capability of the Bayesian estimation method described in Section 3.3 to identify the HMM parameters along with the associated estimation uncertainty for parameters of a HMM with discrete emission symbols. Another simulated dataset consisting of data sequences generated from several HMMs with continuously distributed observations will then be utilized to demonstrate the capability of the monitoring method proposed in Section 3.4. Finally, Bayesian HMM estimation introduced in this chapter will be applied to degradation modeling and monitoring of a semiconductor-manufacturing tool operating in a major 300 mm wafer fabrication facility.

3.4.1 Identification of Non-Ergodic Discrete HMM Parameters via Bayesian Estimation

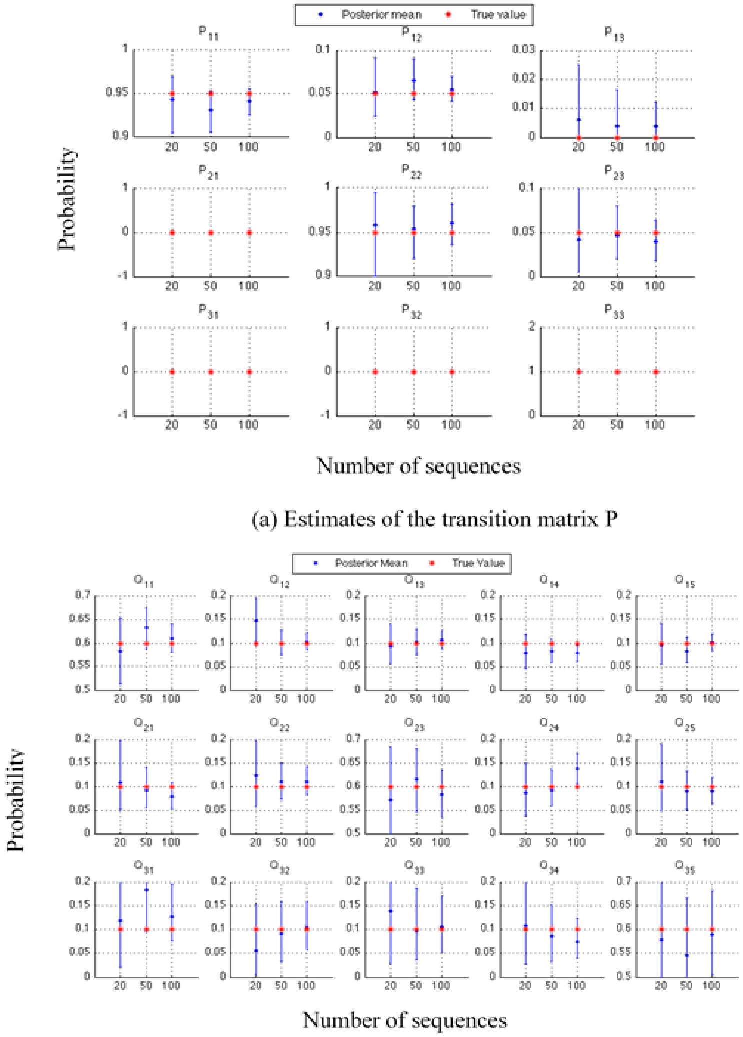
20 sequences, each with 20 observations, were simulated from a HMM with 3 states and 5 observation symbols, with the true model parameters being

$$\boldsymbol{\theta} = (\boldsymbol{\nu}, \mathbf{P}, \mathbf{Q}) = \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.95 & 0.05 & 0 \\ 0 & 0.95 & 0.05 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0.6 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.6 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.6 \end{bmatrix} \right).$$

This is a HMM of a left-to-right state transition structure, which means that it is non-ergodic and multiple sequences are necessary to train the model. We assumed the initial distribution $\boldsymbol{\nu}$ was known and hence the corresponding HMM parameters did not have to be estimated from the simulated sequences. Also, following [137], we assumed independent non-informative prior on each row of the transition matrix P, i.e. we assumed $Dir(0.5, 0.5, 0.5)$, $Dir(0, 0.5, 0.5)$ and $Dir(0, 0, 0.5)$ to the prior distributions for rows 1 to 3 of the transition matrix $\mathbf{P}$, respectively. For emission matrix Q, we assigned $Dir(0.5, 0.5, 0.5, 0.5, 0.5)$ as the prior distribution for each row [137]. Gibbs sampling procedure discussed in Section 3.2 was then evoked to generate a sample of HMM parameters from their posterior distribution, given the 20 observation sequences. It was assumed that the Gibbs sampling converged after the initial 200 iterations and the sample of HMM parameters was observed from the 800 subsequent iterations.

Figure 3 shows the empirical 2.5% and 97.5% error limits (95% credible interval) for each of the HMM parameters as well as the corresponding posterior means, evaluated after more and more observation sequences are presented to the identification algorithm. It is clearly visible that the posterior distribution of the HMM parameters pursued by the newly proposed HMM identification procedure progressively shrink in their variance and concentrate around the true model parameters, as the amount of training data increases. This indicates that the posterior distributions converge towards the point mass at the true model parameters, with variances of the distribution of parameter estimates converging to zero. Moreover, such phenomenon occurred with all types of HMMs we examined in simulations, including HMMs with Gaussian or discrete observations, as well as non-ergodic and non-homogeneous dynamics. To the best of authors' knowledge, formal proofs of convergence property of Bayesian estimation of HMM parameters and its quantitative characterization exist only for ergodic and homogeneous HMM [18] and do not exist for non-ergodic, non-homogeneous HMMs, such as those studied in this chapter

[21]. Though pursuit of these proofs is highly worthy of a focused study, it remains outside the scope of this chapter.



(a) Estimates of the transition matrix P

(b) Estimates of the emission matrix Q

Figure 3: Credible intervals and means of identified HMM parameters using 20/50/100 sequences

[21] As mentioned in Chapter 1, degradation processes are unidirectional and operation mode specific, which means the corresponding degradation HMMs are non-ergodic and non-homogenous.

3.4.2 Degradation Assessment Using Distributions of Slopes of Log-Likelihoods of Observation Sequences

We consider a simulated system whose normal condition can be modeled by a HMM with 3 states and Gaussian observations, with true parameters being

$$\boldsymbol{\theta} = (\boldsymbol{\nu}, \boldsymbol{P}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \left(\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \begin{bmatrix} 0.9 & 0.1 & 0.0 \\ 0.2 & 0.6 & 0.2 \\ 0.0 & 0.1 & 0.9 \end{bmatrix}, \begin{bmatrix} 0 \\ 5 \\ 10 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right).$$

Again, we assumed that the initial distribution $\boldsymbol{\nu}$ is known. Figure 4 shows the time series plots (left) and corresponding histogram plots (right) of five observation sequences $\{\boldsymbol{y}_{[1]}, \boldsymbol{y}_{[2]}, \boldsymbol{y}_{[3]}, \boldsymbol{y}_{[4]}, \boldsymbol{y}_{[5]}\}$, where $\boldsymbol{y}_{[1]}$and $\boldsymbol{y}_{[5]}$ were simulated from the "normal condition", i.e. using "faulty" HMM with parameters $\boldsymbol{\theta}$, while $\boldsymbol{y}_{[2]}, \boldsymbol{y}_{[3]}$and $\boldsymbol{y}_{[4]}$ were simulated using HMMs with parameters different from $\boldsymbol{\theta}$. To be more specific, probability $P_{1,1}$ of remaining in state 1 was changed from 0.9 to 0.8, 0.7 and 0.6, respectively, for $\boldsymbol{y}_{[2]}, \boldsymbol{y}_{[3]}$ and $\boldsymbol{y}_{[4]}$, with the probability P_{12} of transitioning from state 1 to state 2 being adapted so that $P_{1,1} + P_{1,2}$ =1. One can note from Figure 3 that very little difference is visible between the histograms of the five sequences, which means that traditional monitoring methods based on statistical analysis of sensory signatures would have a problem differentiating between those conditions.

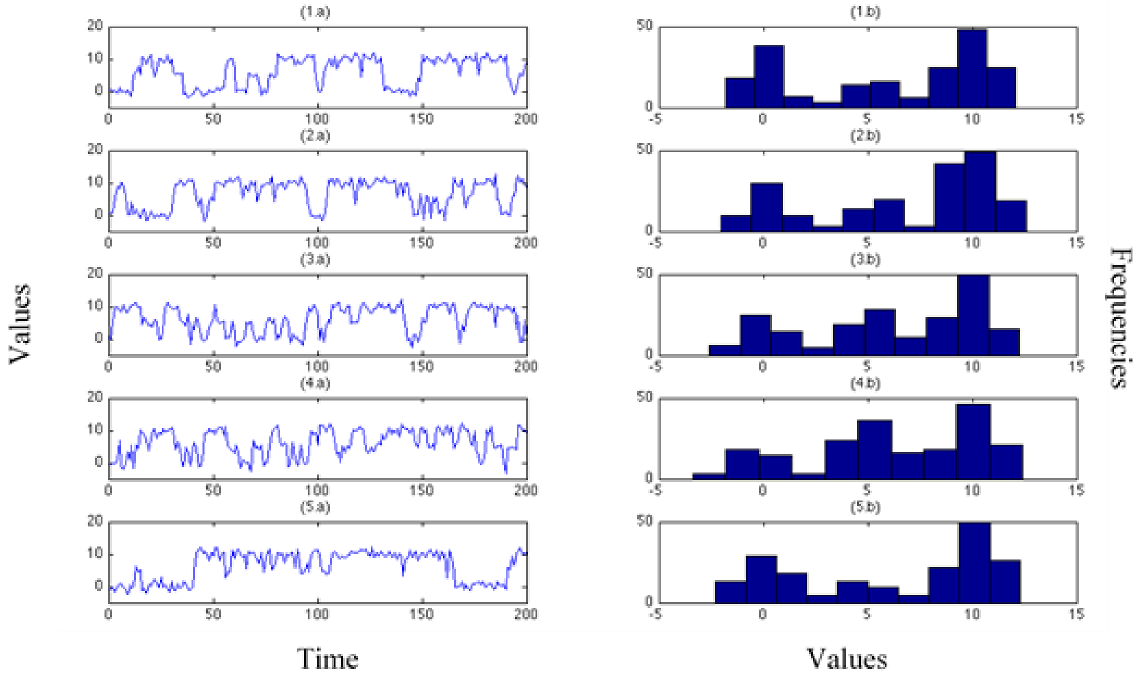


Figure 4: Time series plots (left) and histograms (right) of the simulated sequences: (normal, faulty, faulty, faulty, normal)

In contrast, from Fig. 5 it can be clearly seen that the underlying changes are detectable using the concept of HMMs. Namely, once the HMM parameters were identified from $\boldsymbol{y}_{[1]}$ using Bayesian estimation described in Section 3.2, the monitoring method introduced in Section 3.3 was applied on all five sequences $\{\boldsymbol{y}_{[1]}, \boldsymbol{y}_{[2]}, \boldsymbol{y}_{[3]}, \boldsymbol{y}_{[4]}, \boldsymbol{y}_{[5]}\}$ to investigate abnormalities within them. For these five sequences, we denote samples of the normalized log-likelihood slopes as $\{\tilde{\Lambda}_{[1]}, \tilde{\Lambda}_{[2]}, \tilde{\Lambda}_{[3]}, \tilde{\Lambda}_{[4]}, \tilde{\Lambda}_{[5]}\}$, with each $\tilde{\Lambda}_{[i]}$ evaluated from the corresponding sequence $\boldsymbol{y}_{[i]}$ using the final sample of the HMM model parameters identified based on $\boldsymbol{y}_{[1]}$. It is visible that as the HMM dynamics deviate further from the nominal HMM model, the distribution of log-likelihood slopes deviates further and further away from that corresponding to the normal system condition (corresponding to sequence $\boldsymbol{y}_{[1]}$).

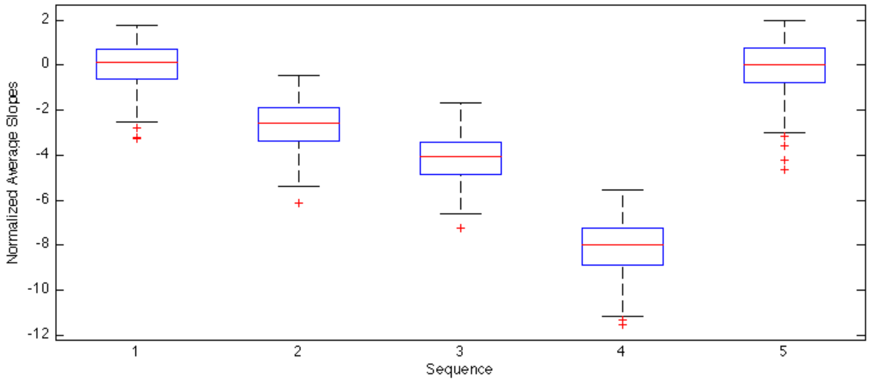


Figure 5: Distribution of normalized slopes evaluated for all observation sequences.

Many metrics exist that can characterize this change [144], including the KS distance mentioned in Section 3.3. As can be seen from the results of the two-sample KS goodness-of-fit test listed in Table 1. Unlike the direct analysis of observation distributions, the normalized slopes can clearly and reliably discriminate between the underlying conditions from which the time-series were generated.

	Purely statistical approach				HMM based approach proposed in this chapter			
First Sample (Normal)	$\boldsymbol{y}_{[1]}$ observation sequence				$\tilde{\Lambda}_{[1]}$ normalized log-likelihood slope			
Second Sample	$\boldsymbol{y}_{[2]}$	$\boldsymbol{y}_{[3]}$	$\boldsymbol{y}_{[4]}$	$\boldsymbol{y}_{[5]}$	$\tilde{\Lambda}_{[2]}$	$\tilde{\Lambda}_{[3]}$	$\tilde{\Lambda}_{[4]}$	$\tilde{\Lambda}_{[5]}$
True Condition of the 2nd samples	Faulty	Faulty	Faulty	Normal	Faulty	Faulty	Faulty	Normal
KS-statistics	0.10	0.11	0.14	0.08	0.86	0.96	1.00	0.09
P-value	0.27	0.22	0.05	0.63	0.00	0.00	0.00	0.38
Significant Difference Detected	No	No	No	No	Yes	Yes	Yes	No
Correctness	No	No	No	Yes	Yes	Yes	Yes	Yes

Table 1: Results of two sample Kolmogorov-Smirnov tests based on the HMM models and pure distribution based characterization of time series simulated in Section 3.4.2.

3.4.3 Application of HMM/Slopes-Based Monitoring Methodology on PECVD Data

In this section, we present the results of applying the proposed method for degradation modeling and monitoring to a Plasma-Enhanced Chemical Vapor Deposition (PECVD) process [145] routinely used in semiconductor manufacturing.

3.4.3.1 Description of the PECVD Dataset

A PECVD process performs deposition of thin films on a silicon wafer substrates via chemical reactions executed in electromagnetic plasma, which facilitates reaction at temperatures low enough not to damage circuits on the wafer. This process requires proper and accurate operations of numerous interacting subsystems on the PECVD tool, including reaction chamber, radio frequency (RF) plasma generation system, gas delivery system, vacuum pump, pendulum valve and seals [147]. The production procedure on a PECVD tool usually involves deposition of thin films onto wafers and periodic in-situ cleans[22] designed to remove residual depositions that accumulated on various parts of the tool (chamber walls, wafer pedestal, showerheads bringing gas into the reaction chamber, etc.). Occasionally, preventative maintenance (PM) in the form of a so-called wet-clean (manual scrubbing of chamber surfaces) is needed to remedy the side effect of imperfections[23] accumulated over multiple in-situ cleans and avoid production of bad wafers. Although numerous sensors are commonly available on the tool to measure the physics of the deposition process, a consequence of the complex interacting phenomena and distributed nature of electro-chemical reactions that take place in this tool is that its true degradation condition is inherently not fully observable and is often barely discernable from the sensory information.

Figure 6 illustrates how operation of a PECVD tool can be related to the terminology of operating regime specific HMMs of its degradation. Namely, each sequence of observations consists of sensory signatures observed between two in-situ cleans (each in situ clean restores the system condition and in between the in-situ cleans, the system degrades following operating regime specific HMMs). Within each sequence, several film thicknesses could be produced on the wafers, with degradation processes (HMM parameters) being different for each of those film thicknesses. In other words, different film thicknesses correspond to different operating regimes of this tool and hence a regime specific (film thickness specific) HMM is needed to describe its degradation.

[22] In-situ cleans on the tool considered here are performed by flowing ionized fluoride into the chamber, which reacts with depositions in the chamber and removes them.

[23] Unfortunately, in-situ cleaning agent (ionized fluoride in this case) also reacts, though much less intensely, with chamber walls and other tool parts, while potentially leaving small residual film depositions in some parts of the tool. All these imperfections accumulate over multiple in-situ clean cycles, resulting in long-term degradation of the tool.

In this study, multiple sensory signals are collected over several months from a PECVD tool operating in a major 300-mm semiconductor manufacturing facility. The tool was used to deposit four possible thicknesses of tetraethyl orthosilicate (TEOS) films onto silicon wafers. Automatic in-situ cleans were triggered based on the total thickness of deposited films since the last in-situ clean. Sampling rate of 10Hz was used to concurrently acquire signals from the tool's radio-frequency (RF) circuitry, as well as temperatures, pressures and flow rates from various parts of the tool. In total, the dataset consisted of signals corresponding to 1662 sequences of wafers, with each sequence containing approximately 25 to 100 wafers that were processed between two consecutive in-situ cleans.

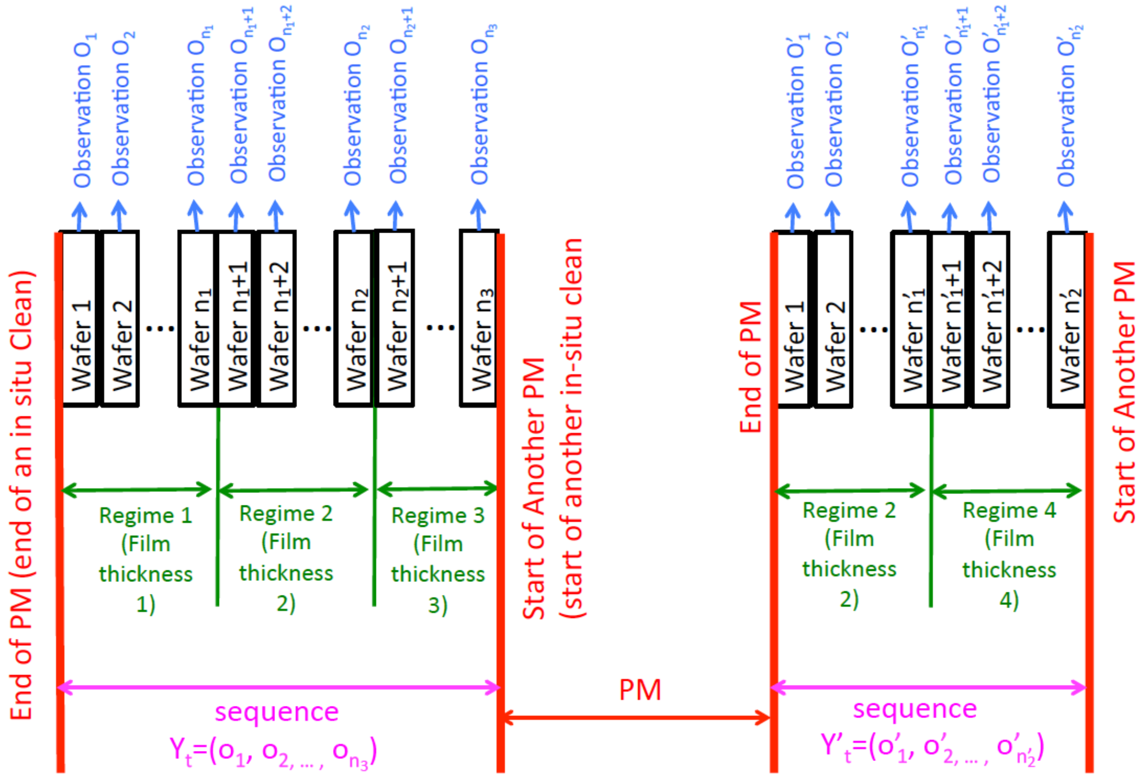


Figure 6: Illustration of PECVD tool operation in the context of operating regime specific HMMs of degradation.

Along with this massive dataset, the corresponding maintenance event logs and metrology data[24] were also available and were used for validation of the monitoring results[25]. Based on those logs, two periods of abnormal tool behavior were identified. Shortly after the first PM, the tool operation was stopped due to dramatically elevated particle counts on the wafers, followed by several repairs on the tool and a return to normal behavior. The interval between the first PM and the last repair on the tool after that PM is treated as the first faulty period, during which 5 faulty wafer sequences were recorded (those were attempts to bring the tool back up, but each of them produced wafers with high particle counts and more repairs needed to be taken). A second faulty period corresponds to a dramatic particle excursion event (Coulomb crystal formations [146]) and corresponds to the last 36 wafer sequences in the dataset (the excursions started soon after the second PM event in the maintenance logs and stretches to the end of this dataset, which is when the tool was finally stopped and taken down for a lengthy repair). According to these two periods, all 1662 sequences of wafers were labeled as either normal or faulty, allowing evaluation of fault detection capabilities of the monitoring methods to be discussed in the next subsection.

3.4.3.2 Process Monitoring on the PECVD Dataset

From the raw sensor readings (traces) collected during processing of each wafer, a set of 40 dynamic and statistical features was extracted, as described in [147]. This feature set is listed in Table 2 and was reduced using Fisher's Linear Discriminate Analysis (LDA) [148], yielding a subset of features that change the most between two in-situ cleans and can thus be seen as the most sensitive to the degradation condition of the

[24] Particle counts, as well as means and ranges of film thicknesses on the wafers output by the tool.

[25] The metrology and maintenance logs were used to identify periods of normal and abnormal tool behavior and evaluate how well the monitoring results conform to those logs.

tool [147]. These features were then discretized using a growing Self-Organizing Map (SOM) [149] constructed on the training dataset, which consisted of the first 499 wafer sequences. This set was selected since both the maintenance and metrology logs indicated that during that period, the tool behaved normally. The SOM based discretization step yielded 1163 sequences of observation symbols, out of which the first 499 (the training set) were used for model building[26], while the monitoring results were evaluated on all 1163 sequences in the dataset.

Signal	Signal Features			
Top Plate Temperature	Mean	Minimum	Amplitude	
Chamber Temperature	Mean	Minimum	Amplitude	
Pedestal 1 Temperature	Mean	Minimum	Amplitude	
LF Forward Power	Steady State Error	Tune Time		
LF Reflected Power	Steady State Error	Tune Time	Maximum	
HF Forward Power	Steady State Error	Tune Time	Overshoot High	Overshoot Low
HF Load Power	Steady State Error	Tune Time	Overshoot High	Overshoot Low
HF Reflected Power	Steady State Error	Tune Time		
Load Capacitor Voltage	Steady State	Tune Time		
Tune Capacitor Voltage	Steady State	Tune Time		
Pendulum Valve Angle	Steady State	Maximum		
Process Chamber Pressure	Steady State Error	Rise Time	Overshoot	Minimum
Liquid Flow Rate TEOS	Steady State Error	Rise Time	Overshoot	

Table 2: Features extracted from the PECVD tool sensors.

[26] Please note that selecting more or fewer wafer sequences for model building would have led to degradation models with more, or less model uncertainty. Explorations of how much data is needed to build a "sufficiently confident model" (model with sufficiently small model uncertainty) are very relevant to the issue of convergence of the Bayesian HMM identification procedure introduced in Section 2, which is outside of the scope of this chapter and will not be discussed here.

A non-ergodic (left-to-right) regime (film-thickness) dependent HMM with 4 regimes, 4 hidden states and 60 observation symbols (size of the SOM) was identified from the training set, using the Bayesian estimation procedure introduced in Section 3.2 and a non-informative prior for its parameters $\boldsymbol{\Theta}^{(R)}$. This HMM was then used for degradation monitoring, using the procedure introduced in Section 3.3. Figure 7(a) shows CIs defined by for all 1163 sequences in the dataset. For contrast, an industry-standard multivariate process monitoring method based on the Principal Component Analysis (PCA) and the use of T^2 statistics [150] was applied to the same dataset, using the same portion of the data for training. In order to enable 1-1 comparison of the two methods, each period between two in-situ cleans[27] was modeled using a distribution of T^2 statistics corresponding to that set of wafers. This enabled monitoring using CI indices based on the T^2 statistics defined by (10) , only not using distributions of normalized log-likelihood slopes, but distributions of T^2 statistics observed on the wafers between any two adjacent in situ cleans. Those CI indices are shown in Figure 7 (b).

[27] One sequence of observations for which the HMM based monitoring method from Section 3 yields one KS distance-based CI.

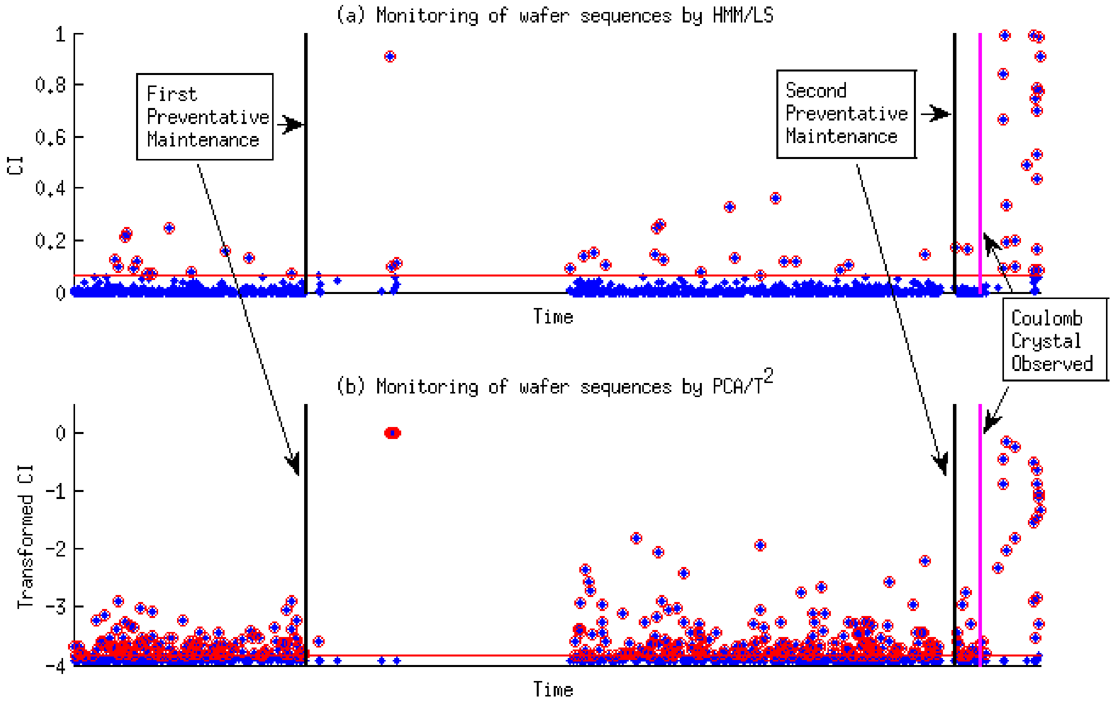


Figure 7: SPC chart of CI for HMM/LS and PCA/ T2. The control limits designated by the red horizontal lines are set based on yielding 65% true alarms in the test sequences in each case. The circled points are out-of-control points. The CI for the PCA case has been logarithmically (monotonically) transformed for visual comparison with the HMM case.

Receiver Operating Characteristic (ROC) curve and associated Area Under the Curve (AUC) [151] are utilized to evaluate the monitoring performance of the newly-proposed HMM-based and traditional PCA/T2 based methods. Figure 8 shows the ROC curves and the associated AUCs for the two methods. It is obvious that the ROC curve yielded by the new method outperforms the one produced by the PCA/ T2 based monitoring method for most potential control limits. Firstly, AUC corresponding to the HMM-based monitoring method is 23.6% larger than that of the PCA/T2 based method. Although the PCA/T2 based method captures the largest outlier sequences with slightly higher true positive rates than the HMM-based method (better performance for false positive rates less than 0.02), for false alarm rates above 0.02, the HMM-based method yields often dramatically higher true positive rates than the PCA/T2 based method. For a commonly used 5% false alarm rate level, the newly proposed approach has the true positive alarm rate of 81%, while the true positive alarm rate for the PCA/T2 based method is at 62%.

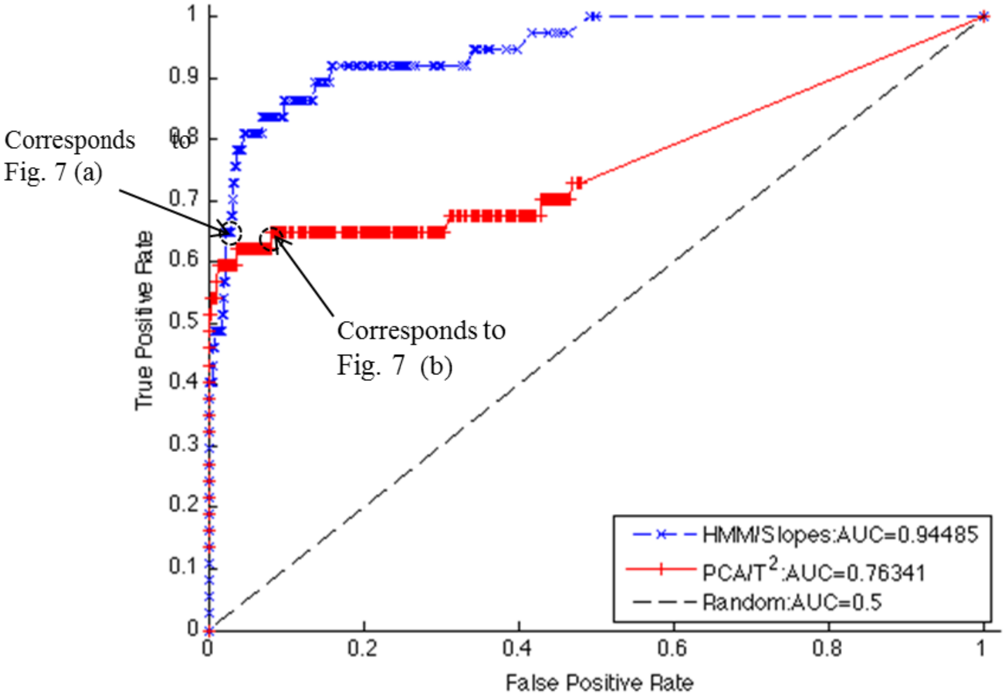


Figure 8: ROC curves and AUCs for HMM/Slopes and PCA/ T2. Each symbol represents the false positive vs. true positive rate of an entire SPC chart resulting from the re-categorization of a single additional sequence due to changing the control limit.

3.5 Conclusion

This chapter presented a novel condition-based monitoring methodology based on operating regime specific HMMs of system degradation, where HMM parameters are identified along with the corresponding estimation uncertainty using a Bayesian estimation approach. The use of the HMM concept overcomes the challenge of partial observability of the underlying condition seen in many engineering processes today, while Bayesian estimation naturally provided understanding of the uncertainties associated with the estimation of parameters of degradation HMMs. Simulations were used to demonstrate capabilities of the Bayesian estimation procedure to identify HMM parameters and the associated parametric uncertainties, as well as to detect changes in the process dynamics using HMMs identified via Bayesian estimation. Monitoring capabilities of the newly proposed degradation modeling and monitoring methods based on HMMs identified via Bayesian estimation were then demonstrated on a vast dataset obtained over several months from a PECVD tool operating in a major semiconductor fabrication facility.

Several directions for potential future work can be extended from this chapter. The newly proposed process monitoring methodology could be adapted and tried on other engineering processes where partial observability of the underlying conditions persists, such as etching process in semiconductor manufacturing or oil rig performance monitoring. Furthermore, statistical asymptotic theory for Bayesian estimation of HMM parameters, which exists for ergodic, homogeneous HMMs, needs to be pursued also for the case of non-ergodic and regime-specific HMMs. In this chapter, we offered empirical results that imply consistency of the newly proposed Bayesian estimator, even for non-ergodic and regime-specific HMMS, but we stopped short of formally proving this property. Besides the obvious theoretical allure, understanding of the asymptotic behavior of posterior distributions of Bayesian estimates of HMM parameters is of high practical importance too. Ascertaining whether more data indeed leads to a better model and enabling one to know how much data is really needed to estimate HMM with sufficient confidence are of high practical value in any discipline in which HMMs are utilized.

Chapter 4: Modeling Imperfect Maintenance and Monitoring with Individual Observation with HMM

4.1 INTRODUCTION

Condition-Based Maintenance (CBM) aims at facilitating maintenance operations exactly where needed and exactly when needed, based on sensor readings that reflect the actual condition of the maintained assets [2]. However, sensor readings obtained from highly complex engineering systems, such as distributed fields (plasma) or systems of many interconnected subsystems (automotive engines) usually provide insufficient information about the underlying conditions due to the insufficiently detailed physical models or the insufficient number and character of sensors. Monitoring of such systems therefore hinges on the development of degradation models capable of handling partial information about the system condition within the available sensory data.

The intuitive relation between the sensor readings and the underlying machine condition can be modeled probabilistically, by associating probabilities of the various levels of system degradation with the observed signatures extracted from the sensor readings. The concept of hidden Markov models (HMMs) [9], with its observable variables modeling the signatures extracted from the sensors mounted on the monitored machine, while its hidden states model the conditions of that machine. Such modeling approach was recently proposed in [81] and [152], and was successfully demonstrated in monitoring of a plasma-based deposition tool operating over multiple months in a major semiconductor-fabrication facility.

Despite the importance of these two studies, they implicitly assumed that after each maintenance action, the monitored system always returns to the state of being as-good-as-new upon completion of the maintenance intervention. However, maintenance actions are not perfect [153], and the post-maintenance condition depends on the effectiveness of that maintenance action. For example, chamber cleaning [154] is a type of periodical maintenance event commonly scheduled on semiconductor manufacturing tools to reestablish purity in the chamber environment. Such operation may leave residue on some surfaces inside the chamber and at the same time etch away some useful surfaces in that chamber. As a consequence, the tool condition after maintenance is a stochastic variable itself [155].

Monitoring of a system whose condition is modeled by hidden states of an HMM can be pursued in multiple ways once the parameters of the underlying HMM become available. One approach is to identify the most likely condition of the system via likelihoods of the newly arrived sensor data, given the HMMs modeling the degradation of the target system. This approach has been applied to diagnose historical wear patterns by Wang et. al [156] and detect deviations from the good-as-new tool by Ocak and Loparo [157]. Alternatively, one can monitor the departure of the dynamics in the new data from the dynamics in the nominal HMM modeling the normal system behavior. Fox et al. [113] and Brown et al. [114] demonstrated the efficacy of this approach in detecting faults when HMM is used for modeling the behavior of a robot and an electric power plant, respectively. Recently, Cholette and Djurdjanovic [81] used the later approach to model the degradation of a semiconductor-manufacturing tool using regime-specific degradation HMMs. Zhang et al. [152] extended the previous work by enabling estimation of parametric uncertainties in estimation of the HMMs that model the system degradation, as well as by introducing a novel HMM based condition monitoring method that incorporates those parametric uncertainties in the degradation HMMs into the fault detection decision.

Recognition of the degradation state in a HMM based model of degradation is a well-known problem about using available observation sequences to identify the corresponding hidden HMM states. A traditional approach to identifying the hidden HMM states is the Viterbi algorithm [158], which finds the sequence of hidden states that maximizes the log-likelihood for a given observation sequence. This algorithm has been applied to detect machine failure [69], and recognize degradation states of bearings [159] as well as the condition of a turbofan engine [160]. Even though entropy of the entire trajectory provided by the Viterbi algorithm was recently analyzed [161], uncertainty information of any individual state is not available through the Viterbi algorithm. On the other hand, estimation of the probability of the most recent state, or filtering, is another approach to the state recognition problem. This approach provides a full distribution of the current hidden state and has been utilized for recognizing degradation condition in a drilling tool [104] and an antenna [105]. However, in both the approaches mentioned above, the HMM parameters are assumed to be perfectly known[28], without any parameter uncertainties in them. As we have argued in [152], the parametric uncertainty of degradation HMM is highly important for modeling and monitoring of engineering systems. Unfortunately, to the best of authors' knowledge, a method capable of recognizing degradation states in an engineering system whose condition is modeled by HMMs with uncertain parameters does not exist.

Despite all the advances in applying HMMs for condition monitoring, modeling the variability in degradation condition caused by the imperfection in maintenance effectiveness has not been addressed. Considering this gap, we extend the condition modeling via hidden states of regime-specific HMMs from condition modeling only in

[28] The well-known and frequently used B-W algorithm for identifying HMM parameters provide results in such form.

the operating regimes where degradation state worsens, to also modeling potentially imperfect maintenance operations as yet another operating regime where degradation state probabilistically recovers, as modeled using right-to-left HMMs. We also propose a new method for performance assessment based on the newly proposed degradation and maintenance HMM whose parameters and corresponding uncertainties are obtained via the Bayesian identification procedure described in [152].

The remainder of this chapter is organized as follows. In Section 4.2, the concept of HMMs is briefly discussed, after which a novel HMM based degradation modeling framework that incorporates models of imperfect maintenance operations is described. A novel fault detection method based on the understanding of parametric uncertainties of the degradation HMM will be presented in Section 4.3. Section 4.4 will show results of degradation modeling and monitoring of an industrial semiconductor-manufacturing process accomplished using the new HMM based degradation modeling and monitoring methods described in Section 4.3. Finally, Section 4.5 offers conclusions of this chapter and outlines some possibilities for future research.

4.2 Hidden Markov Model

Hidden Markov model is a doubly embedded stochastic process $\{X_t, Y_t\}_{t=0}^{\infty}$ with an unobservable Markov chain X_t and the observable process Y_t for which at each time t, the observable variables Y_t are probabilistically related to the hidden state X_t at each time t. Assuming that the set of possible states for the hidden process $X_t$is $S = \{s_1, s_2, \dots, s_N\}$ and the set of possible observable symbols is $O = \{o_1, o_2, \dots, o_M\}$, the HMM can be described by a parameter triplet[29] $\boldsymbol{\theta} = (\boldsymbol{\nu}, \mathbf{P}, \mathbf{Q})$, consisting of the initial state distribution $\boldsymbol{\nu} \in [0,1]^N$, state transition probability matrix $\mathbf{P} \in [0,1]^{N \times N}$ and emission probability matrix $\mathbf{Q} \in [0,1]^{M \times N}$.

In many applications, physics of the process modeled using the HMM can lead to specific patterns in the state transition matrix. For example, if the hidden states $S = \{1,2,3\}$ represent condition of a monitored system, with state 1 denoting the excellent condition, state 2 denoting the OK condition and state 3 representing the bad condition, the state transition matrix P is constrained to be an upper triangular matrix, or $p_{ij} = 0, \forall i > j$, since without a maintenance operation, degradation state of the system can only deteriorate. Such "left-to-right" HMM structure has been utilized for degradation modeling in [81], [152].

Recently, the standard HMM construct described above has been extended to regime-specific HMMs by incorporating time-varying dynamics and observation models, in order to account for variability in the degradation models caused by the potentially variable operating regimes of the monitored system [81], [152]. However in those papers, each maintenance operation was assumed to be perfect meaning that the condition after each maintenance was assumed to be as good as new with probability 1.

In order to model the potential imperfections of maintenance operations, in this chapter, we will model the degradation state recovery caused by a maintenance intervention as yet another Markovian hidden state transition, only this time encoded by a left-to-right structure of the state transition matrix, denoting a stochastic and thus imperfect recovery. Suppose the operating regimes over time are denoted by a sequence

[29] Emission distributions, such as Gaussian distributions, can be conceptualized and parameterized, leading to a vector of state dependent means and variances substituting the emission matrix Q in the parameter triplet $\boldsymbol{\theta}$.

$z_t, t = 0,1,2, \ldots$, with each z_t having a known value from the set of possible operating regimes

$$R = \{r_1, r_2, \ldots, r_L, \rho_1, \rho_2, \ldots, \rho_{L'}\}$$

where r's denote the production regimes improving the system condition and ρ's are the maintenance regimes (system condition restoring). For each regime in the set R, let us allow different HMM dynamics and observation probabilities by introducing a regime-specific HMM concept, which, assuming *N* hidden states $\{s_1, s_2, \ldots, s_N\}$, can be described by parameters

$$\boldsymbol{\theta}^{(R)}$$

$$= \left(\boldsymbol{\nu}, \mathbf{P}^{(r_1)}, \mathbf{Q}^{(r_1)}, \mathbf{P}^{(r_2)}, \mathbf{Q}^{(r_2)}, \ldots, \mathbf{P}^{(r_L)}, \mathbf{Q}^{(r_L)}, \mathbf{P}^{(\rho_1)}, \mathbf{Q}^{(\rho_1)}, \mathbf{P}^{(\rho_2)}, \mathbf{Q}^{(\rho_2)}, \ldots, \mathbf{P}^{(\rho_{L'})}, \mathbf{Q}^{(\rho_{L'})}\right),$$

with initial state probability vector

$$\boldsymbol{\nu} = [\nu_1 \quad \nu_2 \quad \cdots \quad \nu_N]^T; \nu_i = \Pr(X_0 = s_i)\,, i = 1,2, \ldots, N$$

regime-specific left-to-right state transition matrices $\mathbf{P}^{(r)}, r \in \{r_1, r_2, \ldots, r_L\}$

$$\mathbf{P}^{(r)} = \left[p_{i,j}^{(r)}\right]_{i,j=1,2,\ldots,N}, \; p_{i,j}^{(r)} = \Pr\left(X_{t+1} = s_j \middle| X_t = s_i\right), \text{for } z_t = r$$

describing state transitions that degrade the system state[30], "right-to-left" transition matrices $\mathbf{P}^{(\rho)}, \rho \in \{\rho_1, \rho_2, \ldots, \rho_{L'}\}$

[30] These matrices describe production regimes of the system and satisfy $P_{ij}^{(r)} = 0$, for $1 \le i < j \le N, 1 \le l \le L$.

$$\mathbf{P}^{(\rho)} = \left[p_{i,j}^{(\rho)}\right]_{i,j=1,2,\dots,N}, \; p_{i,j}^{(\rho)} = \Pr\left(X_{t+1} = s_j \middle| X_t = s_i\right), \text{for } z_t = \rho$$

describing the maintenance related state transitions that recover the system state[31], regime-specific emission probability matrices $\boldsymbol{Q}^{(r)}, r \in \{r_1, r_2, \dots, r_L, \rho_1, \rho_2, \dots, \rho_{L'}\}$ satisfying

$$\mathbf{Q}^{(r)} = \left[q_{i,j}^{(r)}\right]_{\substack{i=1,2,\dots,N\\ j=1,2,\dots,M}}, q_{i,j}^{(r)} = \Pr\left(Y_t = o_j \middle| X_t = s_i\right), \text{for } z_t = r$$

and the hidden states process X_t progressing according to probabilities[32]

$$\begin{bmatrix} \Pr(X_t = s_1) \\ \Pr(X_t = s_2) \\ \vdots \\ \Pr(X_t = s_N) \end{bmatrix} = \boldsymbol{\nu}\left(\prod_{i=0}^{t} \mathbf{P}^{(z_i)}\right). \tag{11}$$

Let us note that Eq. (11) formalizes the well-known notion of the continuity of degradation, stipulating that the last state of degradation after one operating regime becomes the initial state of degradation for the next one.

The HMM parameters $\boldsymbol{\theta}$ need to be identified from the available realizations of the observable variables (sensor readings), and Chapter 3 described a Bayesian estimation based approach to identification of those parameters.

[31] These matrices describe maintenance regimes of the system and satisfy $P_{ij}^{(\rho)} = 0$, for $1 \le j < i \le N, 1 \le l \le L'$.

[32] These probabilities assert the probability of degradation states in future time t assuming current time is the starting time. If past observations are available, the probability of the current state can be estimated by the so-called forward algorithm [9], and can be similarly used for assessing the predicative degradation condition of the modeled system.

4.3 CONDITION MONITORING

Condition monitoring needs to be done for each newly arrived observation to facilitate on-line condition monitoring of the system without any delay. For a system whose degradation is modeled by HMMs, as proposed by Chollette and Djurdjanovic [81] and Zhang et al. [152], one approach to realize this is to use the well-known Viterbi algorithm [158] to determine the most likely sequence of states

$$\boldsymbol{x}_t^* = \underset{x_t}{\operatorname{argmax}} \Pr(x_t, y_t | \boldsymbol{\theta}^{(R)})$$

Nevertheless, as mentioned earlier this method does not take into account the uncertainty of the model, nor does it offer information on the uncertainties regarding the most likely states $\boldsymbol{x}_t^*$.

As an alternative, let us estimate the probability of the current state x_t being the most degraded state given an observation sequence $\boldsymbol{y}_t$. Following [9], it can be calculated by using forward probabilities $\alpha_t(i)$ defined by

$$\alpha_t(n) = \Pr(x_t = n, \boldsymbol{y}_t | \boldsymbol{\theta}^{(R)})$$

followed by a normalization step

$$\bar{\alpha}_t(n) = \frac{\alpha_t(i)}{\sum_{i=1}^{n} \alpha_t(i)} = \Pr(x_t = n | \boldsymbol{y}_t, \boldsymbol{\theta}^{(R)}) \tag{12}$$

Since the Bayesian HMM estimation procedure introduced in Chapter 3 and utilized in this chapter yields a distribution of model parameters, rather than a point estimate of those parameters, one should monitor the probability of the worst state [33], using the entire distribution of $\bar{\alpha}_t(n)$, rather than a single state probability estimate in (12). Namely, the

[33] State n.

estimate of the degradation model parameters $\boldsymbol{\Theta}^{(R)}$ from the Bayesian estimation procedure is a distribution and the distribution of $\bar{\alpha}_t(n)$ over the entire distribution $\boldsymbol{\Theta}^{(R)}$ can be considered. One possibility is to monitor the expected value for the distribution of $\bar{\alpha}_t(n)$

$$A_t = \int_{\Omega^{(R)}} \bar{\alpha}_t(n)\, \pi\big(\boldsymbol{\theta}^{(R)}\big)\, d\boldsymbol{\theta}^{(R)} \int_{\Omega^{(R)}} \Pr\big(x_t = n \big| \boldsymbol{y}_t, \boldsymbol{\theta}^{(R)}\big)\, \pi\big(\boldsymbol{\theta}^{(R)}\big)\, d\boldsymbol{\theta}^{(R)}$$

which can be estimated as the average obtained through sampling in $\boldsymbol{\Theta}^{(R)}$, as described in [152]. This is the method pursued in the rest of the chapter.

4.4 RESULTS AND DISCUSSION

4.4.1 Description of the PECVD Datasets

The dataset used in this study is collected from a PECVD tool used to deposit thin films of multiple thicknesses onto silicon wafers, with residual depositions in the tool champer removed by periodic in-situ cleans [147], or so-called wet cleans [154], which take place less frequently and remove residual depositions caused by imperfections in the in-situ cleans. Figure 9 illustrates operation of a PECVD tool in terms of operating regime-specific HMMs of its degradation and maintenance operations. Namely, each sequence of observations consists of sensory signatures observed between two in situ cleans, with each in situ clean stochastically improving the system condition, while in between the in situ-cleans, the system degrades according to the operating regime-specific HMMs). Within each sequence, several film thicknesses could be deposited on the wafers (multiple subsequences of film depositions can be observed), with degradation processes being different for each of those film thicknesses[34]. In other words, different

[34] I.e., also the parameters of the corresponding degradation HMMs are different for each film thickness.

film thicknesses correspond to different operating regimes of this tool, and hence, a regime-specific (film thickness specific) HMM is needed to describe its degradation.

Ideally, HMMs for modeling condition recoveries from maintenance operations could be identified from sensory signatures collected during those interventions, just like degradation models are identified from the corresponding sensory signatures. However, in spite of its unique size and granularity[35], this data set does not contain sensory signatures corresponding to the in-situ cleans and hence, an alternative approach was needed. Different regimes of deposition can leave different byproduct or residue levels on the chamber, and thus the effectiveness of each *in situ* clean depends to a large degree on the last deposition sequence executed prior to that *in situ* clean. On the other hand, condition of the PECVD tool at the start of each wafer sequence[36] reflects the condition to which the previous in-situ clean brought the tool. Therefore, the state-transition between the state just after processing the last pre-clean wafer and the state just before processing the first post-clean wafer reflects the maintenance (in-situ clean) activity and is assumed to follow an *in situ* clean regime that is associated with the last pre-clean deposition regime. As described in Sec. 4.2, all in-situ clean regimes are associated with right-to-left state transition matrices, illustrating recoveries of system conditions when those cleans take place. Eventually, the overall regime-specific HMM contains regimes for all deposition thicknesses, as well as in-situ clean regimes.

In this study, multiple sensory signals are collected over several months from a PECVD tool operating in a major 300-mm semiconductor-manufacturing facility. The tool was used to deposit four possible thicknesses of tetraethyl orthosilicate (TEOS) films

[35] Signals from dozens of sensors collected during more than 30,000 depositions, all collected concurrently at 10Hz.

[36] I.e, just after the in-situ clean.

onto silicon wafers. Automatic in situ cleans were triggered based on the total thickness of deposited films since the last in situ clean. Sampling rate of 10 Hz was used to concurrently acquire signals from the tool's RF circuitry, as well as temperatures, pressures, and flow rates from various parts of the tool. In total, the dataset consisted of signals corresponding to 2556 sequences of wafers, with each sequence containing signals from approximately 25 to 100 wafers that were processed between two consecutive in situ cleans.

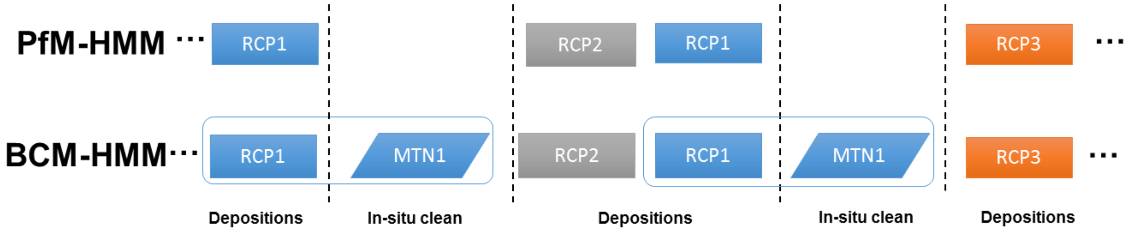


Figure 9: Illustration of regime-specific HMM of system conditions assuming perfect maintenance (referred to as the Perfect Maintenance HMM or PfM-HMM) and regime-specific HMM assuming imperfect HMM (referred to as Backward Coupling Maintenance HMM or BCM-HMM). The terminology is adopted to emphasize the association of regimes between each in situ clean and the last deposition regime before that in-situ clean.

Along with this massive dataset, the corresponding maintenance event logs and metrology data were also available and were used for validation of the monitoring results. Based on those logs, two periods of abnormal tool behavior were identified. Shortly after the first PM, the tool operation was stopped due to dramatically elevated particle counts on the wafers. The interval between the first PM and the last repair on the tool after that PM is treated as the first faulty period,

The second faulty period corresponds to a dramatic particle excursion event caused by Coulomb crystal formations [146] and correspond to the last 36 wafer sequences in the dataset. Consequently, all 2556 sequences of wafers were labeled as either normal or faulty, allowing evaluation of fault detection capabilities of the monitoring methods, which is to be discussed in the next section.

4.4.2 Data Processing and Process Modeling by Regime-Specific HMM

From the raw sensor readings collected during processing of each wafer, a set of 40 dynamic and statistical features was extracted, as described by Bleakie and Djurdjanovic [147]. These features were then discretized using a growing self-organizing map (SOM) [149] constructed on the training dataset. The training dataset consisted of the first 512 wafer sequences and was selected for training since both the maintenance and metrology logs indicated that during that period, the tool behaved normally.

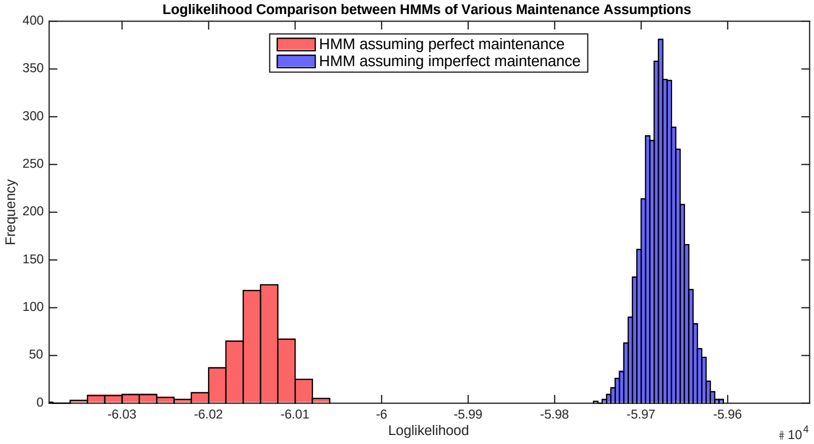


Figure 10: Comparison of distribution of loglikelihood slopes based on the regime-specific HMM assuming perfect maintenance and regime-specific HMM assuming imperfect maintenance.

A regime (film-thickness and in-situ clean) dependent HMM with 8 regimes (4 deposition thicknesses and 4 in-situ clean regimes), 4 hidden states, and 60 observation symbols (size of the SOM) was identified from the training set, along with the corresponding parameter uncertainties, using the Bayesian estimation procedure introduced in Chapter 3 and a non-informative prior for HMM parameters. This HMM will be referred to as Backward Coupling Maintenance HMM, or BCM-HMM for the rest of the chapter. In contrast, the same amount of training data and the same estimation method were used to train regime dependent degradation HMMs assuming perfect maintenance operations, which yielded in 4 degradation HMM regimes with 4 states and 60 observation symbols. This method corresponds to the degradation model used in Chapter 3 and will be referred to as the Perfect Maintenance HMM or PfM-HMM. The distribution of log-likelihoods yielded by these two models, as evaluated on the training set, is shown in Figure 10 and some properties of the corresponding distributions are listed in Table 3. It is clear that the BCM-HMM outperforms the PfM-HMM significantly in terms of likelihood, which indicates that modeling of maintenance imperfections considerably improves the model of degradation dynamics within the PECVD process.

Model	**Mean of Log-likelihood**	**Variance of Log-likelihood**	**Sample Size**	**Improvement in Mean of Log-likelihood**
PfM-HMM	-60156.8	2805.826	500	NA
BCM-HMM	-59675.4	495.5023	4000	0.8%

Table 3: Improvement in log-likelihood based on the HMM with and without modeling of imperfect maintenance, using the same training dataset.

4.4.3 Improvement in Detection Performance for Sequence-Based Process Monitoring

Receiver operating characteristic (ROC) curve, and the associated areas under the curve (AUC) [151] are utilized to evaluate [37] the monitoring performance of the newly proposed BCM-HMM-based method, the PfM-HMM-based method proposed in Chapter 3, as well as the traditional PCA/T^2 based statistical process control monitoring method [150]. Figure 11 shows the ROC curves and the associated AUCs for the three methods. It is evident that the ROC curve yielded by the new method outperforms the other two monitoring methods for almost all potential control limits. Furthermore, AUC corresponding to the BCM-HMM-based monitoring method is 2.75% larger than that of the PfM-HMM-based method, and 27.23% larger than that of the PCA/T^2 based method.

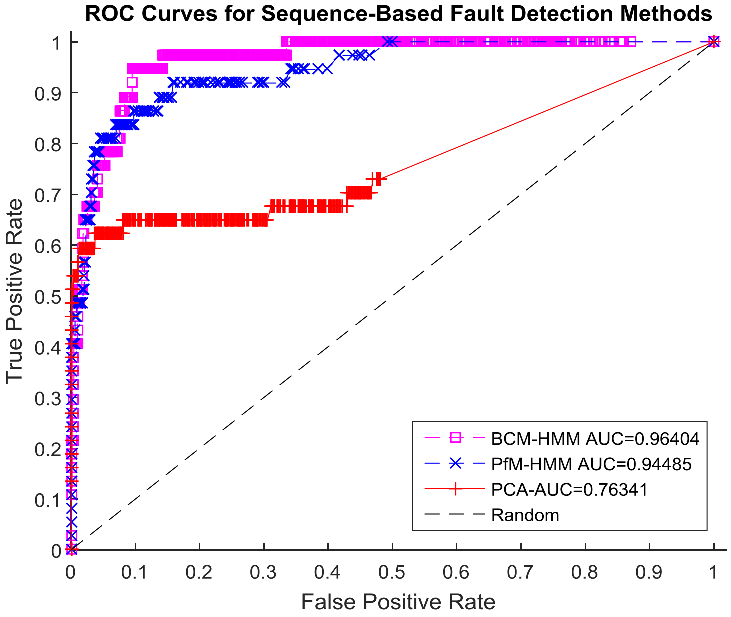


Figure 11: ROC curves for detection of faulty sequences using the models of PCA/T^2 PfM-HMM/LS, BCM-HMM/LS.

[37] The evaluation of the three methods is based only on the 1163 sequences of deposition regimes from the entire 1803 test sequences, because we only know the ground truth about the faultiness of these selected sequences.

4.4.4 Evaluation and Analysis of Wafer-Based Monitoring Methods

In this section, we assess the monitoring performance of the fault detection method based on the use of degradation HMMs that account for maintenance imperfections (BCM-HMM) and individual observations, as described in Section 4.3. This method, denoted as the BCM-HMM/filtering method, was evaluated on the aforementioned PECVD tool data and compered to several benchmark methods.

These methods include the traditional PCA/T2 SPC method based on observations from each individual wafer, monitoring based on HMMs that do not account for maintenance imperfections and the newly proposed filtering that evaluates hidden state probabilities for any given sensory observation (labeled as the PfM-HMM/filtering method), monitoring based on degradation HMMs that assume perfect maintenance operations, but using the mean log-likelihood slopes within a given observation sequance for monitoring, as suggested in [81] and [152](labeled as the PfM-HMM/slope method) and finally, method based on the newly proposed degradation HMMs that model maintenance imperfections, but using mean log-likelihood slopes of observation sequences (labeled as BCM HMM/slope method). Figure 12 shows results of this comparison and it is evident that the BCM-HMM/filtering monitoring method outperforms all the other approaches for all false positive alarm rates. It is interesting to note that the PfM-HMM/filtering method has dramatically worse performance than the counterpart method that uses the BCM-HMM degradation model (or any other method for that matter). Such poor performance may be attributed to the fact that the accuracy of the probabilities of the hidden state sequence relies heavily on the accuracy of recognition of the initial condition for each sequence. Within the PfM-HMM degradation model, the initial conditions were always assumed to be as-good-as-new and that deteriorated the resulting monitoring performance based on state filtering. On the other hand, the PfM-HMM degradation model coupled with monitoring based on the mean log-likelihood slopes for any given sequence provides a slightly better (higher) AUC value than the BCM-HMM degradation model coupled with monitoring based on the mean log-likelihood slopes. This advantage can be attributed to the fact that the log-likelihood slopes in the degraded states become steeper when the initial wafer state is modeled as perfect, as opposed to being recognized as random, which is the case with the BCM-HMM degradation model.

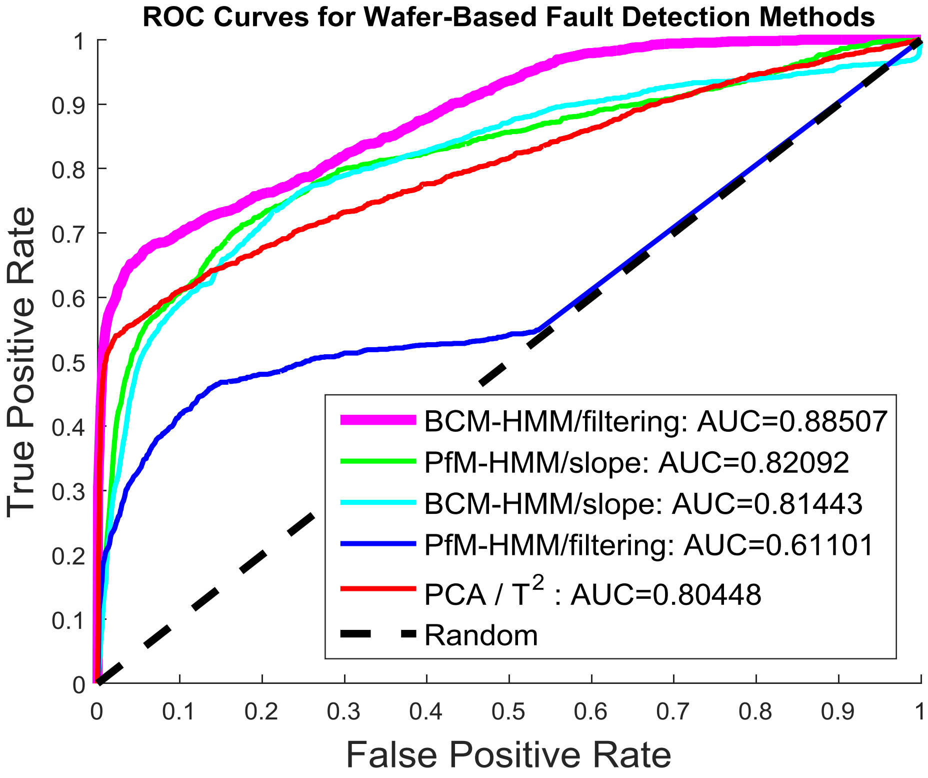


Figure 12: ROC curves for detection of faults in individual wafers from the PECVD dataset.

4.5 CONCLUSIONS

This chapter introduced a new method for modeling of degradation in complex systems using regime-specific HMMs that model imperfections in maintenance activities. Furthermore, a novel monitoring method based on the estimation of probabilities of hidden condition states using degradation HMMs with uncertain parameters was also proposed. Unlike the HMM-based monitoring methods introduced in [81] and [152], the newly proposed method enables on-line performance evaluation based on each individual observation symbol, rather than monitoring solely based on an entire sequence of observations.

Using a large-scale semiconductor manufacturing production dataset, it was demonstrated clearly that the newly proposed model yields significantly higher data likelihoods compared to the previously reported degradation models that assumed perfect maintenance operations, thus indicating better representation of the data when the new method is used. Furthermore, the newly proposed monitoring method based on the degradation HMMs that are aware of maintenance imperfections and fault detection based on estimating probabilities of hidden degradation states using uncertain HMMs of system degradation yielded significantly and consistently better performance compared to a set of benchmark methods.

Many extensions to the research presented in this chapter are possible. The methodology seems to be obviously applicable monitoring of plasma etch processes in semiconductor manufacturing, where the periodic yet imperfect chamber cleans take place after periods of production. Furthermore, other complex and insufficiently observable systems, such as Li-ion battery, or oil/gas extraction systems could be monitored using HMM-based models of degradation. In addition, sensory signatures collected during maintenance operations could be used to estimate maintenance-related

HMMs of condition dynamics (condition recoveries), similarly to how degradation HMMs were estimated in [81], [152] and in this chapter. Finally, let us note that ultimate benefits of the work presented in this chapter and even this thesis would be realized once degradation information from multiple machines in a system gets collected, coordinated and utilized for cost-effective operational decision-making. In a recent thesis [164], Celen proposed optimized operational decision-making for systems of machines whose degradations followed operating regime dependent HMMs described in this thesis. Nevertheless, degradation HMMs in [164] were assumed to be perfectly known and were not obtained from any realistic piece of equipment. Hence, full integration of degradation modeling described in this chapter and operational decision-making described in [164] remains to be done in the future.

Chapter 5: Analysis of Convergence Properties of Bayesian Estimator for HMM Parameters

5.1 FORMULATION

Hidden Markov model with finite state space is a doubly embedded stochastic process $\{(X_t,\ Y_t)\}_{t=0}^{\infty}$ following a probability law $\mathbb{P}$ [38] on a measurable space $(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}, \mathcal{B}_x^{\otimes\mathbb{N}} \otimes \mathcal{B}_y^{\otimes\mathbb{N}})$ [39], based on a probability space $(\mathcal{S}, \mathcal{A}, \mu_{\mathcal{S}})$, where $\{X_t\colon \mathcal{S} \to \mathcal{X}, \forall t \geq 0\}$ is a set of (hidden) random variables with values in a finite state space $\mathcal{X} = \{s_1, s_2, \dots, s_N\}$, and $\{Y_t\colon \mathcal{S} \to \mathcal{Y}, \forall t \geq 0\}$ is a set of (observable) random variables with values in the observation space $\mathcal{Y}$ that could be discrete or continuous[40]. We assume that both $\mathcal{X}$ and $\mathcal{Y}$ are equipped with appropriate σ-finite positive measures μ_X and μ_Y, as reference measures based on which density functions can be defined[41]. In addition, we use abbreviation $\boldsymbol{X}_{t_1:t_2}$ for the subsequence $(X_{t_1}, X_{t_1+1}, \dots, X_{t_2})$ of the state sequence $\{X_t\}_{t=0}^{\infty}$, along with $\boldsymbol{x}_{t_1:t_2}$for its realization (when $t_1 = 0$, we simply denote $\boldsymbol{X}_t$ and $\boldsymbol{x}_t$), and such convention is applied to the observations $\{Y_t\}_{t=0}^{\infty}$ as well. Using the above notation, we assume the following Markov properties:

M1) $f_{X_t|\boldsymbol{X}_{t-1},\boldsymbol{Y}_{t-1}}(x_t|\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) = f_{X_t|X_{t-1}}(x_t|x_{t-1}), \forall x_t, x_{t-1} \in \mathcal{X}, \forall t \geq 1.$

M2) $f_{Y_t|\boldsymbol{X}_t,\boldsymbol{Y}_{t-1}}(y_t|\boldsymbol{x}_t, \boldsymbol{y}_{t-1}\ ) = f_{Y_t|X_t}(y_t|x_t), \forall x_t \in \mathcal{X}, y_t \in \mathcal{Y}, \forall t \geq 0.$

Then the probability law $\mathbb{P}$ can be completely defined by the initial density $v\colon \mathcal{X} \to [0,1]$ such that $\nu(s_i) = f_{X_0}(s_i), \forall i$, transition densities $\{p^t\}_{t=1}^{\infty}$such that $p^t\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and $p^t(s_i, s_j) = f_{X_{t+1}|X_t}(s_j|s_i), \forall t \geq 0$, and conditional densities $\{q^t\}_{t=1}^{\infty}$ that

[38] $\mathbb{P}$ is a generic notation for probability measure, and will be specified using subscript in specific contexts.

[39] This infinite product measurable space is commonly used in literature on asymptotic analysis for ergodic HMM [10], [11], [18], [128].

[40] Observations don't have to be scalar and could be multidimensional, e.g. when $\mathcal{Y}$ could be $\mathbb{R}^2$.

[41] E.g., Lebesgue measure or counting measure.

relate hidden states and observations so that $q^t\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ and $q^t(s_i, y) = f_{Y_t|X_t}(y|s_i), \forall t \geq 0$, where $q^t(s_i, y)$ is measurable in y for any s_i.

Let $\mathbb{P}_{\boldsymbol{\theta}}$ be the law that governs a time-homogeneous [42] HMM $\{(X_t,\, Y_t)\}_{t=0}^{\infty}$ and is parameterized by a d-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\nu}, \boldsymbol{P}, \boldsymbol{\phi})$ composed of a stochastic vector $v \in \mathbb{R}_+^N$, a stochastic matrix $\boldsymbol{P} \in \mathbb{R}_+^{N\times N}$ with $P_{ij} = p(s_i, s_j)$, as well as a parameter vector[43] $\phi \in \mathbb{R}^{d_\phi}$ that determines the function q such that the density functions $f_{Y|X}(y|s_i), 1 \leq i \leq N$ are from the same parametric family. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\nu_0}, \boldsymbol{P_0}, \boldsymbol{\phi_0})$ be an unknown, but fixed parameter. Let $\mathbb{P}_0$ be the law determined by $\boldsymbol{\theta}_0$ [44] and $\mathbb{E}_0[\cdot]$ be the corresponding expectation of some random variable. In the Bayesian statistics framework, $\boldsymbol{\theta}$ can be viewed as realization of a random vector $\boldsymbol{\Theta}\colon \mathcal{S} \to \Omega$ with a prior distribution Π, which is induced from $(\mathcal{S}, \mathcal{A}, \mu)$ and has density $\pi(\cdot)$ with respect to some reference measure μ_Ω.

Traditional studies of large sample properties of parameter estimates for HMM parameters $\boldsymbol{\theta}_0$ use an observation sequence generated from a single process $\{(X_t,\, Y_t)\}_{t=0}^{\infty}$ of infinite length [10], [11], [18], [128], [133], [165]. Instead, multiple independent observation sequences of finite length are typically collected from degrading manufacturing system and need to be used to build a HMM that models its degradation [81]. Hence, we consider a growing number of independent identically distributed (i.i.d) processes

[42] Despite the non-homogeniety of the regime-specific HMM discussed in previous chapters, we limit ourselves to the homogeneous HMM where both p^ts and $q^t s$ are constant over time and can be denoted, as p and q.

[43] For instance, $Q^t \in \mathbb{R}_+^{N\times M}$ and $Q_{ij}^t = q^t(s_i, o_j), \forall t, i, j$, when the observation space is $\mathcal{Y} = \{o_1, o_2, \dots, o_M\}$.

[44] Since we can use $\boldsymbol{\theta}_0 = (\boldsymbol{\nu_0}, \boldsymbol{P_0}, \boldsymbol{\phi_0})$ to define any finite dimensional measure on $(\mathcal{X}^j \times \mathcal{Y}^j, \mathcal{B}_x^{\otimes j} \otimes \mathcal{B}_y^{\otimes j}), \forall j > 0$,, a unique $\mathbb{P}_0$ must exist for the measure space $(\mathcal{X}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}, \mathcal{B}_x^{\otimes \mathbb{N}} \otimes \mathcal{B}_y^{\otimes \mathbb{N}}, \mathbb{P}_0)$ by the Kolmogorov consistency theorem. Also, for each realization ω, we consider (X_i, Y_i) is the coordinate projection, i.e. $X_i(\omega) = \omega_{n,1}, Y_i(\omega) = \omega_{n,2}$.

$$\left\{\left\{\left(X_{k,t},\ Y_{k,t}\right)\right\}_{t=1}^{T}\right\}_{k=1}^{\infty},$$

each of which follows the marginal law $\mathbb{P}_0$ on the first T component $\left\{\left(X_{k,t},\ Y_{k,t}\right)\right\}_{t=1}^{T}$. In addition, the infinite product measure followed by these processes will be denoted by $\mathbb{P}_0^{\infty}$. In general, the marginal density function for the kth observation sequence $\boldsymbol{y}_{k,T} = \left[y_{k,1}, y_{k,2}, \dots, y_{k,T}\right]$ is

$$f_{\boldsymbol{Y}_{k,T}|\Theta}\left(\boldsymbol{y}_{k,T}|\boldsymbol{\theta}\right) = \sum_{\boldsymbol{x}_T \in \mathcal{X}^T} \nu(x_1) f(y_1|x_1) \prod_{i=1}^{T-1} p_{x_i x_{i+1}} f(y_{i+1}|x_{i+1}).$$

Accordingly, the log-likelihood function of $\boldsymbol{\theta}$ given $\boldsymbol{y}_{k,T}$ is

$$\tilde{\ell}_k\left(\boldsymbol{\theta}, \boldsymbol{y}_{k,T}\right) = \log f_{\boldsymbol{Y}_{k,T}|\boldsymbol{\theta}}\left(\boldsymbol{y}_{k,T}|\boldsymbol{\theta}\right), \forall k \geq 1.$$

For the first k sequences $\left\{\boldsymbol{y}_{i,T}\right\}_{i=1}^{k}$, let the likelihood function be

$$\ell_k\left(\boldsymbol{\theta}, \left\{\boldsymbol{y}_{i,T}\right\}_{i=1}^{k}\right) = \log f_{\left\{\boldsymbol{Y}_{i,T}\right\}_{i=1}^{k}|\boldsymbol{\theta}}\left(\left\{\boldsymbol{y}_{i,T}\right\}_{i=1}^{k}|\boldsymbol{\theta}\right), \forall k \geq 1. \tag{13}$$

We define the information matrix $I(\boldsymbol{\theta}) = \left[I_{ij}(\boldsymbol{\theta})\right]_{1 \leq i,j \leq d}$[45] where

$$I_{ij}(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}}\left[\frac{\partial \ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)}{\partial \theta_i} \cdot \frac{\partial \ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)}{\partial \theta_j}\right], \forall 1 \leq i, j \leq d,$$

Given the prior density $\pi(\boldsymbol{\theta})$, Bayes' theorem yields the posterior densities with respect to μ_Ω as (Theorem 1.31 in [166])

$$\pi_k\left(\boldsymbol{\theta}\middle|\left\{\boldsymbol{y}_{i,T}\right\}_{i=1}^{k}\right) = \frac{f\left(\left\{\boldsymbol{y}_{i,T}\right\}_{i=1}^{k}|\boldsymbol{\theta}\right)\pi(\boldsymbol{\theta})}{\int_\Omega f\left(\left\{\boldsymbol{y}_{i,T}\right\}_{i=1}^{k}|\boldsymbol{\theta}\right)\pi(\boldsymbol{\theta}) d\mu_\Omega} \propto e^{\ell_k(\theta)}\pi(\boldsymbol{\theta}), \forall k,$$

[45] The definition of this matrix is different from the Fisher information matrix defined by eq. (12.25) in [17] defined for ergodic HMM, and it follows the limiting covariance matrix in Theorem 4 in [165].

Furthermore, we use $|\cdot|$ to denote the norm of an object in the rest of the chapter: absolute value for scalars, Euclidean norm for vectors, and Frobenius norm[46] for matrices.

The following assumptions will be made.

A1. Ω is a compact set[47] in $\mathbb{R}^d$, and $\boldsymbol{\theta}_0$ is an interior point of Ω.

A2. The initial distribution $\boldsymbol{\nu}$ is known to be $v_1 = 1, \text{and}\, v_i = 0, \forall i \neq 1$. In other words, $\forall k \geq 1\; X_{k,1} = s_1$.

A3. For all $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$, the maps $\boldsymbol{\theta} \to v_x$, $\boldsymbol{\theta} \to p_{xx'}$, and $\boldsymbol{\theta} \to f_{Y|X,\boldsymbol{\theta}}(y|x, \boldsymbol{\theta})$ have continuous second derivatives on Ω.

A4. The support of $f_{Y_1|X_1,\boldsymbol{\theta}}(y|x, \boldsymbol{\theta})$ is $\mathcal{Y}$ for every $x \in \mathcal{X}$, and for all $x, \boldsymbol{\theta}, y$,there exists deterministic and integrable functions $g_1: \mathcal{Y} \to \mathbb{R}_+$ and $g_2: \mathcal{Y} \to \mathbb{R}_+$, and a positive constant M_g, such that [48]

$$g_1(y) \leq f_{Y_1|X_1,\boldsymbol{\theta}}(y|x, \boldsymbol{\theta}) \leq g_2(y)\,,$$

and

$$\int_{\mathcal{Y}} \left|\log\big(g_i(y)\big)\right| g_2(y) d\mu_y \leq M_g < \infty, i = 1 \text{ or } 2.$$

A5. For all $1 \leq i \leq d$ and all $x \in \mathcal{X}$, there exists a function $g_3: \mathcal{Y} \to \mathbb{R}_+$ and $\delta > 0$ such that

$$\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0| \leq\, \delta} \left|\frac{\partial}{\partial \theta_i} \log f_{Y_1|X_1,\boldsymbol{\theta}}\,(y|x, \boldsymbol{\theta})\right| \leq g_3(y)$$

and

$$\int_{\mathcal{Y}} g_3(y)\, g_2(y) d\mu_y < \infty, \text{and} \int_{\mathcal{Y}} g_3(y)^2 g_2(y)\, d\mu_y < \infty.$$

[46] The Frobenious norm has the property that $|A \cdot B| \leq |A| \cdot |B|$ for all multipliable matrices A and B.

[47] Study of the case when Ω is non-compact is out of the scope of this thesis.

[48] The conditional density functions $f_{Y_1|X_1,\Theta}$ in A4 to A6 are for every sequence and thus the sequence subscript for X, Y are omitted.

A6. For all $1 \leq i, j \leq d$ and all $x \in \mathcal{X}$, there exists a function $g_4 : \mathcal{Y} \to \mathbb{R}_+$ and $\delta > 0$ such that

$$\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0| \leq \delta} \left| \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{Y_1|X_1,\Theta}(y|x, \boldsymbol{\theta}) \right| \leq g_4(y),$$

and

$$\int_{\mathcal{Y}} g_4(y) g_2(y) \, d\mu_y < \infty.$$

A7. For each $\boldsymbol{\theta}$ for which the distributions $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_0$ agree almost everywhere on the product observation space, it must be true that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (up to a possible permutation of states).

A8. For the components $\{\phi_i\}_{i=1}^{N}$ of ϕ, each of which is a parameter for a different conditional density function $f_{Y_1|X_1,\Theta}(y|s, \boldsymbol{\theta})$, e.g. $\phi_i = \mu_{s_i}$ is the conditional mean given the state s_i, $\{\phi_i\}_{i=1}^{N}$ satisfies the order constraint, i.e.,

$$\phi_i < \phi_j, \forall 1 \leq i < j \leq N.$$

A9. The prior density π is positive and continuous for all $\boldsymbol{\theta} \in \Omega$.

A10. $I(\boldsymbol{\theta}_0)$ is nonsingular and finite.

Remarks regarding assumptions A1-A10:

- **Assumption A1** is assumed to avoid pathological cases, such as when $\boldsymbol{\theta}_0$ is on the boundary of Ω or $\mathbb{R}^d$.
- **Assumption A2** on the initial state distribution of each sequence is similar to what we see in several ergodic HMM studies [10], [128].
- **Assumptions A3-A6** are adapted from [128], and these conditions regulate the boundedness of the likelihood functions, as well as their first and second order derivatives. Similar conditions can be found in theorems for consistency and asymptotic normality of MLE of HMM parameters in [10], [11], [18].

- **Assumptions A7** and **A8** together ensure the true parameters $\boldsymbol{\theta}_0$ can be uniquely identified from $\mathcal{X}^T \times \mathcal{Y}^T$. Examples of conditional distributions that satisfy **Assumption A7** are Gaussian and Poisson distributions. The order constraint in **Assumption A8** is mentioned in [128], [133], and [17], and allows unique labeling of the hidden states.
- **Assumption A9** pertains to the prior distribution and is typically assumed in the formulation of BvMT, including those for i.i.d. models [165] and ergodic HMMs [18].
- **Assumption A10** ensures the finiteness of the limiting covariance matrix of the normal approximation of the posterior distribution of $\boldsymbol{\theta}$.

The following theorem corresponds to the scenario when observation sequences of equal length T are collected. In the context of machine monitoring, this corresponds to the situation when a fixed-schedule Preventive Maintenance (PM) policy is implemented and each sequence $\boldsymbol{y}_{i,T}$ terminates when a scheduled PM action is executed [49].

[49] This means the sample space for each process is $(\mathcal{Y}^T, \mathcal{B}_Y^{\otimes T}, \mu_Y^T)$, where the joint density $f_{Y_T|\boldsymbol{\Theta}}(\boldsymbol{y}_T|\boldsymbol{\theta})$is defined on.

Theorem 1 (PM-inspired Bernstein-von Mises Theorem) *Let*

$$\boldsymbol{\tau}_k = \boldsymbol{\theta}_0 + \frac{1}{k} I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k(\boldsymbol{\theta}_0), \forall k > 0,$$

and let $\pi_k^*\left(\boldsymbol{u}\middle|\{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)$ *be the posterior density of* $\boldsymbol{u} = \sqrt{k}(\boldsymbol{\theta} - \boldsymbol{\tau}_k)$. *Assume A1-A10 hold, then for any* $\epsilon > 0$

$$\lim_{k\to\infty} \mathbb{P}_0^\infty\left(\int_{\mathbb{R}^d}\left|\pi_k^*\left(\boldsymbol{u}\middle|\{\boldsymbol{y}_{i,T}\}_{i=1}^k\right) - \frac{1}{\sqrt{|2\pi I^{-1}(\boldsymbol{\theta}_0)|}} e^{-\boldsymbol{u}^T I(\boldsymbol{\theta}_0)\,\boldsymbol{u}/2}\right| d\boldsymbol{u} \le \epsilon\right) = 1.$$

Outline of the proof:

To prove the above-formulated BvMT, a prerequisite is the establishment of several limiting properties of lower order derivatives for the log-likelihood functions, such as those for ergodic HMMs [10], [11]. We will prove those properties for left-to-right HMMs in Lemmas 6-11, and then follow Bickel's strategy from [165] by showing through Lemmas 12-14 the desired convergence in L_1 distance.

Remark:

The BvMT formulated above claims that under certain regularity conditions, a sequence of normalized posterior distributions given a set of observation sequences $\{\boldsymbol{y}_{i,T}\}_{i=1}^k$ converges to a fixed Gaussian distribution in total variation distance. In this limiting process, the centers of the posterior distributions are asymptotically efficient, while the posterior standard deviation decreases to 0 at the rate of $1/\sqrt{k}$. The resulting asymptotic normal approximation to the posterior distributions enables analysis of the dependency of model uncertainty on the number of observation sequences, allowing solution to the

Sample Size Determination (SSD) problems [169], which is of utmost important for condition-monitoring problems.

5.2 Proof

The following two propositions are already proven results from the literature.

Proposition 1: (Theorem 2 in [168]) Let $\boldsymbol{X}$ be a random vector defined on a probability space (Ω, Σ, μ), where $\boldsymbol{X}$ induces a law $\mathbb{P}_{\boldsymbol{X}}$ on a measurable space $(\mathbb{R}^n, \mathcal{B}^{\otimes n})$. Let g be a real-valued function on $\mathbb{R}^n \times \Theta$. Assume that

(i) Θ is a compact subset of $\mathbb{R}^m$.

(ii) $g(\boldsymbol{x}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ given each $\boldsymbol{x}$, and is a measurable function of $\boldsymbol{x}$ given each $\boldsymbol{\theta}$.

(iii) $|g(\boldsymbol{x}, \boldsymbol{\theta})| \leq h(\boldsymbol{x})$ for almost every $\boldsymbol{x}$ in $\mathbb{R}^n$, where $h(\boldsymbol{x})$ is measureable and has finite expectation.

Then the following holds.

$$\mathbb{P}_X^{\infty}\left(\lim_{k\to\infty} \sup_{\boldsymbol{\theta}\in\Omega} \left|\frac{1}{k}\sum_{i=1}^{k} g(\boldsymbol{x}_i, \boldsymbol{\theta}) \ - \mathbb{E}_X[g(\boldsymbol{x}, \boldsymbol{\theta})]\right| = 0\right) = 1.$$

Proposition 2: (Theorem 16.8 in [171]) Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose $g(\omega, \theta)$ is a real valued function on the Cartesian product space $\Omega \times (a, b)$, where (a, b) is a finite open interval in $\mathbb{R}$. Assume g satisfies the following:

(i) For each fixed $\theta \in (a, b)$, the function $g(\omega, \theta)$ is a Borel function, i.e., measurable w.r.t. the Borel sigma-algebra, and

$$\int |g(\omega, \theta)| d\omega < \infty.$$

(ii) There's a null-set $\mathcal{N}$ such that for all $\omega \neq \mathcal{N}$, the derivative $\partial g(\omega, \theta)/\partial\theta$ exists for all $\theta \in (a, b)$.

(iii) There is an integrable function $G: \Omega \to \mathbb{R}$ such that for all $\omega \notin \mathcal{N}$ and all (a, b)

$$\left|\frac{\partial g}{\partial \theta}(\omega, \theta)\right| \leq G(\omega).$$

Then for each fixed $\theta \in (a, b)$, $\partial g(\omega, \theta)/\partial\theta$ is integrable w.r.t. μ and the derivative and integration are interchangeable, i.e.,

$$\frac{dg}{d\theta}\int_{\Omega} g(\omega, \theta)\, d\mu(\omega) = \int_{\Omega} \frac{dg}{d\theta}(\omega, \theta)\, d\mu(\omega).$$

The notations in each of the Lemmas 1-4 are self-contained, and the results in those Lemmas are generic in the sense that they are not restricted to the usage for proving HMM-related properties.

Lemma 1: Consider two parameterized sequences of random vectors $\{X_n(t)\}$ where $X_n(t) \in \mathbb{R}^{m\times l}$ and $\{Y_n(t)\}$, where $Y_n(t) \in \mathbb{R}^{l\times k}$. If $\sup_{t\in\tau}|X_n(t) - C| \overset{\mathbb{P}}{\to} 0$ and $\sup_{t\in\tau}|Y_n(t) - D| \overset{\mathbb{P}}{\to} 0$ as $n \to \infty$, where C and D are two constant vectors, then
(i) $\sup_{t\in\tau}|X_n(t) \cdot Y_n(t) - C \cdot D| \overset{\mathbb{P}}{\to} 0$;
(ii) $\sup_{t\in\tau}|X_n(t) + Y_n(t) - C - D| \overset{\mathbb{P}}{\to} 0$.

Proof: (i) First notice that the assumed uniform convergence in probability of $Y_n(t)$ implies uniform boundedness in probability of $Y_n(t)$, i.e., for any positive $\epsilon > 0$, there exists $N \in \mathbb{N}$ and $M \in \mathbb{R}_+$ such that for all $n \geq N$, we have $\mathbb{P}(\sup_{t\in\tau}|Y_n(t)| \leq M) > 1 - \epsilon$.

Assuming $C \neq 0$, then for any positive $\epsilon_1, \epsilon_2 > 0$, there exists $N_1, N_2, N_3 \in \mathbb{N}$ and $M \in \mathbb{R}_+$ such that $\mathbb{P}(\sup_{\mathrm{t}\in\tau}|Y_n(t)| \leq M) \geq 1 - \epsilon_2$, $\mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|X_n(t) - C| \leq \frac{\epsilon_1}{2M}\right) \geq 1 - \epsilon_2$, and $\mathbb{P}\left(\sup_{t\in\tau}|Y_n(t) - D| \leq \frac{\epsilon_1}{2|C|}\right) \geq 1 - \epsilon_2$. Hence, for all $n \geq \max\{N_1, N_2, N_3\}$,

$$\mathbb{P}(\sup_{\mathrm{t}\in\tau}|X_n(t) \cdot Y_n(t) - C \cdot D| \leq \epsilon_1)$$

$$\geq \mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|X_n(t) - C| \cdot |Y_n(t)| \leq \frac{\epsilon_1}{2} \text{ and } \sup_{\mathrm{t}\in\tau}|Y_n(t) - D| \cdot |C| \leq \frac{\epsilon_1}{2}\right)$$

$$\geq \mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|X_n(t) - C| \leq \frac{\epsilon_1}{2M}, \sup_{\mathrm{t}\in\tau}|Y_n(t)| \leq M, \sup_{\mathrm{t}\in\tau}|Y_n(t) - D| \cdot |C| \leq \frac{\epsilon_1}{2}\right)$$

$$\geq \mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|X_n(t) - C| \leq \frac{\epsilon_1}{2M}\right) + \mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|Y_n(t)| \leq M\right) + \mathbb{P}\left(\sup_{t\in\tau}|Y_n(t) - D| \leq \frac{\epsilon_1}{2|C|}\right) - 2$$

$$\geq 1 - \epsilon_2.$$

The case when $C = 0$ can be proven similarly.

(ii) For all $\epsilon_1, \epsilon_2 > 0$, there exists $N_1, N_2 \in \mathbb{N}$ such that for $n \geq N_1$, $\mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|X_n(t) - C| \leq \frac{\epsilon_1}{2}\right) \geq 1 - \frac{\epsilon_2}{2}$, and for $n \geq N_2$, $\mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|Y_n(t) - D| \leq \frac{\epsilon_1}{2}\right) \geq 1 - \frac{\epsilon_2}{2}$. It follows that

$$\mathbb{P}(\sup_{\mathrm{t}\in\tau}|X_n(t) + Y_n(t) - C - D| \leq \epsilon_1)$$

$$\geq \mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|X_n(t) - C| \leq \frac{\epsilon_1}{2} \text{ and } \sup_{\mathrm{t}\in\tau}|Y_n(t) - D| \leq \frac{\epsilon_1}{2}\right)$$

$$\geq \mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|X_n(t) - C| \leq \frac{\epsilon_1}{2}\right) + \mathbb{P}\left(\sup_{\mathrm{t}\in\tau}|Y_n(t) - D| \leq \frac{\epsilon_1}{2}\right) - 1$$

$$\geq 1 - \epsilon_2,$$

Q.E.D. □

Lemma 2: If a sequence of parameterized random vectors $\{X_n(\theta)\}$ converges to a constant vector C uniformly in probability on a compact set Θ in a Euclidean space $\mathbb{R}^n$, i.e., if $\lim_{n\to\infty} \mathbb{P}(\sup_{\theta\in\Theta} |X_n(\theta) - C\,| < \epsilon) = 1 \;\; \forall\epsilon > 0$, and if $g(x)$ is a continuous real-valued function in x, then $\forall\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\sup_{\theta\in\Theta} |g(X_n(\theta)) - g(C)\,| < \epsilon) = 1.$$

Proof: For any positive M such that $-M < C < M$, $g(x)$ is uniformly continuous on $[-M, M]$. Thus, for any $\delta > 0$, there exists $\epsilon > 0$, such that for all $|x_1 - x_2| \leq \epsilon$, there is $|g(x_1) - g(x_2)| \leq \delta$. It follows that for any $\theta \in \Theta$,

$$\mathbb{P}(|X_n(\theta) - c| \leq \epsilon) \leq \mathbb{P}(|g(X_n(\theta)) - g(c)| \leq \delta),$$

and therefore

$$\mathbb{P}(\sup_{\theta\in\Theta}|X_n(\theta) - c| \leq \epsilon) \leq \mathbb{P}(\sup_{\theta\in\Theta}|g(X_n(\theta)) - g(c)| \leq \delta).$$

Then the right hand side converges to 1 as well, Q.E.D. □

Lemma 3: If a sequence of random matrices $\{Y_n\}$ is bounded in $\mathbb{P}$-probability and if $\{C_n\}$ is a sequence of random matrices tending to 0 in $\mathbb{P}$-probability, then $C_nY_n \overset{\mathbb{P}}{\rightarrow} 0$, assuming C_nY_n are compatible in matrix multiplication.

Proof: For any $\epsilon_1, \epsilon_2 > 0$, there exists N and M such that for all $n > N$,

$$\mathbb{P}(|Y_n| \leq M) \geq 1 - \tfrac{\epsilon_2}{2}, \text{and } \mathbb{P}\left(|C_n| \leq \tfrac{\epsilon_1}{M}\right) \geq 1 - \tfrac{\epsilon_2}{2},$$

and hence,

$$\mathbb{P}(|C_n \cdot Y_n| \le \epsilon_1) \ge \mathbb{P}\left(|C_n| \le \tfrac{\epsilon_1}{M} \; and \; |Y_n| \le M\right)$$

$$\ge \mathbb{P}\left(|C_n| \le \tfrac{\epsilon_1}{M}\right) + \mathbb{P}(|Y_n| \le M) - 1$$

$$\ge 1 - \epsilon_2,$$

Q.E.D. □

Lemma 4: Suppose a sequence of random vectors $\{Y_n\}$ is bounded by a sequence of nonnegative numbers $\{f_n\}$ with probability 1, i.e., $\lim_{n\to\infty} \mathbb{P}(|Y_n| \le f_n) = 1$. Then $\lim\limits_{n\to\infty} f_n = 0$ implies $Y_n \overset{\mathbb{P}}{\to} 0$.

Proof: For any $\epsilon_1, \epsilon_2 > 0$, there exists N_1 such that for all $n > N_1$, $0 \le f_n \le \epsilon_1$, and N_2 such that for all $n > N_2$, $\mathbb{P}(|Y_n| \le f_n) \ge 1 - \epsilon_2$. Therefore, for all $n \ge \max(N_1, N_2)$,

$$\mathbb{P}(|Y_n| \le \epsilon_1) \ge \mathbb{P}(|Y_n| \le f_n) \ge 1 - \epsilon_2,$$

which proves the lemma. □

Lemma 5: Define

$$\omega_k(\boldsymbol{u}) = \ell_k\left(\tfrac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right) - \ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)$$

$$-\tfrac{1}{2k}\, \nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)^{\mathrm{T}} I^{-1}(\boldsymbol{\theta}_0) \nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)$$

where ℓ_k is defined by (13) and

$$R_k(\boldsymbol{\theta}) = -\nabla^2\ell_k\left(\boldsymbol{\theta}_k^*, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right) - kI\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right),$$

where $\boldsymbol{\theta}_k^*$ is a point between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ that satisfies[50]

[50] This point exists per (14) being a Taylor series expansion.

$$\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)=\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)+(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)$$

$$+\tfrac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\nabla^2\ell_k\left(\boldsymbol{\theta}_k^*,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)(\boldsymbol{\theta}-\boldsymbol{\theta}_0). \tag{14}$$

Then, we have the following equality

$$\omega_k(\boldsymbol{u})=-\frac{\boldsymbol{u}^T I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2}-\frac{1}{2k}\left(\boldsymbol{u}+\frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)^{\mathrm{T}}R_k\left(\frac{\boldsymbol{u}}{\sqrt{k}}+\boldsymbol{\tau}_k\right)$$

$$\left(\boldsymbol{u}+\tfrac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right). \tag{15}$$

Proof:

$$\omega_k(\boldsymbol{u})$$

$$=\left(\tfrac{\boldsymbol{u}}{\sqrt{k}}+\tfrac{1}{k}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)^{\mathrm{T}}\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)$$

$$-\tfrac{1}{2}\left(\tfrac{\boldsymbol{u}}{\sqrt{k}}+\tfrac{1}{k}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)^{\mathrm{T}}\cdot$$

$$\left(R_k\left(\tfrac{\boldsymbol{u}}{\sqrt{k}}+\boldsymbol{\tau}_k\right)+kI(\boldsymbol{\theta}_0)\right)\left(\tfrac{\boldsymbol{u}}{\sqrt{k}}+\tfrac{1}{k}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)$$

$$-\tfrac{1}{2k}\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)^{\mathrm{T}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)$$

$$=\tfrac{\boldsymbol{u}^{\mathrm{T}}\nabla\ell_k(\boldsymbol{\theta}_0)}{\sqrt{k}}+\tfrac{1}{k}\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)^{\mathrm{T}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)$$

$$-\tfrac{1}{2k}\left(\boldsymbol{u}+\tfrac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)^{\mathrm{T}}R_k\left(\tfrac{\boldsymbol{u}}{\sqrt{k}}+\boldsymbol{\tau}_k\right)$$

$$\left(\boldsymbol{u}+\tfrac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)$$

$$-\tfrac{1}{2}\left(\boldsymbol{u}+\tfrac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)^{\mathrm{T}}I(\boldsymbol{\theta}_0)$$

$$\left(\boldsymbol{u}+\tfrac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)$$

$$-\tfrac{1}{2k}\,\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)^{\mathrm{T}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)$$

$$=-\frac{\boldsymbol{u}^{\mathrm{T}}I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2}-\tfrac{1}{2k}\left(\boldsymbol{u}+\tfrac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right)^{\mathrm{T}}R_k\left(\tfrac{\boldsymbol{u}}{\sqrt{k}}+\boldsymbol{\tau}_k\right)$$

$$\left(\boldsymbol{u}+\tfrac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)\right),$$

Q.E.D. □

Lemma 6: Assume A1-A10 hold, then

$$\mathbb{P}_0^{\infty}\left(\lim_{k\to\infty}\sup_{\boldsymbol{\theta}\in\Omega}\left|\tfrac{1}{k}\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\mathbb{E}_0[\ell_1(\boldsymbol{\theta},\boldsymbol{y}_T)]\right|=0\right)=1,$$

Proof: We will show that conditions of Proposition 1 are satisfied, which directly leads to the proof. Consider the log-likelihood function

$$\ell_1(\boldsymbol{\theta},\boldsymbol{y}_T)=\log\left(\sum_{\boldsymbol{x}_T\in\mathcal{X}^T}\prod_{i=1}^{T-1}\nu_{x_1}f(y_1|x_1)p_{x_ix_{i+1}}f(y_{i+1}|x_{i+1})\right),\tag{16}$$

which is defined on $\Omega\times\mathbb{R}^T$. Prop. 1 (i) is clear by A1. On the right hand side of (16), all initial probabilities ν_{s_i}, transition probabilities $p_{x_ix_{i+1}}$, and conditional densities $f(y|s_i)$ are continuous function on Ω by A3, and are therefore all measurable functions as probability density or mass functions. Since continuity and measurability are preserved under algebraic operations [172], the likelihood function satisfies Prop. 2 (ii).

Let us now show the boundedness of $\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)$. When $\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T) \leq 0$, we have

$$
\begin{aligned}
|\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)| &= \left|\log \textstyle\sum_{\boldsymbol{x}_T \in \mathcal{X}^T} \nu_{x_1} f(y_1|x_1) \prod_{i=1}^{T-1} p_{x_i x_{i+1}} f(y_{i+1}|x_{i+1})\right| \\
&\leq \left|\log \textstyle\sum_{\boldsymbol{x}_T \in \mathcal{X}^T} \nu_{x_1} \left(\prod_{i=1}^{T-1} p_{x_i x_{i+1}}\, g_1(y_i)\right)\right| \\
&\leq \textstyle\sum_{i=1}^{T}(|\log g_1(y_i)| + |\log g_2(y_i)|)\,,
\end{aligned}
$$

When $\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T) > 0$,

$$
\begin{aligned}
|\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)| &= \left|\log \textstyle\sum_{\boldsymbol{x}_T \in \mathcal{X}^T} \nu_{x_1} f(y_1|x_1) \prod_{i=1}^{T-1} p_{x_i x_{i+1}} f(y_{i+1}|x_{i+1})\right| \\
&\leq \left|\log \textstyle\sum_{\boldsymbol{x}_T \in \mathcal{X}^T} \nu_{x_1} \prod_{i=1}^{T} p_{x_i x_{i+1}} g_2(y_i)\right| \\
&\leq \textstyle\sum_{i=1}^{T}(|\log g_1(y_i)| + |\log g_2(y_i)|).
\end{aligned}
$$

In either case, the log-likelihood function is then bounded by

$$
M_\ell(\boldsymbol{y}_T) = \textstyle\sum_{i=1}^{T}(|\log g_1(y_i)| + |\log g_2(y_i)|).
$$

Since $\mathbb{E}_0[M_\ell] \leq T \cdot 2M_g < \infty$ according to A4, $M_\ell(\boldsymbol{y}_T)$ is $\mathbb{P}_0$-integrable and therefore, Prop. 1 (iii) also holds, Q.E.D. □

Lemma 7: Assume A1-A10 hold, then $\mathbb{E}_0[\ell_1(\boldsymbol{\theta})]$ is a continuous function in $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \Omega$. Furthermore, define Kullback-Leibler information

$$
\mathcal{K}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \mathbb{E}_0\left[\log \frac{f(\boldsymbol{y}_T|\boldsymbol{\theta}_0)}{f(\boldsymbol{y}_T|\boldsymbol{\theta})}\right].
$$

Then $\mathcal{K}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \geq 0$ with equality if only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Proof: This proof follows [128]. It is known from Lemma 6 that

$$\mathbb{E}_0[\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)] \leq \mathbb{E}_0[M_\ell] < \infty.$$

For any $\boldsymbol{y}_T$, any $\boldsymbol{\theta}_* \in \Omega$, and any sequence converging sequence $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_*$, the continuity of $\ell_1(\boldsymbol{\theta})$ guarantees that $\ell_1(\boldsymbol{\theta}_n, \boldsymbol{y}_T) \to \ell_1(\boldsymbol{\theta}_*, \boldsymbol{y}_T)$. Then the continuity of $\mathbb{E}_0[\ell_1(\boldsymbol{\theta})]$ is given by

$$\lim_{n\to\infty} \int \ell_1(\boldsymbol{\theta}_n, \boldsymbol{y}_T) f(\boldsymbol{y}_T|\boldsymbol{\theta}_0) d\boldsymbol{y}_T = \int \lim_{n\to\infty} \ell_1(\boldsymbol{\theta}_n, \boldsymbol{y}_T) f(\boldsymbol{y}_T|\boldsymbol{\theta}_0)\, d\boldsymbol{y}_T$$

$$= \int \ell_1(\boldsymbol{\theta}_*, \boldsymbol{y}_T) f(\boldsymbol{y}_T|\boldsymbol{\theta}_0) d\boldsymbol{y}_T$$

as a result of Lebesgue's Dominated Convergence Theorem.

Due to Jensen's inequality,

$$-\mathcal{K}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \leq \log 1 = 0,$$

and $\mathcal{K}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = 0$ if and only if $f(\boldsymbol{y}_T|\boldsymbol{\theta}) = f(\boldsymbol{y}_T|\boldsymbol{\theta_0})$ almost everywhere on $\mathcal{Y}^T$, which implies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Q.E.D. □

Lemma 8: Assume A1-A10 hold, then for any $\delta > 0$ there exists $\epsilon > 0$ such that

$$\lim_{k\to\infty} \mathbb{P}_0^\infty \left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\delta} \frac{1}{k} \left(\ell_k \left(\boldsymbol{\theta}, \{\boldsymbol{y}_{i,T}\}_{i=1}^k \right) - \ell_k \left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k \right) \right) \leq -\epsilon \right) = 0.$$

Proof: This proof follows the line of [18]. First we show that there exists a positive constant C such that

$$\mathrm{Sup}_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\delta}\, \mathbb{E}_0[\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)] - \mathbb{E}_0[\ell_1(\boldsymbol{\theta}_0, \boldsymbol{y}_T)] \leq -C < 0. \tag{17}$$

Suppose such C does not exist, then for any sequence $\{C_n\}_{n=1}^\infty$ such that $\lim_{n\to\infty} C_n = 0$, there is a sequence $\{\boldsymbol{\theta}_n\}_{n=1}^\infty$ from $\{\boldsymbol{\theta}: |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \delta\}$ such that

$$\mathbb{E}_0[\ell_1(\boldsymbol{\theta}_n, \boldsymbol{y}_T)] - \mathbb{E}_0[\ell_1(\boldsymbol{\theta}_0, \boldsymbol{y}_T)] > -C_n.$$

This sequence has a limit point θ_* due to the compactness of $\{\boldsymbol{\theta}: |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \delta\}$.

Following continuity of $\mathbb{E}_0[\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)]$ proven in Lemma 7, we have $\lim_{n\to\infty} \mathbb{E}_0[\ell_1(\boldsymbol{\theta}_n, \boldsymbol{y}_T)] = \mathbb{E}_0[\ell_1(\boldsymbol{\theta}_*, \boldsymbol{y}_T)] \geq \mathbb{E}_0[\ell_1(\boldsymbol{\theta}_0, \boldsymbol{y}_T)]$. This, however, contradicts with the fact that $\boldsymbol{\theta}_0$ is the unique maximum of $E_0[\ell_1(\boldsymbol{\theta})]$, as per by Lemma 7, and therefore C exists for (17). Now, following Lemma 6, for any $\epsilon > 0$ there exits K_1 such that for all $k > K_1$

$$\mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\delta}\left|\frac{1}{k}\left(\ell_k\left(\boldsymbol{\theta}, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right) - \mathbb{E}_0[\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)]\right)\right| \leq \frac{C}{4}\right) \geq 1 - \frac{\epsilon}{2}, \qquad (18)$$

and there exists K_2 such that for all $k > K_2$,

$$\mathbb{P}_0^\infty\left(\left|\frac{1}{k}\left(\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right) - \mathbb{E}_0[\ell_1(\boldsymbol{\theta}_0, \boldsymbol{y}_T)]\right]\right| \leq \frac{C}{4}\right) \geq 1 - \frac{\epsilon}{2}, \qquad (19)$$

Combining (17), (18), and (19), we have

$$\mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\delta}\frac{1}{k}\left(\ell_k\left(\boldsymbol{\theta}, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right) - \ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\right) \leq -\frac{C}{2}\right)$$

$$= \mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\delta}\left[\left(\frac{1}{k}\ell_k\left(\boldsymbol{\theta}, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right) - \mathbb{E}_0[\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)]\right) + \left(\mathbb{E}_0[\ell_1(\boldsymbol{\theta}_0, \boldsymbol{y}_T)] - \right.\right.\right.$$

$$\left.\left.\left.\frac{1}{k}\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\right) + (\mathbb{E}_0[\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)] - \mathbb{E}_0[\ell_1(\boldsymbol{\theta}_0, \boldsymbol{y}_T)])\right] \leq -\frac{C}{2}\right)$$

$$\geq \mathbb{P}_0^\infty\left(\sup_{|\theta-\theta_0|\geq\delta}\frac{1}{k}\left|\ell_k\left(\boldsymbol{\theta}, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right) - \mathbb{E}_0[\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)]\right| \leq \frac{C}{4} \text{ and } \frac{1}{k}\left|\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right) - \right.\right.$$

$$\left.\left.\mathbb{E}_0[\ell_1(\boldsymbol{\theta}_0, \boldsymbol{y}_T)]\right| \leq \frac{C}{4} \text{ and } \sup_{|\theta-\theta_0|\geq\delta}(\mathbb{E}_0[\ell_1(\boldsymbol{\theta}, \boldsymbol{y}_T)] - \mathbb{E}_0[\ell_1(\boldsymbol{\theta}_0, \boldsymbol{y}_T)]) \leq -C\right)$$

$$\geq 1 - \epsilon,$$

Q.E.D. □

Lemma 9: Assume A1-A10 hold, then

(i) $\frac{1}{\sqrt{k}}\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right) \xrightarrow{\mathcal{L}} N(0, I(\boldsymbol{\theta}_0))$.

(ii) $\frac{1}{k}\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right) \xrightarrow{\mathbb{P}_0^\infty} 0$.

(iii) The centering sequence $\{\boldsymbol{\tau}_k\}$ defined by $\boldsymbol{\tau}_\mathrm{k} = \boldsymbol{\theta}_0 + \frac{1}{\mathrm{k}} I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_\mathrm{k}(\boldsymbol{\theta}_0)$ satisfies

$$\lim_{k\to\infty} \mathbb{P}_0^\infty(|\boldsymbol{\tau}_k - \boldsymbol{\theta}_0| \le \epsilon) = 1,$$

for any $\epsilon > 0$.

Proof: (i) This proof has similar content as in Lemma 5.1 from [133]. By assumption A10, $I(\boldsymbol{\theta}_0)$ is finite, and it remains to show that $\mathbb{E}_0[\nabla\ell_1(\boldsymbol{\theta}_0)] = 0$, which can be proven as the consequence of

$$\int_{\mathcal{Y}^T} \frac{\partial}{\partial\theta_i} f(\boldsymbol{y}_T|\boldsymbol{\theta}_0) d\mu_Y^T = \frac{\partial}{\partial\theta_i}\int_{\mathcal{Y}^T} f(\boldsymbol{y}_T|\boldsymbol{\theta}_0) d\mu_Y^T = 0, \forall 1 \le i \le d. \qquad (20)$$

by the Weak Law of Large Numbers.

To show (20), we will verify that $f(\boldsymbol{y}_T|\boldsymbol{\theta})$, which is constrained on $\mathcal{Y}^T \times \{\boldsymbol{\theta}: |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \le \delta\}$, where $\delta = \min\{\delta_0/2, (1-\delta_0)/2\}$, satisfies the three conditions in Proposition 2. For Prop. 2 (i), $f(\boldsymbol{y}_T|\boldsymbol{\theta})$ for each fixed $\boldsymbol{\theta}$ is a density function that is integrable and is measurable with respect to $\mathcal{B}_Y^T$. As for Prop. 2 (ii), considering A4 and eq. (16), $f(\boldsymbol{y}_T|\boldsymbol{\theta})$ is continuously differentiable w.r.t each $\theta_i, 1 \le i \le d$, and therefore Prop 2. (ii) is satisfied. For Prop. 2 (iii), define the set of all probable state sequences of length T by

$$\mathcal{X}_+^T \coloneqq \{\boldsymbol{x}_T \in \mathcal{X}^T : \mathbb{P}_0(x_T) > 0\}.$$

Then let $\xi_1(\boldsymbol{x}_T) = \nu(x_1) f(y_1|x_1)$ and $\xi_t(\boldsymbol{x}_T) = P_{x_{t-1}x_t} f(y_t|x_t), \forall 2 \le t \le T$, so that $f(\boldsymbol{y}_T|\boldsymbol{\theta}) = \sum_{\boldsymbol{x}_T \in \mathcal{X}_+^T} \prod_{t=1}^T \xi_t(\boldsymbol{x}_T)$ and that

$$\frac{\partial f(\boldsymbol{y}_T|\boldsymbol{\theta})}{\partial \theta_i} = \sum_{\boldsymbol{x}_T \in \mathcal{X}_+^T} \sum_{t=1}^{T} \frac{\partial \xi_t(\boldsymbol{x}_T)}{\partial \theta_i} \cdot \prod_{t' \neq t} \xi_{t'}(\boldsymbol{x}_T). \tag{21}$$

Note that

$$\left|\frac{\partial \xi_1}{\partial \theta_i}\right| = \left|\frac{\partial \xi_1/\partial \theta_i}{\xi_1}\xi_1\right| = \left|\left[\frac{\partial v_{x_1}/\partial \theta_i}{v_{x_1}} + \frac{\partial f(y_1|x_1)/\partial \theta_i}{f(y_1|x_1)}\right]\xi_1\right| \leq [C_P + g_3(y_1)]\xi_1,$$

and for $\forall t, 2 \leq t \leq T$,

$$\left|\frac{\partial \xi_t}{\partial \theta_i}\right| = \left|\frac{\partial \xi_t/\partial \theta_i}{\xi_t}\xi_t\right| = \left|\left[\frac{\partial P_{x_{t-1}x_t}/\partial \theta_i}{P_{x_{t-1}x_t}} + \frac{\partial f(y_t|x_t)/\partial \theta_i}{f(y_t|x_t)}\right]\xi_t\right| \leq [C_P + g_3(y_t)]\xi_t,$$

where $C_P > \max\{2/\delta_0,\ 2/(1-\delta_0)\}$. It follows that for all $\boldsymbol{y}_T$ and all $\boldsymbol{\theta} \in \{\boldsymbol{\theta}: |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \leq \delta\}$,

$$\begin{aligned}\left|\frac{\partial f(\boldsymbol{y}_T|\boldsymbol{\theta})}{\partial \theta_i}\right| &\leq \sum_{\boldsymbol{x}_T \in \mathcal{X}_+^T} \sum_{t=1}^{T} [C_P + g_3(y_t)] \prod_{t'=1}^{T} \xi_{t'}(\boldsymbol{x}_T)\,, \\ &= \sum_{\boldsymbol{x}_T \in \mathcal{X}_+^T} \mathbb{P}_0(\boldsymbol{x}_T) \sum_{t=1}^{T} [C_P + g_3(y_t)] \prod_{t'=1}^{T} f(y_{t'}|x_{t'}) \\ &\leq \sum_{t=1}^{T} [C_P + g_3(y_t)] \prod_{t'=1}^{T} g_2(y_{t'}) \end{aligned} \tag{22}$$

where the right-hand side is integrable based on Fubini's theorem and Assumption A5. Hence, Prop 2. (iii) holds and therefore the sufficient condition (20) is satisfied.

As for condition (ii), the corresponding convergence in probability is implied by (i) according to [173]. Finally, when it comes to (iii), this is obvious according to $\boldsymbol{\tau}_{\mathrm{k}} = \boldsymbol{\theta}_0 + \frac{1}{\mathrm{k}}\mathrm{I}^{-1}(\boldsymbol{\theta}_0)\nabla \ell_{\mathrm{k}}(\boldsymbol{\theta}_0)$, Q.E.D. □

Lemma 10: Assume A1-A10 hold, then for any $\epsilon > 0$,

$$\lim_{k\to\infty} \mathbb{P}_0^{\infty}\left(\left|\frac{1}{k}\nabla^2 \ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right) + I(\boldsymbol{\theta}_0)\right| \leq \epsilon\right) = 1.$$

Proof: By the Weak Law of Large Numbers, it is sufficient to show that

$$\int_{\mathcal{Y}^T} \frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\boldsymbol{y}_T|\boldsymbol{\theta}_0) d\mu_Y^T = \frac{\partial}{\partial\theta_j} \int_{\mathcal{Y}^T} \frac{\partial}{\partial\theta_i} f(\boldsymbol{y}_T|\boldsymbol{\theta}_0) d\mu_Y^T = 0, \forall i,j, 1 \leq i,j \leq d. \quad (23)$$

so that $\mathbb{E}_0[-\nabla^2\ell_1(\boldsymbol{\theta}_0)] = \mathbb{E}_0[\nabla\ell_1(\boldsymbol{\theta}_0)\nabla\ell_1(\boldsymbol{\theta}_0)^{\mathrm{T}}] = I(\boldsymbol{\theta}_0) < \infty$. Let us check for the three conditions in Proposition 2 for $\frac{\partial}{\partial\theta_i} f(\boldsymbol{y}_T|\boldsymbol{\theta})$. It is integrable by (22) and is measurable, which is implied by the fact that $f(\boldsymbol{y}_T|\boldsymbol{\theta})$ is measurable, and hence (i) holds. Due to (21) and A3, (ii) also holds because $\partial^2 f(\boldsymbol{y}_T|\boldsymbol{\theta})/\partial\theta_i\partial\theta_j$ exists almost everywhere on $\{\boldsymbol{\theta}: |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \leq \delta\}$ for every $\boldsymbol{y}_T$. As for condition (iii), let $\mathcal{X}_+^T$, ξ_t, and C_P be defined in the same way as in the proof of Lemma 9. One immediately obtains that

$$\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\xi_t\right| \leq \left(g_4(y_t) + \left(C_P + g_3(y_t)\right)^2\right) \cdot \xi_t$$

and then,

$$\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j} f(\boldsymbol{y}_T|\boldsymbol{\theta})\right|$$

$$= \left|\frac{\partial}{\partial_j}\left(\sum_{\boldsymbol{x_T}\in\mathcal{X}_+^T}\sum_{t=1}^T \frac{\partial\xi_t(\boldsymbol{x_T})}{\partial\theta_i} \cdot \prod_{t'\neq t}\xi_{t'}(\boldsymbol{x_T})\right)\right|$$

$$= \left|\sum_{\boldsymbol{x_T}\in\mathcal{X}_+^T}\sum_{t=1}^T \left(\frac{\partial^2\xi_t}{\partial\theta_i\partial\theta_j}\prod_{t'\neq t}\xi_{t'}(\boldsymbol{x_T}) + \frac{\partial\xi_t(\boldsymbol{x_T})}{\partial\theta_i}\sum_{t'\neq t}\frac{\partial\xi_{t'}(\boldsymbol{x_T})}{\partial\theta_i}\prod_{t"\neq t,t'}\xi_{t"}(\boldsymbol{x_T})\right)\right|$$

$$\leq \sum_{\boldsymbol{x_T}\in\mathcal{X}_+^T}\sum_{t=1}^T \left(\left|\frac{\partial^2\xi_t}{\partial\theta_i\partial\theta_j}\right|\prod_{t'\neq t}\xi_{t'}(\boldsymbol{x_T}) + \left|\frac{\partial\xi_t(\boldsymbol{x_T})}{\partial\theta_i}\right|\left|\sum_{t'\neq t}\frac{\partial\xi_{t'}(\boldsymbol{x_T})}{\partial\theta_i}\prod_{t"\neq t,t'}\xi_{t"}(\boldsymbol{x_T})\right|\right)$$

$$\leq \sum_{\boldsymbol{x_T}\in\mathcal{X}_+^T}\prod_{t'=1}^T\xi_{t'}(\boldsymbol{x}_T)\sum_{t=1}^T\left(g_4(y_t) + \left(C_P + g_3(y_t)\right)^2 + \sum_{t'\neq t}\left(C_P + g_3(y_{t'})\right)\left(C_P + g_3(y_t)\right)\right)$$

$$\leq \sum_{t=1}^T\left(g_4(y_t) + \left(C_P + g_3(y_t)\right)^2 + \sum_{t'\neq t}\left(C_P + g_3(y_{t'})\right)\left(C_P + g_3(y_t)\right)\right)\prod_{t"=1}^T g_2(y_{t"})$$

where the right hand side of the inequality is integrable based on Fubini's theorem and assumptions A4-A6. Hence, (iii) holds and therefore, the sufficient condition (23) is satisfied, Q.E.D. □

Lemma 11 [51]**:** Assume A1-A10 hold, then for any sequence of positive numbers $\delta_n \to 0$, we have

$$\lim_{n\to\infty}\lim_{k\to\infty}\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\delta_n}\left|\frac{\nabla^2\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)}{k}+I(\boldsymbol{\theta}_0)\right|\xrightarrow{\mathbb{P}_0^\infty}0.$$

Proof: Following Lemma 2 in [11] and Lemma 5.2 in [133], we will prove that

$$\lim_{n\to\infty}\lim_{k\to\infty}\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\delta_n}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)+I_{ij}(\boldsymbol{\theta}_0)\right|\xrightarrow{\mathbb{P}_0^\infty}0,\forall i,j.$$

Due to Lemma 10, it is sufficient to prove

$$\lim_{n\to\infty}\lim_{k\to\infty}\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\delta_n}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_k(\boldsymbol{\theta}_0)\right|\xrightarrow{\mathbb{P}_0^\infty}0,\forall i,j.$$

For any $\epsilon > 0$,

$$\limsup_{n\to\infty}\limsup_{k\to\infty}\mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\delta_n}\frac{1}{k}\left|\frac{\partial^2\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)}{\partial\theta_i\partial\theta_j}-\frac{\partial^2\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)}{\partial\theta_i\partial\theta_j}\right|\ge\epsilon\right)$$

$$\le\limsup_{n\to\infty}\limsup_{k\to\infty}\mathbb{P}_0^\infty\left(\frac{1}{k}\sum_{i=1}^{k}\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\le\delta_n}\left|\frac{\partial^2\tilde{\ell}_i(\boldsymbol{\theta},\boldsymbol{y}_{i,T})}{\partial\theta_i\partial\theta_j}-\frac{\partial^2\tilde{\ell}_i(\boldsymbol{\theta}_0,\boldsymbol{y}_{i,T})}{\partial\theta_i\partial\theta_j}\right|\ge\epsilon\right)$$

[51] This lemma stipulates about a locally uniform convergence, which is strictly weaker than the uniform convergence in assumption (B2) for [5, Theorem 8.2], which requires a fixed δ for the supremum tending to 0 as k tends to infinity.

$$\leq \limsup_{n\to\infty} \limsup_{k\to\infty} \frac{\left(\sum_{i=1}^{k} \mathbb{E}_0\left[\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\delta_n}\left|\frac{\partial^2\tilde{\ell}_i(\boldsymbol{\theta},\boldsymbol{y}_{i,T})}{\partial\theta_i\partial\theta_j} - \frac{\partial^2\tilde{\ell}_i(\boldsymbol{\theta}_0,\boldsymbol{y}_{i,T})}{\partial\theta_i\partial\theta_j}\right|\right]\right)}{k\epsilon}$$

$$= \limsup_{n\to\infty} \frac{\mathbb{E}_0\left[\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\delta_n}\left|\frac{\partial^2\ell_1(\boldsymbol{\theta},\boldsymbol{y}_T)}{\partial\theta_i\partial\theta_j} - \frac{\partial^2\ell_1(\boldsymbol{\theta}_0,\boldsymbol{y}_T)}{\partial\theta_i\partial\theta_j}\right|\right]}{\epsilon}$$

$$\leq \frac{\mathbb{E}_0\left[\limsup_{n\to\infty} \sup_{|\theta-\theta_0|\leq\delta_n}\left|\frac{\partial^2\ell_1(\boldsymbol{\theta},\boldsymbol{y}_T)}{\partial\theta_i\partial\theta_j} - \frac{\partial^2\ell_1(\boldsymbol{\theta}_0,\boldsymbol{y}_T)}{\partial\theta_i\partial\theta_j}\right|\right]}{\epsilon}$$

$$= 0.$$

Please note that the third inequality is derived by Markov inequality, while the fourth inequality is obtained by Fatou's Lemma and using the integrable bound by which for all $n \geq 1$,

$$\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\delta_n}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_1(\boldsymbol{\theta},\boldsymbol{y}_T) - \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_1(\boldsymbol{\theta}_0,\boldsymbol{y}_T)\right|$$

$$\leq 2\cdot \sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\sup\{\delta_n\}}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_1(\boldsymbol{\theta},\boldsymbol{y}_T)\right|$$

$$\leq 2\cdot\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\sup\{\delta_n\}}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(\boldsymbol{y}_T|\boldsymbol{\theta})\right| + \sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\sup\{\delta_n\}}\left|\frac{\partial f(\boldsymbol{y}_T|\boldsymbol{\theta})}{\partial\theta_i}\right|\cdot\right.$$

$$\left.\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\sup\{\delta_n\}}\left|\frac{\partial f(\boldsymbol{y}_T|\boldsymbol{\theta})}{\partial\theta_j}\right| \sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\sup\{\delta_n\}}\left|\frac{1}{f(\boldsymbol{y}_T|\boldsymbol{\theta})}\right|\right)$$

and the last equation is given by the uniform continuity of $\nabla^2\ell_1(\boldsymbol{\theta})$ over the closed set $|\boldsymbol{\theta}-\boldsymbol{\theta}_0| \leq \sup\{\delta_n\}$, Q.E.D. □

Lemma 12: Assume A1 – A10 hold, then for any $\mathcal{M} < \infty$ and for any $\epsilon > 0$,

$$\lim_{k\to\infty} \mathbb{P}_0^\infty \left(\int_{|\boldsymbol{u}|\leq\mathcal{M}} \left| \pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})} - \pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\frac{\boldsymbol{u}^\mathrm{T} I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2}} \right| \mathrm{d}\boldsymbol{u} \leq \epsilon \right) = 1.$$

Proof: This proof follows Theorem 8.2 in [165]. Since by Lemma 9, $\boldsymbol{\tau}_k$ converges to $\boldsymbol{\theta}_0$ in probability and since $\sup_{|\boldsymbol{u}|\leq\mathcal{M}} k^{-\frac{1}{2}}\boldsymbol{u}$ converges to 0 as $k \to \infty$, we have

$$\sup_{|u|\leq\mathcal{M}} \left|\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k - \boldsymbol{\theta}_0\right| \xrightarrow{\mathbb{P}_0^\infty} 0. \tag{24}$$

Then by A9 and Continuous Mapping Theorem (Lemma 2),

$$\sup_{|\boldsymbol{u}|\leq\mathcal{M}} \left|\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) - \pi(\boldsymbol{\theta}_0)\right| \xrightarrow{\mathbb{P}_0^\infty} 0$$

By Lemma 9, $I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)/\sqrt{k}$ converges in distribution and is therefore bounded in probability. It follows that $\sup_{|u|\leq\mathcal{M}} \left(\boldsymbol{u} + I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)/\sqrt{k}\right)^2$ is also bounded in probability. Furthermore, Lemma 11 implies that

$$\sup_{|u|\leq\mathcal{M}} \left|\frac{1}{2k} R_k\left(\boldsymbol{\tau}_k + \frac{\boldsymbol{u}}{\sqrt{k}}\right)\right| \xrightarrow{\mathbb{P}_0^\infty} 0. \tag{25}$$

This is because for any $\epsilon > 0$ and for any sequence $\delta_n \to 0$,

$$\limsup_{k\to\infty} \mathbb{P}_0^\infty \left(\sup_{|\boldsymbol{u}|\leq\mathcal{M}} \left|\frac{1}{2k} R_k\left(\boldsymbol{\tau}_k + \frac{\boldsymbol{u}}{\sqrt{k}}\right)\right| \geq \epsilon \right)$$

$$\leq \limsup_{k\to\infty} [\mathbb{P}_0^\infty \left(\sup_{|\boldsymbol{u}|\leq\mathcal{M}} \left|\frac{1}{2k} R_k\left(\boldsymbol{\tau}_k + \frac{\boldsymbol{u}}{\sqrt{k}}\right)\right| \geq \epsilon,\ \sup_{|\boldsymbol{u}|\leq\mathcal{M}} \left|\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k - \boldsymbol{\theta}_0\right| \leq \right.$$

$$\left. \delta_n \right) + \mathbb{P}_0^\infty \left(\sup_{|\boldsymbol{u}|\leq\mathcal{M}} \left|\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k - \boldsymbol{\theta}_0\right| > \delta_n \right)]$$

$$\leq \limsup_{k\to\infty} \mathbb{P}_0^\infty \left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\leq\, \delta_n} \left| \frac{1}{2k} R_k(\boldsymbol{\theta}) \right| \geq \epsilon \right)$$

$$+ \limsup_{k\to\infty} \mathbb{P}_0^\infty \left(\sup_{|\boldsymbol{u}|\leq\mathcal{M}} \left| \frac{1}{\sqrt{k}} \boldsymbol{u} + \boldsymbol{\tau}_k - \boldsymbol{\theta}_0 \right| > \delta_n \right).$$

Notice that following (24), the second term is 0 for any $\delta_n > 0$, while the first term vanishes as $n \to \infty$ by the locally uniform convergence in Lemma 10. Thus, (24) is proven.

It follows by Lemma 3 that

$$\sup_{|\boldsymbol{u}|\leq\mathcal{M}} \left| \frac{1}{2k} \left(\boldsymbol{u} + \frac{I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)}{k} \right)^{-1} R_k\left(\boldsymbol{\tau}_k + \frac{\boldsymbol{u}}{\sqrt{k}}\right) \left(\boldsymbol{u} + \frac{I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)}{k} \right) \right| \xrightarrow{\mathbb{P}_0} 0,$$

and consequently, by Lemma 5 we have that

$$\sup_{|u|\leq M} \left| \omega_k(\boldsymbol{u}) + \frac{\boldsymbol{u}^{\mathrm{T}} I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2} \right| \xrightarrow{\mathbb{P}_0^\infty} 0.$$

Hence, per Lemma 1, we know that

$$\sup_{|\boldsymbol{u}|\leq M} \left| \pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u}) + \frac{\boldsymbol{u}^{\mathrm{T}} I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2}} - \pi(\boldsymbol{\theta}_0) \right| \xrightarrow{\mathbb{P}_0^\infty} 0,$$

which proves the lemma, Q.E.D. □

Lemma 13: Assume A1 – A10 hold, then for any positive sequence $\{a_k\}$ such that $a_k/\sqrt{k} \to 0$, there exists an integrable function $\mathcal{H}(u)$ such that

$$\lim_{k\to\infty} \mathbb{P}_0^\infty \left(\int_{|\boldsymbol{u}|\leq a_k} \left| \pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})} - \pi(\boldsymbol{\theta}_0) \mathrm{e}^{-\frac{\boldsymbol{u}^{\mathrm{T}} I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2}} \right| d\boldsymbol{u} \leq \int_{|\boldsymbol{u}|\leq a_k} H(\boldsymbol{u})\, d\boldsymbol{u} \right) = 1$$

Proof: The proof follows Theorem 8.2 in [165]. Since $\pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\boldsymbol{u}^{\mathrm{T}}I(\boldsymbol{\theta}_0)\boldsymbol{u}/2}$ is integrable by itself, it suffices to show that some integrable function $H(\boldsymbol{u})$ exists such that

$$\lim_{k\to\infty} \mathbb{P}_0^{\infty}\left(\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u}+\boldsymbol{\tau}_k\right)e^{\omega_k(\boldsymbol{u})} \leq H(\boldsymbol{u}), \forall|\boldsymbol{u}| \leq a_k\right) = 1, \tag{26}$$

Since τ_k converges to θ_0 in probability and by Lemma 9 $\sup_{|\boldsymbol{u}|\leq a_k}\left|k^{-\frac{1}{2}}\boldsymbol{u}\right| \xrightarrow{k\to\infty} 0$, we have

$$\sup_{|\boldsymbol{u}|\leq a_k}\left|\frac{1}{\sqrt{k}}\boldsymbol{u}+\boldsymbol{\tau}_k-\boldsymbol{\theta}_0\right| \xrightarrow{\mathbb{P}_0^{\infty}} 0,$$

and then per Lemma 2, $\sup_{|\boldsymbol{u}|\leq a_k}\left|\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u}+\boldsymbol{\tau}_k\right)-\pi(\boldsymbol{\theta}_0)\right| \xrightarrow{\mathbb{P}_0^{\infty}} 0$.

For $C_1 = \epsilon + \pi(\boldsymbol{\theta}_0)$ where ϵ is any postive number,

$$\lim_{k\to\infty} \mathbb{P}_0^{\infty}\left(\sup_{|\boldsymbol{u}|\leq a_k}\left|\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u}+\boldsymbol{\tau}_k\right)\right| \leq C_1\right) = 1. \tag{27}$$

Based on the fact that $I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\left\{\boldsymbol{y}_{i,T}\right\}_{i=1}^{k}\right)/\sqrt{k}$ is bounded in probability, as proven in Lemma 9, and the fact that

$$\sup_{|\boldsymbol{u}|\leq a_k}\left|\frac{1}{2k}R_k\left(\boldsymbol{\tau}_k+\frac{\boldsymbol{u}}{\sqrt{k}}\right)\right| \xrightarrow{\mathbb{P}_0^{\infty}} 0,$$

which follows from (25) and Lemma 11 [52]. Applying Lemma 3, we obtain

$$\sup_{|\boldsymbol{u}|\leq a_k}\left|\frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\left\{\boldsymbol{y}_{i,T}\right\}_{i=1}^{k}\right)\right|^2\left|\frac{1}{2k}R_k\left(\boldsymbol{\tau}_k+\frac{\boldsymbol{u}}{\sqrt{k}}\right)\right| \xrightarrow{\mathbb{P}_0^{\infty}} 0.$$

[52] Considering that for any $\delta > 0$, k and τ_k, $\left\{\boldsymbol{\theta}\colon \boldsymbol{\theta} = \left|\frac{1}{\sqrt{k}}\boldsymbol{u}+\boldsymbol{\tau}_k\right| \text{ where } \sup_{\mathcal{M}\leq|\boldsymbol{u}|\leq\mathcal{M}+a_k}\left|\frac{1}{\sqrt{k}}\boldsymbol{u}+\boldsymbol{\tau}_k-\boldsymbol{\theta}_0\right| \leq \delta\right\} \subseteq \{\boldsymbol{\theta}\colon \boldsymbol{\theta} = \ |\boldsymbol{\theta}-\boldsymbol{\theta}_0| \leq \delta\}$.

Let $\lambda_{min} > 0$ be the smallest eigenvalue of the nonsingular matrix $I(\boldsymbol{\theta}_0)$. Then, for any $\epsilon, \epsilon_1 > 0$, there exists K such that $\forall k \geq K$,

$$\mathbb{P}_0^\infty\left(\left|\frac{1}{2k}R_k\left(\boldsymbol{\tau}_k + \frac{\boldsymbol{u}}{\sqrt{k}}\right)\right| \leq \frac{1}{8}\lambda_{min}, \forall|\boldsymbol{u}| \leq a_k\right) > 1 - \epsilon, \tag{28}$$

and

$$\mathbb{P}_0^\infty\left(\left|\frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k(\boldsymbol{\theta}_0)\right|^2\left|\frac{1}{2k}R_k\left(\boldsymbol{\tau}_k + \frac{\boldsymbol{u}}{\sqrt{k}}\right)\right| \leq \epsilon_1, \forall|\boldsymbol{u}| \leq a_k\right) > 1 - \epsilon. \tag{29}$$

Since $\forall|\boldsymbol{u}| \leq a_k$, we have

$$\left|\frac{1}{2k}\left(\boldsymbol{u} + \frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\right)^{\mathrm{T}} R_k\left(\frac{\boldsymbol{u}}{\sqrt{k}} + \boldsymbol{\tau}_k\right)\left(\boldsymbol{u} + \frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\right)\right|$$

$$\leq 2|\boldsymbol{u}|^2\left|\frac{1}{2k}R_k\left(\frac{\boldsymbol{u}}{\sqrt{k}} + \boldsymbol{\tau}_k\right)\right| + 2\left|\frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\right|^2\left|\frac{1}{2k}R_k\left(\boldsymbol{\tau}_k + \frac{\boldsymbol{u}}{\sqrt{k}}\right)\right|. \tag{30}$$

In addition, we have

$$\omega_k(\boldsymbol{u}) = -\frac{\boldsymbol{u}^T I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2} - \frac{1}{2k}\left(\boldsymbol{u} + \frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\right)^{\mathrm{T}} R_k\left(\frac{\boldsymbol{u}}{\sqrt{k}} + \boldsymbol{\tau}_k\right)\Bigg(\boldsymbol{u} +$$

$$\frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\Bigg)$$

$$\leq -\frac{1}{2}|\boldsymbol{u}|^2\lambda_{min} + \Bigg|\frac{1}{2k}\left(\boldsymbol{u} + \frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\right)^{\mathrm{T}} R_k\left(\frac{\boldsymbol{u}}{\sqrt{k}} + \boldsymbol{\tau}_k\right)\Bigg(\boldsymbol{u} +$$

$$\frac{1}{\sqrt{k}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0, \{\boldsymbol{y}_{i,T}\}_{i=1}^k\right)\Bigg)\Bigg|. \tag{31}$$

It follows from eqs. (28-31) that

$$\mathbb{P}_0^\infty\left(\omega_k(\boldsymbol{u}) \leq -\frac{1}{4}|\boldsymbol{u}|^2\lambda_{min} + \epsilon_1, \forall|\boldsymbol{u}| \leq a_k\right) \geq 1 - 2\epsilon,$$

which leads to

$$\lim_{k\to\infty} \mathbb{P}_0^\infty \left(e^{\omega_k(\boldsymbol{u})} \leq e^{-\frac{1}{4}|\boldsymbol{u}|^2\lambda_{min}+\epsilon_1}, \forall|\boldsymbol{u}| \leq a_k\right) = 1,$$

Therefore, eq. (26) is satisfied by choosing $H(u) = e^{\log C_1-\frac{1}{4}|\boldsymbol{u}|^2\lambda_{min}+\epsilon_1}$, Q.E.D. □

Lemma 14. Assume A1-A10 hold**,** and denote

$$J_k(\delta) = \int_{\sqrt{k}\delta\leq|\boldsymbol{u}|} \left|\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})} - \pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\frac{\boldsymbol{u}^{\mathrm{T}}I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2}}\right| d\boldsymbol{u}.$$

Then, the following is true:
(i) For any fixed $\delta > 0, J_k(\delta) \xrightarrow{\mathbb{P}_0^\infty} 0$.
(ii) There exists a sequence of positive numbers δ_k such that $\delta_k \to 0, \delta_k\sqrt{k} \to \infty$, and $J_k(\delta_k) \xrightarrow{\mathbb{P}_0^\infty} 0$.
Proof: The first part of the proof follows Theorem 8.2 in [165], and the second part of the proof follows Theorem 3.2 in [18].
(i) The expression $\pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\frac{\boldsymbol{u}^{\mathrm{T}}I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2}}$ is proportional to some Guassian density and therefore is negligible in the integration for sufficiently large k. It suffices to prove that

$$\int_{\delta\sqrt{k}\leq|\boldsymbol{u}|} \pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u}$$

converges to zero in $\mathbb{P}_0^\infty$-probability. One can observe that

$$\int_{\delta\sqrt{k}\leq|\boldsymbol{u}|} \pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u}$$

$$= \sqrt{k} \int_{|\boldsymbol{\theta}-\boldsymbol{\tau}_k|\geq\delta} \pi(\boldsymbol{\theta}) e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\frac{\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)^{\mathrm{T}} I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)}{2k}} d\boldsymbol{\theta}.$$

According to Lemma 8, for any $\delta > 0$, there exists ϵ_δ such that

$$\lim_{k\to\infty} \mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\frac{\delta}{2}} e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)} \leq e^{-k\epsilon_\delta}\right) = 1.$$

Since for any k,

$$\mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\frac{\delta}{2}} e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)} \leq e^{-k\epsilon_\delta}\right)$$

$$\leq \mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta}_0|\geq\frac{\delta}{2}} e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)} \leq e^{-k\epsilon_\delta} \text{ and } |\boldsymbol{\tau}_k-\boldsymbol{\theta}_0| < \frac{\delta}{2}\right) +$$

$$\mathbb{P}_0^\infty\left(|\boldsymbol{\tau}_k-\boldsymbol{\theta}_0| < \frac{\delta}{2}\right)$$

$$\leq \mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\tau}_k|\geq\delta} e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)} \leq e^{-k\epsilon_\delta}\right) + \mathbb{P}_0^\infty\left(|\boldsymbol{\tau}_k-\boldsymbol{\theta}_0| < \frac{\delta}{2}\right), \tag{32}$$

and since $\boldsymbol{\tau}_k \xrightarrow{\mathbb{P}_0^\infty} \boldsymbol{\theta}_0$, we obtain that

$$\lim_{k\to\infty} \mathbb{P}_0^\infty\left(\sup_{|\boldsymbol{\theta}-\boldsymbol{\tau}_k|\geq\delta} e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)} \leq e^{-k\epsilon_\delta}\right) = 1.$$

Since $\sup_{|\boldsymbol{\theta}-\boldsymbol{\tau}_k|\geq\delta} e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)} \leq e^{-k\epsilon_\delta}$ implies that

$$\sqrt{k}\int_{|\boldsymbol{\theta}-\tau_k|\geq\delta}\pi(\boldsymbol{\theta})e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\frac{\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)^{\mathrm{T}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)}{2k}}\,d\boldsymbol{\theta}$$

$$\leq\sqrt{k}\sup_{|\boldsymbol{\theta}-\tau_k|\geq\delta}e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)}\int_{|\boldsymbol{\theta}-\tau_k|\geq\delta}\pi(\boldsymbol{\theta})e^{-\frac{\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)^{\mathrm{T}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)}{2k}}\,d\boldsymbol{\theta}$$

$$\leq C_3\sqrt{k}e^{-k\epsilon}\int_{|\boldsymbol{\theta}-\boldsymbol{\tau}_k|\geq\delta}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

$$\leq C_3\sqrt{k}e^{-k\epsilon}.$$

Hence, for a given δ, the $\mathbb{P}_0^\infty$-probability that

$$\sqrt{k}\int_{|\boldsymbol{\theta}-\tau_k|\geq\delta}\pi(\boldsymbol{\theta})e^{\ell_k\left(\boldsymbol{\theta},\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)-\frac{\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)^{\mathrm{T}}I^{-1}(\boldsymbol{\theta}_0)\nabla\ell_k\left(\boldsymbol{\theta}_0,\{\boldsymbol{y}_{i,T}\}_{i=1}^{k}\right)}{2k}}\,d\boldsymbol{\theta}$$

$$\leq C_3\sqrt{k}e^{-k\epsilon}$$

converges to 1. The proof is complete per Lemma 4. Q.E.D.

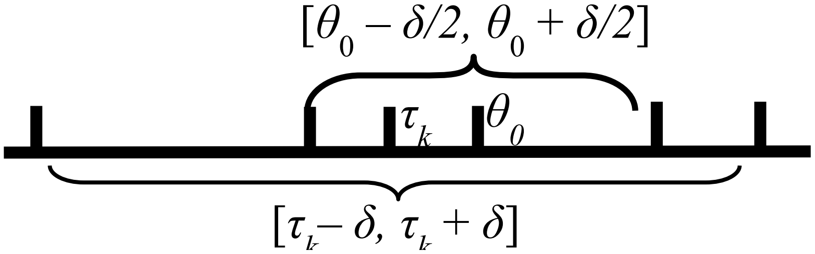

$[\theta_0 - \delta/2,\ \theta_0 + \delta/2]$
τ_k
θ_0
$[\tau_k - \delta,\ \tau_k + \delta]$

Figure 13. A univariate illustration of the probability events considered in the chain of inequalities (32) used in the proof of Lemma 14, part (i).

(ii) According to (i), for any strictly decreasing sequence of positive constants $\{\delta_i\}_{i=1}^{\infty}$ that converges to 0, there exists an integer sequence $\{N(\delta_i)\}_{i=1}^{\infty}$ such that for each $i \geq 1$ and for all $n \geq N(\delta_i)$,

$$\mathbb{P}_0^{\infty}(J_n(\delta_i) \geq \delta_i) \leq \delta_i. \tag{33}$$

Then, let us construct a monotonically increasing sequence $\{N'(\delta_i)\}_{i=1}^{\infty}$ such that $N'(\delta_1) = N(\delta_1)$ and

$$N'(\delta_i) = \max(N(\delta_i), N'(\delta_{i-1}) + 1, [\delta_i^{-2-\alpha}] + 1)\,, \forall i \geq 2,$$

for some $0 < \alpha < 1$. Since $\delta_i\sqrt{N'(\delta_i)} > \delta_i^{-\alpha/2}$. we have

$$\delta_i\sqrt{N'(\delta_i)} \to \infty, \text{as i} \to \infty. \tag{34}$$

Considering eq. (33), for any $\epsilon > 0$ there exists I such that for all $i \geq I$, we have $\delta_i < \epsilon$ and

$$\mathbb{P}_0^{\infty}\left(J_{N'(\delta_i)}(\delta_i) \geq \epsilon\right) \leq \mathbb{P}_0^{\infty}\left(J_{N'(\delta_i)}(\delta_i) \geq \delta_i\right) \leq \delta_i.$$

Therefore,

$$\lim_{i\to\infty} \mathbb{P}_0^{\infty}\left(J_{N'(\delta_i)}(\delta_i) \geq \epsilon\right) = 0, \tag{35}$$

Let us now construct a sequence $\{\delta_k'\}_{k=1}^{\infty}$ by choosing $\delta_k' = \delta_i$ for $N'(\delta_i) \leq k < N'(\delta_{i+1}), \forall k$, Then we have:

1) $\delta_k' \to 0$, since $\delta_i \to 0$;

2) $\delta_k'\sqrt{k} \to \infty$, since $\delta_k'\sqrt{k} > \delta_i\sqrt{N'(\delta_i)}$ (where $\delta_k' = \delta_i$) and (34);

3) $J_k(\delta_k) \overset{\mathbb{P}_0}{\to} 0$, since $\mathbb{P}_0(J_k(\delta_k') \geq \epsilon) > \mathbb{P}_0\left(J_{N'(\delta_i)}(\delta_i) \geq \epsilon\right)$ and (35).

Hence, $\{\delta_k'\}_{k=1}^{\infty}$ is the desired sequence, for Lemma 14 (ii), Q.E.D. □

Proof of Theorem 1: The proof follows the line of the proofs of Theorem 8.2 in [165] and Theorem 2.1 in [18]. Let

$$\mathcal{C}_k = \int_{\mathbb{R}^d} e^{\omega_k(\boldsymbol{u})}\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) d\boldsymbol{u}.$$

One readily shows that

$$\pi_k^*(\boldsymbol{u}|\boldsymbol{w}_k) = \frac{1}{\mathcal{C}_k}\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})}.$$

It follows that

$$\mathcal{C}_k \cdot \int_{\mathbb{R}^d}\left|\pi_{\mathrm{k}}^*(\boldsymbol{u}|\boldsymbol{w}_k) - \frac{1}{\sqrt{2\pi I^{-1}(\boldsymbol{\theta}_0)}} e^{-\frac{I(\boldsymbol{\theta}_0)\boldsymbol{u}^2}{2}}\right| d\boldsymbol{u}$$

$$\leq \int_{\mathbb{R}^d}\left|\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})} - \pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\frac{I(\boldsymbol{\theta}_0)\boldsymbol{u}^2}{2}}\right| d\boldsymbol{u}$$

$$+\int_{\mathbb{R}^d}\left|\frac{\mathcal{C}_k}{\sqrt{2\pi I^{-1}(\boldsymbol{\theta}_0)}}\mathrm{e}^{-I(\boldsymbol{\theta}_0)\frac{\boldsymbol{u}^2}{2}} - \pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\frac{I(\boldsymbol{\theta}_0)\boldsymbol{u}^2}{2}}\right| d\boldsymbol{u}. \qquad (36)$$

It now suffices to show that the first expression on the right hand side of (36) converges, i.e.

$$\int_{\mathbb{R}^d}\left|\pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})} - \pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\frac{I(\boldsymbol{\theta}_0)\boldsymbol{u}^2}{2}}\right| \mathrm{d}u \xrightarrow{\mathbb{P}_0^\infty} 0, \qquad (37)$$

because (37) immediately implies the convergence of $\mathcal{C}_k$ in probability $\mathbb{P}_0^\infty$ to a finite value, as well as the convergence in probability $\mathbb{P}_0^\infty$ of the second expression on the right hand side of inequality (36) to 0.

To show that (37) holds, let $J_k(\boldsymbol{u})$ be the integrand in the integral (36), i.e.,

$$J_k(\boldsymbol{u}) = \int_{\mathbb{R}^d} \left| \frac{\mathcal{C}_k}{\sqrt{2\pi I^{-1}(\boldsymbol{\theta}_0)}} \mathrm{e}^{-I(\boldsymbol{\theta}_0)\frac{\boldsymbol{u}^2}{2}} - \pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\frac{I(\boldsymbol{\theta}_0)\boldsymbol{u}^2}{2}} \right| d\boldsymbol{u}.$$

Then, for any positive $\epsilon \geq 0$, choose $\mathcal{M}$ according to the integrable function $H(\boldsymbol{u})$ from Lemma 13 such that $\int_{\mathcal{M}}^{\infty} H(\boldsymbol{u})\, d\boldsymbol{u} \leq \epsilon/3$. In addition, following Lemma 14 (ii), we can choose a sequence $\{\delta_k\}$ such that

$$\int_{\sqrt{k}\delta_k \leq |\boldsymbol{u}|} \left| \pi\left(\frac{1}{\sqrt{k}}\boldsymbol{u} + \boldsymbol{\tau}_k\right) e^{\omega_k(\boldsymbol{u})} - \pi(\boldsymbol{\theta}_0)\mathrm{e}^{-\frac{\boldsymbol{u}^{\mathrm{T}} I(\boldsymbol{\theta}_0)\boldsymbol{u}}{2}} \right| d\boldsymbol{u} \xrightarrow{\mathbb{P}_0^\infty} 0.$$

Then let $a_k = \max\{0, \sqrt{k}\delta_k - \mathcal{M}\}, \forall k$. Finally, let us choose a common K from Lemma 12, 13, and 14 such that for all $k \geq K$,

$$\mathbb{P}_0^\infty \left(\int_{|\boldsymbol{u}| \leq \mathcal{M}} J_k(\boldsymbol{u})\, d\boldsymbol{u} \leq \frac{\epsilon}{3} \right) \geq 1 - \frac{\epsilon}{3},$$

$$\mathbb{P}_0^\infty \left(\int_{\mathcal{M} \leq |\boldsymbol{u}| \leq \mathcal{M} + a_k} J_k(\boldsymbol{u})\, d\boldsymbol{u} \leq \frac{\epsilon}{3} \right) \geq 1 - \frac{\epsilon}{3},$$

$$\mathbb{P}_0^\infty \left(\int_{\mathcal{M} + a_k \leq |\boldsymbol{u}|} J_k(\boldsymbol{u})\, d\boldsymbol{u} \leq \frac{\epsilon}{3} \right) \geq 1 - \frac{\epsilon}{3}.$$

Such choices lead to (22) and the proof is now completed. □

5.3 ILLUSTRATIVE EXAMPLE

We provide here an example to demonstrate the applicability of Theorem 1.[53] Consider using the data $\{\boldsymbol{y}_{k,T}\}_{k=1}^{\infty}$ with $T = 2$ to estimate a 2-state Gaussian HMM with transition probability matrix

[53] This example could be extended for HMM with 3, 4, or even larger number of states. However, the mathematical expressions in the fulfilment of the assumptions would become intractable and therefore avoided in this chapter.

$$\boldsymbol{P}_0 = \begin{bmatrix} \alpha_0 & 1-\alpha_0 \\ 0 & 1 \end{bmatrix},$$

conditional distributions $Y_t|X_t = i \sim \mathcal{N}(\mu_{0i}, 1), \forall t \geq 0$ [54], with a known initial distribution $\nu_0 = [1,0]$, The HMM parameter vector $\boldsymbol{\theta_0} = (\alpha_0, \mu_{01}, \mu_{02})$ is an interior point of the parameter space defined by

$$\Omega = \{\boldsymbol{\theta} \in \mathbb{R}^3 : \alpha_{min} \leq \alpha \leq \alpha_{max} \, ; \mu_{min} + \Delta_\mu \leq \mu_1 + \Delta_\mu \leq \mu_2 \leq \mu_{max}\}, \qquad (38)$$

where $0 < \alpha_{min} < \alpha_{max} < 1, -\infty < \mu_{min} < \mu_{max} < \infty, \Delta_\mu > 0$. Assume the prior for α is scaled Beta(a, b) distribution with support on $[\alpha_{min}, \alpha_{max}]$, and each prior for the μ_i's is independently from normal distribution $N(\mu_0', {\sigma_0'}^2)$. It follows that the joint prior distribution on Ω satisfies

$$\pi(\alpha, \mu_1, \mu_2) = \frac{(\alpha - \alpha_{min})^{a-1}(\alpha_{max} - \alpha)^{b-1}}{(\alpha_{max} - \alpha_{min})B(a,b)} \times \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_0' C_\pi} e^{\frac{(\mu_i - \mu_0')^2}{2{\sigma_0'}^2}},$$

where C_π is a normalizing constant so that $\prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_0' C_\pi} e^{\frac{(\mu_i - \mu_0')^2}{2{\sigma_0'}^2}}$ integrates to 1 on $\{(\mu_1, \mu_2) : \mu_{min} + \Delta_\mu \leq \mu_1 + \Delta_\mu \leq \mu_2 \leq \mu_{max}\}$.

Now we check the assumptions needed for Theorem 1. A1-A3 hold obviously. For A4-A6, the bounding functions g_i's for $i = 1,2,3,4$ can be constructed in the following manner.

[54] Please note that in this example we assume known conditional variances for observation distributions.

$$g_1(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_{max})^2}{2}} & y < \mu_{min}, \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu_{max}-\mu_{min})^2}{2}} & \mu_{min} \leq y \leq \mu_{max}, \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_{min})^2}{2}} & \mu_{max} < y, \end{cases}$$

$$g_2(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_{min})^2}{2}} & y < \mu_{min}, \\ \frac{1}{\sqrt{2\pi}} & \mu_{min} \leq y \leq \mu_{max}, \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu_{max})^2}{2}} & \mu_{max} < y, \end{cases}$$

$$g_3(y) = \max\{|y-\mu_{01}| + \delta, |y-\mu_{02}| + \delta\},$$

and $g_4(y) = 1$. It is clear that $\int g_2(y)\, dy = 1 + \frac{(\mu_{max}-\mu_{min})}{\sqrt{2\pi}} < \infty$ and therefore

$$\int_{\mathbb{R}} |\log g_1| g_2 \, dy$$

$$\leq \int_{y<\mu_{min}} |\log g_1| g_2 \, dy + \int_{[\mu_{min},\mu_{max}]} |\log g_1| g_2 \, dy + \int_{\mu_{max}<y} |\log g_1| g_2 \, dy$$

$$\leq \int_{y<\mu_{min}} \left(\frac{(y-\mu_{max})^2}{2} + C_1\right) g_2 \, dy + \int_{[\mu_{min},\mu_{max}]} \left(\frac{(\mu_{max}-\mu_{min})^2}{2} + C_1\right) g_2 \, dy$$

$$+ \int_{\mu_{max}<y} \left(\frac{(y-\mu_{min})^2}{2} + C_1\right) g_2 \, dy$$

$$\leq \int_{y<\mu_{min}} \left(\frac{(y-\mu_{min})^2 - 2(\mu_{max}-\mu_{min})(y-\mu_{min}) + (\mu_{max}-\mu_{min})^2}{2} + C_1\right) g_2 \, dy$$

$$+ \int_{[\mu_{min},\mu_{max}]} \left(\frac{(\mu_{max}-\mu_{min})^2}{2} + C_1\right) g_2 \, dy$$

$$+ \int_{\mu_{max}<y} \left(\frac{(y-\mu_{max})^2 + 2(\mu_{max}-\mu_{min})(y-\mu_{max}) + (\mu_{max}-\mu_{min})^2}{2} + C_1\right) g_2 \, dy$$

$$\leq \int_{\mathbb{R}} \frac{\left(y^2+2(\mu_{max}-\mu_{min})|y|\right)}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy + \left(\frac{(\mu_{max}-\mu_{min})^2}{2} + C_1\right) \int_{\mathbb{R}} g_2\, dy$$

$$\leq \frac{1}{2} + \frac{2(\mu_{max}-\mu_{min})}{\sqrt{2\pi}} + \left(\frac{(\mu_{max}-\mu_{min})^2}{2} + C_1\right)\left(1 + \frac{(\mu_{max}-\mu_{min})}{\sqrt{2\pi}}\right)$$

$$< \infty,$$

where $C_1 = \log\sqrt{2\pi}$. Based on [167], the identifiability in assumption A7 can be established using the same argument as in Section 3 in [128]. Assumptions A8 and A9 are obviously satisfied. For A10, it is sufficient to show that $I(\boldsymbol{\theta}_0)$ is positive definite, or equivalently to show that for any a_1, a_2, a_3,

$$Var_{\theta_0}\left[a_1 \frac{\partial \ell_1}{\partial \alpha} + a_2 \frac{\partial \ell_1}{\partial \mu_1} + a_3 \frac{\partial \ell_1}{\partial \mu_2}\right] = 0, \text{if and only if } a_1 = a_2 = a_3 = 0, \qquad (39)$$

since $\mathbb{E}_0[\nabla \ell_1(\boldsymbol{\theta}_0)] = 0$ and $I(\boldsymbol{\theta}_0) = Var_{\boldsymbol{\theta}_0}[\nabla \ell_1(\boldsymbol{\theta}_0)]$. Note that

$$Var_{\boldsymbol{\theta}_0}\left[a_1 \frac{\partial \ell_1}{\partial \alpha} + a_2 \frac{\partial \ell_1}{\partial \mu_1} + a_3 \frac{\partial \ell_1}{\partial \mu_2}\right]$$

$$= \int \left(a_1 \left(\frac{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)}{\partial \alpha}\right)^2 + a_2 \left(\frac{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)}{\partial \mu_1}\right)^2 + a_3 \left(\frac{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)}{\partial \mu_2}\right)^2\right) \frac{1}{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)}\, dy_1 y_2$$

where is based on

$$f_{\boldsymbol{\theta}_0}(y_1, y_2) = \alpha_0 f(y_1|\mu_{01}) f(y_2|\mu_{01}) + (1-\alpha_0) f(y_1|\mu_{01}) f(y_2|\mu_{02}),$$

it can be shown that $\int a_1 \left(\frac{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)}{\partial \alpha}\right)^2 \frac{1}{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)} dy_1 y_2 > 0$, $\int a_2 \left(\frac{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)}{\partial \mu_1}\right)^2 \frac{1}{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)} dy_1 y_2 > 0$, and $\int a_3 \left(\frac{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)}{\partial \mu_2}\right)^2 \frac{1}{\partial f_{\boldsymbol{\theta}_0}(y_1,y_2)} dy_1 y_2 > 0$. Therefore (24) is proven, and hence the nonsingularity of $I(\boldsymbol{\theta}_0)$ is established. Consequently, Theorem 1 applies to this HMM.

5.4 ILLUSTRATION OF THEOREM 1

Let us now consider a 2-state degradation (left-to-right) HMM of the form discussed in Sec. 5.3, with parameters $\boldsymbol{\theta}_0 = \left(P_{0,11}, \mu_{0,1}, \mu_{0,2}\right) = (0.8, 5, 10)$ and known variances $\sigma_{0,1} = \sigma_{0,2} = 1$ per discussion. In Section 5.3, this HMM satisfies assumption A1-A10. Conjugate priors were set for each of the three parameters, including Beta(1, 1) as the flat prior for α, Gaussian $N(0, 10)$ as the prior for μ_1 and μ_2. The model parameters were then estimated using the Gibbs sampling algorithm presented in Chapter 3. This estimation procedure was repeated 30 times, with the number of observation sequences increasing from 100 to 3100 sequences, each of which contains 5 observations.

Figure 14 shows the sequence of posterior means with two posterior standard deviations above and below the means, for each parameter. It is clear that as the number of sequences grows, each sequence of posterior means becomes closer to the corresponding actual parameter value, while the posterior standard deviation decreases towards 0. Figure 15 shows the posterior standard deviations for each parameters, as well as the least square approximation to the progression of those standard deviations by a curve decreasing at the rate of $1/\sqrt{k}$.

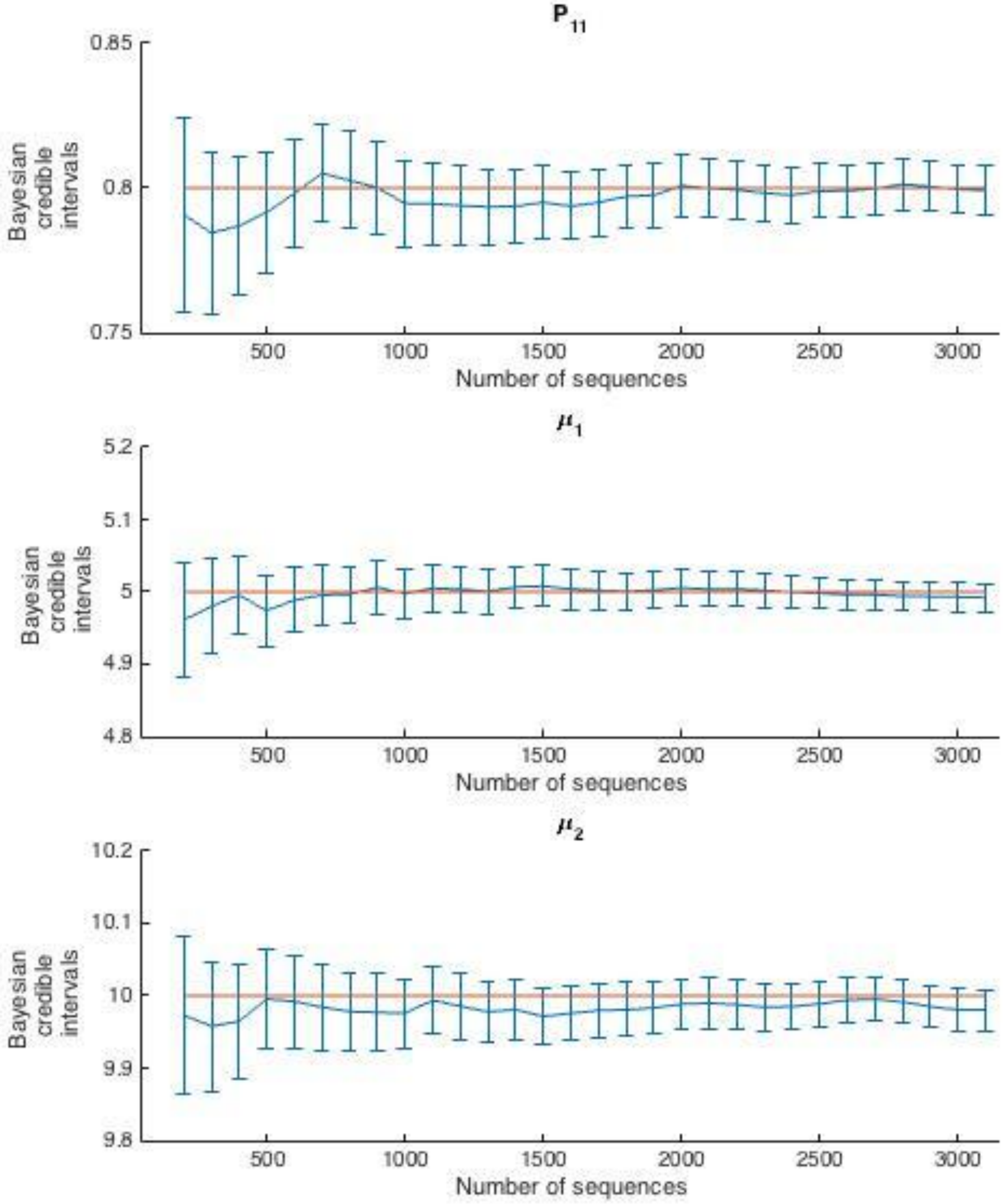


Figure 14: Posterior mean with 2-sigma error limits based on approximated posterior distribution as the number of observation sequences increases.

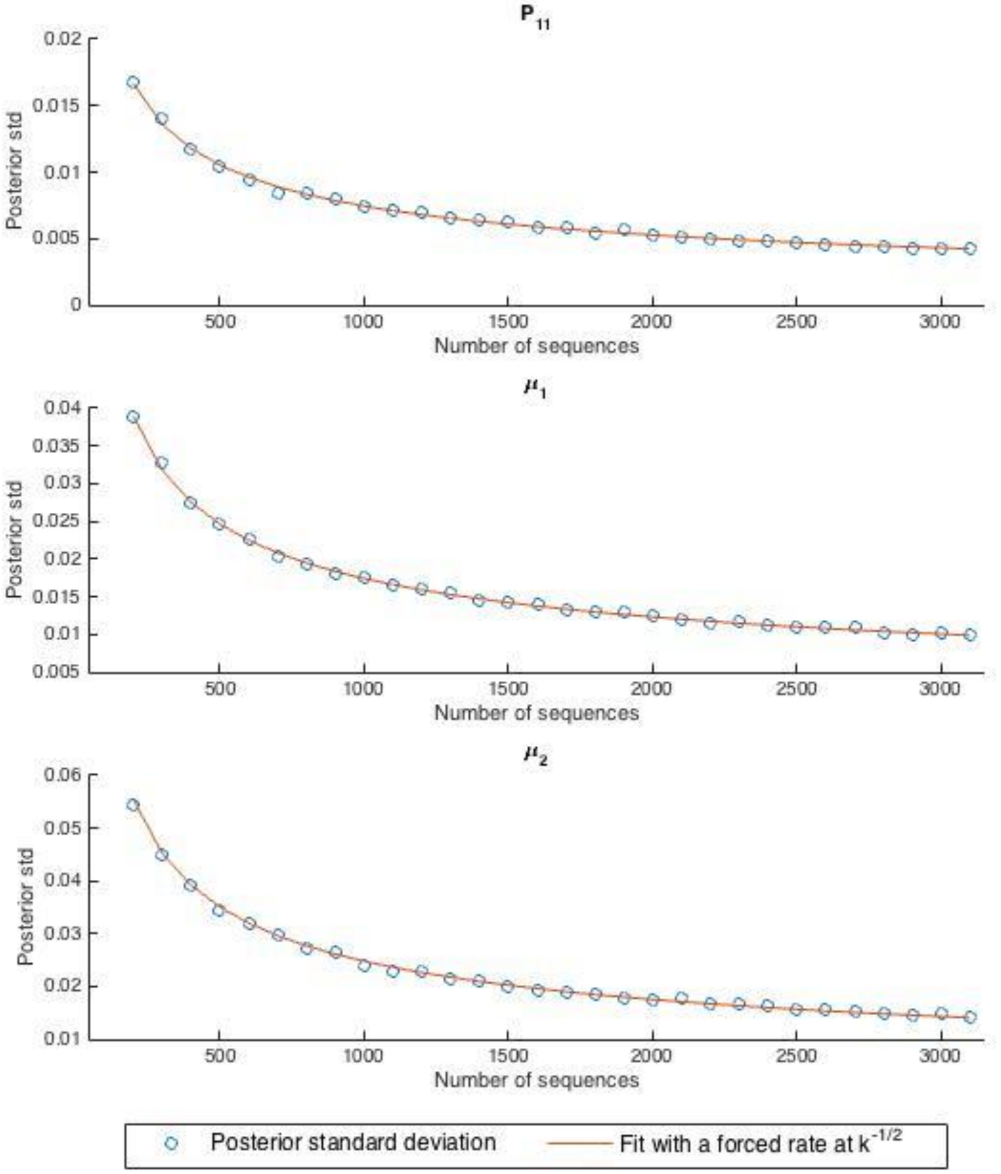


Figure 15: Posterior standard deviations based on the approximated posterior distribution as the number of observations varies. For each parameter, the sequence of standard deviations is fitted by a curve proportional to $1/\sqrt{k}$ in a least-square sense.

5.5 DISCUSSION

The BvMT formulated as Theorem 2.1 in [18] is similar to Theorem 1 proven in this chapter. However, there are significant distinctions between these two theorems. While the BvMT in [18] depicts asymptotic behavior of posterior distribution as the number of observations tends to infinity which limits it to applicability only to ergodic HMMs. Theorem 1 asserts the asymptotic posterior normality given an infinite amount of observation sequences of finite length, which makes it applicable even to non-ergodic HMMs. There is also a clear distinction between our formulation and the one in [18] in the way the limiting Fisher information matrix is considered. The information matrix $I(\boldsymbol{\theta}_0)$ defined in this chapter represents the information gain about HMM parameters $\boldsymbol{\theta}$ per sequence, whereas the information matrix from [18] expresses that gain with respect to each observation symbol.

Finally, it should be noted that the character and rate of decline of the posterior standard deviations derived by Theorem 1 enables prediction of how many observation sequences are needed until uncertainties in the posterior distribution decline under a certain tolerance. This is of utmost importance if one wants to build and use HMMs for modeling of degradation of machine conditions in condition based-maintenance.

Chapter 6: Conclusions and Future Work

6.1 Summary of Accomplishments

This dissertation focuses on a scientific study and engineering application of potentially non-stationary HMMs, including uncertainty quantification of Bayesian estimation of HMM parameters, condition-based monitoring of complex machines using HMM models with uncertain parameters as well as derivation of an asymptotic theory for Bayesian HMM parameter estimation and the associated uncertainty in the case of unidirectional time-homogeneous HMMs.

In Chapter 3, we proposed a novel Markov Chain Monte Carlo method that produces a probability distribution of model parameters for a non-homogeneous and non-ergodic HMM, and such distribution was utilized further in a novel degradation condition monitoring method that tracks discrepancy between the new data and the nominal HMM. Monitoring capabilities of the newly proposed degradation modeling and monitoring methods based on HMMs were then demonstrated on a massive dataset collected over several months from a PECVD tool operating in a major semiconductor fab. Results of this work were published in a recent journal paper [152].

In the degradation modeling approach in Chapter 3, each maintenance action was assumed to recover the degradation level[55] to the "as-good-as-new" level. This assumption was relaxed in Chapter 4, where imperfect maintenance operations were considered. A novel non-homogeneous and non-ergodic HMM was proposed to probabilistically model the recovery of the degradation level due to each maintenance

[55] I.e. perfecst maintenance operations were assumed.

event, which was, unlike what was considered in Chapter 3, potentially imperfect. Experiments with the large PECVD data set showed using that the newly-proposed HMM of system degradation that acknowledged maintenance imperfections yields significantly higher likelihood, as well as fault detection capabilities than the HMM assuming perfect maintenance operations. A novel filtering method was also proposed to provide degradation condition for each observation, rather than only for each sequence of observations, as suggested in [152]. Once again, dramatic improvements in terms of fault detection performance were observed when compared to fault detection based on the traditional PCA/T^2 based multivariate SPC method.

Motivated by the empirical convergences of HMM parameter distribution demonstrated in Chap. 3, a rigorous theoretical framework was proposed in Chap. 5 for studying the asymptotic behavior of Bayesian posterior distributions of parameters of HMMs, including left-to-right HMMs, obtained as more and more observation sequences became available during a Bayesian estimation process. Under a set of regularity conditions, the expected value of the posterior distribution was proven to convergence toward actual parameters, and the posterior standard deviation was proven to converge to 0 at approximately the rate of $1/\sqrt{n}$, where n is the number of sequences of observations used for modeling. This theoretical rate was shown to be consistent with the empirical convergence rate via a simulated example that was shown to satisfy all the assumptions for such convergence. Although convergence studies with infinite sequence for ergodic HMMs [18], we are not aware of any studies regarding the convergence of Bayesian estimation using infinitely many finite sequences which addresses unidirectional and hence non-ergodic HMMs.

6.2 Scientific Contributions

Quantification of model uncertainty is paramount to any application of data-driven models in CBM, and the newly proposed Bayesian estimation method delivers such confidence information about model parameters for several types of non-stationary HMMs. To the best of author's knowledge, this work was the first to obtain confidence evaluation in estimation of non-ergodic and non-homogeneous HMMs.

Furthermore, the fault detection method proposed in this thesis can robustly detect behavior changes in complex systems whose degradation dynamics are not perfectly (deterministically) observable. The newly introduced monitoring methods are the first HMM-based fault detection methods that incorporate the degradation model uncertainties into the decision-making process as to whether a fault occurred or not.

Finally, the theoretical analysis of the Bayesian estimation procedure provides further understanding of Bayesian estimation of parameters without assuming “usual” HMMs (ergodic) and can provide performance guarantees of the identified model in terms of its uncertainty levels. Such understanding can enable formal determination of the number of observation samples needed to achieve a desired level of model uncertainty, which in turn would enable economical considerations for the model identification that would balance the model precision and modeling cost.

6.3 Publications

The publications already produced or anticipated based on this doctoral research are as follows:

- Deyi Zhang, Andrew D. Bailey III, and Dragan Djurdjanovic, "Bayesian Identification of Hidden Markov Models and Their Use for Condition-Based

Monitoring," *IEEE Transactions on Reliability*, vol. 65, no. 2, pp. 1471-1482, 2016.

- Deyi Zhang and Dragan Djurdjanovic, “A hidden Markov model based approach to modeling and monitoring of processes with imperfect maintenance,” *the 16th IFAC Symposium on Information Control Problems in Manufacturing,* INCOM 2018*,* submitted.
- Deyi Zhang and Dragan Djurdjanovic, “A Bernstein-von Mises theorem for hidden Markov models,” anticipated journal paper based on Chapter 5.

6.4 POSSIBLE FUTURE WORK

Firstly, maintenance events in advanced manufacturing, as well as many other areas, entail collation of signals that reflect the character and quality of those interventions. Utilization of those signals for evaluating the quality of maintenance operations (or lack thereof) is a tremendous opportunity for future research that would improve HMM based CBM. Unfortunately, the datasets considered in this thesis, though they indeed reflected real-life manufacturing processes, did not contain any signals collected during maintenance operations and hence this work remained outside the scope of this research and should be considered in the future. In addition, wider scale implementation of the HMM based monitoring methods to CBM of other complex processes, such as semiconductor etching or downhole condition monitoring in oil/gas extraction, as well as utilization of such models for optimal scheduling of production, logistic and maintenance operations, as considered in [164], remains a promising direction for future research.

When it comes to theoretical considerations of Bayesian estimation of HMM parameters, various extensions to Theorem 1 from Chapter 5 are foreseeable. The single-regime left-to-right HMM was addressed by Theorem 1, and one future direction would be to show a BvMT for a more general class of nonhomogeneous (multiple-regime), non-ergodic HMMs. Another potentially fruitful avenue is to consider the convergence of posterior distributions of HMM parameters based on observation sequences of variable-lengths. Many practical maintenance schemes would necessitate the length of each observation sequence to be non-constant and related to some properties of the sample path for the hidden states. Finally, further refinement of posterior normality can be pursued via quantification of the convergence rate in terms of the L1 distance between the posterior distributions and their Gaussian appoximants, if higher accuracy of the approximation is of interest.

Acronyms

AUC	Area Under the Curve
BvMT	Bernstein-von Mises Theorem
CBM	Condition Based Maintenance
CI	Confidence Interval
CI	Confidence Index
EM	Expectation Maximization
FD	Fault Diagnosis
HMM	Hidden Markov Model
i.i.d	independent and identically distributed
KS	Kolmogorov-Smirnov
LDA	Linear Discriminate Analysis
MAP	Maximum A Posteriori
MCMC	Monte Carlo Markov Chain
MLE	Maximum Likelihood Estimator
PECVD	Plasma-Enhanced Chemical Vapor Deposition
SMC	Sequential Monte Carlo
SOM	Self Organizing Map
SPC	Statistical Process Control
TPM	Transition Probability Matrix
ROC	Receiver’s Operating Characteristics

Summary of Notations

N	Number of hidden states
M	Number of observations
L	Number of regimes
$O = \{o_1, o_2, \dots, o_M\}$	Set of observation symbols
$S = \{s_1, s_2, \dots, s_N\}$	Set of hidden states
$R = \{r_1, r_2, \dots, r_L\}$	Set of operating regimes
$\boldsymbol{\nu} = [\nu_1 \quad \nu_2 \quad \cdots \quad \nu_N]^T$	Probability vector as initial distribution
$\mathbf{P}^{(r)} = \left[p_{i,j}^{(r)}\right]_{i,j=1,2,\dots,N}$	Transition probability matrix for regime $r \in R$
$\mathbf{Q}^{(\mathrm{r})} = \left[q_{i,j}^{(r)}\right]_{\substack{i=1,2,\dots,N\\ j=1,2,\dots,M}}$	Emission probability matrix for regime $r \in R$
$\boldsymbol{\theta}^{(R)} = \left(\boldsymbol{\nu}, \mathbf{P}^{(r_1)}, \mathbf{Q}^{(r_1)}, \mathbf{P}^{(r_2)}, \mathbf{Q}^{(r_2)}, \dots, \mathbf{P}^{(r_L)}, \mathbf{Q}^{(r_L)}\right)$	Parameter of a regime-specific hidden Markov model
$\Omega^{(R)}$	Parameter space for a regime-specific hidden Markov model
$\boldsymbol{x}_T$	Sequence of hidden states
$\boldsymbol{y}_T$	Sequence of observations
$\boldsymbol{z}_T$	Sequence of operating regimes
λ_h	Log-likelihood slopes
$\mathbb{R}_+$	Set of positive real numbers
$\ell_k(\cdot,\cdot)$	Log-likelihood of first *k* observation sequences
$\pi_k(\cdot \mid \cdot)$	Posterior density given first *k* observation sequences

References

[1] A. Heng, S. Zhang, A. Tan, and J. Mathew, “Rotating machinery prognostics: State of the art, challenges and opportunities,” *Mech. Syst. and Signal Process.*, vol. 23, no. 3, pp. 724–739, Apr. 2009.

[2] J. Lee, J. Ni, D. Djurdjanovic, H. Qiu, and H. Liao, “Intelligent prognostics tools and e-maintenance,” *Comput. Ind.*, vol. 57, no. 6, pp. 476-489, 2006.

[3] Z. Yang, “Dynamic maintenance scheduling using online information about system condition,” Ph.D. Dissertation, Dept. Mech. Eng., Univ. Michigan, Ann Arbor, MI, 2005.

[4] Z. Yang, D. Djurdjanovic, and J. Ni, “Maintenance scheduling for a manufacturing system of machines with adjustable throughput,” *IIE Trans.*, vol. 39, no. 12, pp. 1111-25, 2007.

[5] A. K. S. Jardine, D. Lin, and D. Banjevic, “A review on machinery diagnostics and prognostics implementing condition-based maintenance,” *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, Oct. 2005.

[6] L. Gardner, “The quantum hydrodynamic model for semiconductor devices,” *SIAM J. Appl. Math.*, vol. 54, no. 2, pp. 409–427, 1994.

[7] M. Doyle, T. Fuller, and J. Newman, “Modeling of Galvanostatic charge and discharge of the lithium/polymer/insertion cell,” *J. Electrochem. Soc.*, vol. 140, no. 6, pp. 15-26, 1993.

[8] Y. W. Kim, G. Rizzoni, and V. Utkin, "Automotive engine diagnosis and control via nonlinear estimation," *IEEE Control Systems*, vol. 18, no. 5, pp. 84-99, Oct. 1998.

[9] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[10] B. Leroux, “Maximum-likelihood estimation for hidden Markov models,” *Stoch. Process. their Appl.*, vol. 40, pp. 127–143, 1992.

[11] P. Bickel, Y. Ritov, and T. Ryden, “Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models,” *Ann. Statist.*, vol. 26, no. 4, pp. 1614–1635, 1998.

[12] P. Dempster, N. M. Lair, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *R. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1976.

[13] D. Montgomery, *Introduction to Statistical Quality Control*, 7th ed. Hoboken, NJ: Wiley, 2013.

[14] M. E. Cholette, M. Celen, D. Djurdjanovic, and J. D. Rasberry, “Condition monitoring and operational decision making in semiconductor manufacturing,” *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 4, pp. 454–464, 2013.

[15] P. Diaconis and D. Freedman, “On the consistency of Bayes estimates,” *Ann. Statist.*, vol. 14, no. 1, pp. 63–67, 1986.

[16] A. Wald, “Note on the consistency of the maximum likelihood estimate,” *Ann. Math. Statist.*, vol. 20, no. 4, pp. 595–601, 1949.

[17] O. Cappe, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models.* New York-Berlin: Springer, 2005.

[18] M. C. M. De Gunst and O. Shcherbakova, “Asymptotic behavior of Bayes estimators for hidden Markov models with application to ion channels,” *Math. Methods Statist.*, vol. 17, no. 4, pp. 342–356, 2009.

[19] M. Kano and Y. Nakagawa, “Data-based process monitoring, process control, and quality improvement: recent developments and applications in steel industry,” *Comput. Chem. Eng.*, vol. 32, no. 1, pp. 12–24, 2008.

[20] S. J. Qin, “Survey on data-driven industrial process monitoring and diagnosis,” *Annu. Rev. Control*, vol. 36, no. 2, pp. 220–234, 2012.

[21] C. Pantelides and J. G. Renfro, “The online use of first-principles models in process operations: review, current status and future needs,” *Comput. Chem. Eng.*, vol. 51, pp. 136–148, 2013.

[22] M. E. Cholette and D. Djurdjanovic, “Precedent-free fault isolation in a diesel engine exhaust gas recirculation system,” *J. Dyn. Syst. Meas. Control*, vol. 134, no. 3, pp. 031007, 2012.

[23] M. Musselman and D. Djurdjanovic, “Time-frequency distributions in the classification of epilepsy from EEG signals,” *Expert Syst. Appl.*, vol. 39, no. 13, pp. 11413–11422, 2012.

[24] A. Bleakie and D. Djurdjanovic, “Feature extraction, condition monitoring, and fault modeling in semiconductor manufacturing systems,” *Comput. Ind.*, vol. 64, no. 3, pp. 203–213, 2013.

[25] T. A. Johansen and B. A. Foss, “Operating regime based process modeling and identification,” *Comput. Chem. Eng.*, vol. 21, no. 2, pp. 159–176, 1997.

[26] J. Lee, E. Lapira, B. Bagheri, and H.-a. Kao, “Recent advances and trend in predictive manufacturing systems in big data environment,” *Manuf. Lett.*, vol. 1, no. 1, pp. 38–41, 2013.

[27] C. Aldrich and L. Auret, *Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods*. Springer, 2013.

[28] B. M. Colosimo and E. D. Castillo, *Bayesian Process Monitoring, Control and Optimization*. Chapman and Hall/CRC, 2007.

[29] M. S. Nikulin, N. Limnios, and N. Balakrishnan, *Advances in Degradation Modeling: Applications to Reliability, Survival Analysis, and Finance*. Birkhäuser, 2009.

[30] C. Bunks, D. McCarthy, and T. Al-Ani, "Condition-based maintenance of machines using hidden Markov models," *Mech. Syst. Signal Process.*, vol. 14, no. 4, pp. 597–612, 2000.

[31] H. M. Ertunc and C. Oysu, "Drill wear monitoring using cutting force signals," *Mechatronics*, vol. 14, no. 5, pp. 533–548, 2004.

[32] Z. Li, Z. Wu, Y. He, and C. Fulei, "Hidden Markov model-based fault diagnostics method in speed-up and speed-down process for rotating machinery," *Mech. Syst. Signal Process.*, vol. 19, no. 2, pp. 329–339, 2005.

[33] V. Purushotham, S. Narayanan, and S. A. N. Prasad, "Multi-fault diagnosis of rolling bearing elements using wavelet analysis and hidden Markov model based fault recognition," *NDT&E Int.*, vol. 38, no. 2, pp. 654–664, 2005.

[34] G. J. Hahn and W. Q. Meeker, *Statistical Intervals: A Guide for Practitioners*. Wiley-Interscience, 1991.

[35] T. C. Lystig and J. P. Hughes, "Exact computation of the observed information matrix for hidden Markov models," *J. Comput. Graph. Statist.*, vol. 11, no. 3, pp. 678–689, 2002.

[36] T. Aittokallio and E. Uusipaikka, "Computation of standard errors for maximum-likelihood estimates in hidden Markov models", Turku, Finland, Turku Centre for Computer Science, Tech. Rep. 379, 2000.

[37] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan, "A scalable bootstrap for massive data," *J. Roy. Statist. Soc.: Ser. B (Statist. Methodol.),* vol. 76, no. 4, pp. 795-816, 2014.

[38] I. Visser, M. E. Raijmakers, and P. C. Molenaar, "Confidence intervals for hidden Markov model parameters," *Br. J. Math. Statist. Psychol.*, vol. 53, pp. 317–327, 2000.

[39] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statist. Sci.*, vol. 1, no. 1, pp. 54 –77, 1986.

[40] W. Zucchini and I. MacDonald, *Hidden Markov Models for Time Series: an Introduction using R*. CRC, 2009.

[41] C. P. Robert, *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. New York: Springer Verlag, 2007.

[42] O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York-Berlin: Springer, 1985.

[43] C. P. Robert, G. Celeux, and J. Diebolt, “Bayesian estimation of hidden Markov chains: a stochastic implementation,” *Statist. Probab. Lett.*, vol. 16, no. 1, pp. 77–83, Jan. 1993.

[44] C. P. Robert, T. Rydén, and D. M. Titterington, “Convergence controls for MCMC algorithms with applications to hidden Markov chains,” *J. Statist. Comput. Simul.*, vol. 64, no. 4, pp. 327–355, 1999.

[45] H. Teigen and M. Jørgensen, “When 90% confidence intervals are 50% certain: On the credibility of credible intervals,” *Appl. Cogn. Psychol.*, vol. 19, no. 4, pp. 455–475, 2005.

[46] T. Rydén, “EM versus Markov chain Monte Carlo for estimation of hidden Markov models: a computational perspective,” *Bayesian Anal.*, vol. 3, no. 4, pp. 659–688, Dec. 2008.

[47] Á. Martínez-Beneito, D. Conesa, A. López-Quílez, and A. López-Maside, “Bayesian Markov switching models for the early detection of influenza epidemics,” *Statist. Med.*, vol. 28, no. 27, pp. 4455–4468, 2008.

[48] J. D. Chodera, P. Elms, F. Noé, B. Keller, C. M. Kaiser, A. Ewall-Wice, S. Marqusee, C. Bustamante, and N. S. Hinrichs, “Bayesian hidden Markov model analysis of single-molecule force spectroscopy: Characterizing kinetics under measurement uncertainty,” *arXiv preprint arXiv:* 1108.1430, 2011.

[49] C. F. H. Nam, J. A. D. Aston, and A. M. Johansen, “Quantifying the uncertainty in change points,” *J. Time Ser. Anal.*, vol. 33, no. 5, pp. 807–823, 2012.

[50] C. F. H. Nam, J. A. D. Aston, I. A. Eckley, and R. Killick, “The uncertainty of storm season changes: quantifying the uncertainty of autocovariance changepoints,” *Technometrics*, vol. 57, no. 2, pp. 194–206, June 2014.

[51] J. F. MacGregor and T. Kourti, “Statistical process control of multivariate processes,” *Control Eng. Pract.*, vol. 3, no. 3, pp. 403–414, 1995.

[52] G. Box and T. Kramer, “Statistical process monitoring and feedback adjustment: a discussion,” *Technometrics*, vol. 34, no. 3, pp. 251–267, 1992.

[53] F. Zhang, “Statistical data control chart for stationary process,” *Technometrics*, vol. 40, no. 1, pp. 24–38, 1998.

[54] S. Gu, J. Ni, and J. Yuan, “Non-stationary signal analysis and transient machining process condition monitoring,” *Int. J. Mach. Tools Manuf.*, vol. 42, no. 1, pp. 41–51, 2002.

[55] H. Qiu, H. Liao, and J. Lee, "Degradation assessment for machinery prognostics using hidden Markov models", in *Proc. ASME Int. Conf. Mech. Vibr. Noise*, Long Beach, CA, Sep. 24-28, 2005, pp. 531-637.

[56] J. Yu and S. J. Qin, "Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models," *AIChE J.*, vol. 54, no. 7, pp. 1811–1829, May 2008.

[57] S. J. Qin, G. Cherry, R. Good, J. Wang, and C. A. Harrison, "Semiconductor manufacturing process control and monitoring: A fab-wide framework," *J. Process Control*, vol. 16, no. 3, pp. 179–191, 2006.

[58] R. I. Leine, D. H. van Campen, and W. J. G. Keultjes, "Stick-slip whirl interaction in drillstring dynamics," *J. Vib. Acoust.*, vol. 124, no. 2, p. 209, Apr. 2002.

[59] W. Bartelmus and R. Zimroz, "A new feature for monitoring the condition of gearboxes in non-stationary operating conditions," *Mech. Syst. Signal Process.*, vol. 23, no. 5, pp. 1528–1534, Jul. 2009.

[60] B. Sin and J. H. Kim, "Nonstationary hidden Markov model," *Signal Processing*, vol. 46, no. 1, pp. 31–46, 1995.

[61] Zhu, X. D. Zhang, Y. F. Hu, and D. Xie, "Nonstationary hidden Markov models for multiaspect discriminative feature extraction from radar targets," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 2203–2214, 2007.

[62] K. Bae, B. K. Mallick, and C. G. Elsik, "Prediction of protein interdomain linker regions by a nonstationary hidden Markov model," *J. Amer. Statist. Assoc.*, vol. 103, no. 483, pp. 1085–1099, Sept. 2008.

[63] P. M. Djuric and J.-H. Chun, "An MCMC sampling approach to estimation of nonstationary hidden Markov models," *IEEE Trans. Signal Process.*, vol. 50, no. 5, pp. 1113–1123, 2002.

[64] V. S. Barbu and N. Limnios, *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications: Their Use in Reliability and DNA Analysis by*. Springer, 2009.

[65] S.-Z. Yu, "Hidden semi-Markov models," *Artif. Intell.*, vol. 174, no. 2, pp. 215–243, Feb. 2010.

[66] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed. Academic Press, 1975.

[67] F. Camci and R. B. Chinnam, "Health-state estimation and prognostics in machining processes," *IEEE Trans. Autom. Sci. Eng.*, vol. 7, pp. 581-597, 2010.

[68] R. I. A. Davis, B. C. Lovell, and T. Caelli, "Improved estimation of hidden Markov model parameters from multiple observation sequences," *Proc. 16th Int'l Conf. Pattern Recognition (ICPR)*, 2002.

[69] D. A. Tobon-Mejia, K. Medjaher, N. Zerhouni, and G. Tripot, “A data-driven failure prognostics method based on mixture of Gaussians hidden Markov models,” *IEEE Trans. Rel.*, vol. 61, no. 2, pp. 491–503, Jun. 2012.

[70] E. Bibbona and S. Ditlevsen, “Estimation in discretely observed diffusions killed at a threshold,” *Scand. J. Statist.*, vol. 40, no. 2, pp. 274–293, 2013.

[71] A. Jasra, N. Kantas, and A. Persing, “Bayesian parameter inference for partially observed stopped processes,” *Statist. Comput.*, vol. 24, no. 1, pp. 1–20, Jan. 2014.

[72] K. Banachewicz, A. Lucas, and A. Van Der Vaart, “Modelling portfolio defaults using hidden Markov models with covariates,” *Econ. J.*, vol. 11, no. 1, pp. 155–171, 2008.

[73] S. F. Gray, “Modeling the conditional distribution of interest rates as a regime-switching process,” *J. Finan. Econom.*, vol. 42, pp. 27–62, Sep. 1996.

[74] T. Smith and P. Vounatsou, “Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models,” *Statist. Med.*, vol. 22, no. 10, pp. 1709–1724, 2003.

[75] V. Anisimov, H. Mass, M. Danhof, and O. Della Pasqua, “Analysis of responses inmigraine modelling using hidden Markov models,” *Statist. Med.*, vol. 26, no. 22, pp. 4163–4178, 2007.

[76] J. P. Hughes, P. Guttorp, and S. P. Charles, “A non-homogeneous hidden Markov model for precipitation occurrence,” *J. Roy. Statist. Soc.: Ser. C (Appl. Statist.)*, vol. 48, pp. 15–30, Feb. 1999.

[77] R. Mehrotra and A. Sharma, “A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences,” *J. Geophys. Res. D Atmos.*, vol. 110, no. 16, pp. 1–13, 2005.

[78] B. Rajagopalan, U. Lall, and D. G. Tarboton, “Nonhomogeneous Markov model for daily precipitation,” *J. Hydrologic Eng.*, vol. 1, no. 1. pp. 33–40, 1996.

[79] F. X. Diebold, J. H. Lee, and G. C. Weinbach, “Regime switching with time-varying transition probabilities,” in *Nonstationary Time Series Analysis and Cointegration*, C. Hargreaves, Ed. Oxford, U.K.: Oxford Univ. Press, 1993.

[80] J. Filardo and S. F. Gordon, “Business cycle durations,” *J. Econometrics*, vol. 85, no. 1, pp. 99–123, 1998.

[81] E. Cholette and D. Djurdjanovic, “Degradation modeling and monitoring of machines using operation-specific hidden Markov models,” *IIE Trans.*, vol. 46, no. 10, pp. 1107–1123, Mar. 2014.

[82] F. Bartolucci and A. Farcomeni, “Information matrix for hidden Markov models with covariates,” *Statist. Comput.*, vol. 25, no. 3, pp. 515–526, Feb. 2014.

[83] L. Meligkotsidou and P. Dellaportas, “Forecasting with non-homogeneous hidden Markov models,” *Statist. Comput.*, vol. 21, no. 3, pp. 439–449, May 2010.

[84] R. Isermann, *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*, Springer, 2005.

[85] J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*. 1998.

[86] L. H. Chiang, E. L. Russell, and R. D. Braatz, *Fault Detection and Diagnosis in Industrial Systems.* Springer, 2000.

[87] K. J. Aström, P. Albertos, M. Blanke, A. Isidori, and R. Schaufelberger, Walther Sanz, *Control of Complex Systems*. Springer-Verlag London, 2001.

[88] G. J. Vachtsevanos, F. Lewis, M. Roemer, A. Hess, and B. Wu, *Intelligent Fault Diagnosis and Prognosis for Engineering Systems.* Springer, 2006.

[89] M. Lee, S. J. Kim, Y. Hwang, and C. S. Song, “Diagnosis of mechanical fault signals using continuous hidden Markov model,” *Journal of Sound and Vibration*, vol. 276, no. 3–5, pp. 1065–1080, Sep. 2004.

[90] H. M. Ertunc, K. A. Loparo, and H. Ocak, “Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs),” *Int. J. Mach. Tools Manuf.*, vol. 41, no. 9, pp. 1363–1384, Jul. 2001.

[91] P. Baruah and R. B. Chinnam, “HMMs for diagnostics and prognostics in machining processes,” *Int. J. Prod. Res.*, vol. 43, no. 6, pp. 1275–1293, Mar. 2005.

[92] L. Wang, M. G. Mehrabi, and J. E. Kannatey-Asibu, “Hidden Markov model-based tool wear monitoring in turning,” *J. Manuf. Sci. Eng.*, vol. 124, no. 3, pp. 651–658, Aug. 2002.

[93] Y. Xu and M. Ge, “Hidden Markov model-based process monitoring system,” *J. Intell. Manuf.*, vol. 15, no. 3, pp. 337–350, Jun. 2004.

[94] C. Kwan, X. Zhang, R. Xu, and L. Haynes, “A novel approach to fault diagnostics and prognostics,” *Proc. lEEE ICRA*, Sep. 2003, vol. 1, no. 3, pp. 604–609.

[95] J. C. Wong, K. A. McDonald, and A. Palazoglu, “Classification of process trends based on fuzzified symbolic representation and hidden Markov models,” *J. Process Control*, vol. 8, no. 5–6, pp. 395–408, 1998.

[96] J. C. Wong, K. A. McDonald, and A. Palazoglu, “Classification of abnormal plant operation using multiple process variable trends,” *J. Process Control*, vol. 11, no. 4, pp. 409–418, 2001.

[97] W. Sun, A. Palazoğlu, and J. A. Romagnoli, “Detecting abnormal process trends by wavelet-domain hidden Markov models,” *AIChE J.*, vol. 49, no. 1, pp. 140–150, 2003.

[98] J. Chen and W.-J. Chang, "Applying wavelet-based hidden Markov tree to enhancing performance of process monitoring," *Chem. Eng. Sci.*, vol. 60, no. 18, pp. 5129–5143, 2005.

[99] J. Chen, T.-Y. Hsu, C.-C. Chen, and Y.-C. Cheng, "Online predictive monitoring using dynamic imaging of furnaces with the combinational method of multiway principal component analysis and hidden Markov model," *Ind. Eng. Chem. Res.*, vol. 50, no. 5, pp. 2946–2958, 2011.

[100] S. Zhou, J. Zhang, and S. Wang, "Fault diagnosis in industrial processes using principal component analysis and hidden Markov model," in *Proc. American Control Conf.*, Boston, MA, 2004, vol. 6, pp. 5680–5685.

[101] M. M. Rashid and J. Yu, "Hidden Markov model based adaptive independent component analysis approach for complex chemical process monitoring and fault detection," *Ind. Eng. Chem. Res.*, vol. 51, no. 15, pp. 5506-5514, Apr. 2012.

[102] K.-C. Kwon and J.-H. Kim, "Accident identification in nuclear power plants using hidden Markov models," *Eng. Appl. Artif. Intell.*, vol. 12, no. 4, pp. 491–501, Aug. 1999.

[103] E. Dorj, C. Chen, and M. Pecht, "A Bayesian hidden Markov model-based approach for anomaly detection in electronic systems," in *IEEE Aerosp. Conf. Proc.*, 2013, pp. 1–10.

[104] L. P. Heck and J. H. McClellan, "Mechanical system monitoring using hidden Markov models," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 1991, vol. 3, pp. 1697–1700.

[105] P. Smyth, "Hidden Markov-models for fault-detection in dynamic systems," *Pattern Recog.,* vol. 27, pp. 149-164, Jan. 1994.

[106] P. Li and V. Kadirkamanathan, "Particle filtering based likelihood ratio approach to fault diagnosis in nonlinear stochastic systems," *IEEE Trans. Syst. Man. Cybern.*, vol. 31, no. 3, pp. 337–343, 2001.

[107] P. Smyth, "Markov monitoring with unknown states," *IEEE J. Sel. Areas Commun.*, vol. 12, no. 9, pp. 1600–1612, 1994.

[108] C. Stauffer and W.E.L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pp. 246-252, 1999.

[109] S. Lee, L. Li, and J. Ni, "Online degradation assessment and adaptive fault detection using modified hidden Markov model," *J. Manuf. Sci. Eng.*, vol. 132, no. 2, p. 021010, 2010.

[110] P. M. Frank, "Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: a survey and some new results," *Automatica*, vol. 26, no. 3, pp. 459–474, 1990.

[111] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis, Part III: Process history based methods," *Comput. Chem. Eng.*, vol. 27, no. 3, pp. 327–346, 2003.

[112] E. Cholette, J. Liu, D. Djurdjanovic, and K. A. Marko, "Monitoring of complex systems of interacting dynamic systems," *Appl. Intell.*, vol. 37, no. 1, pp. 60–79, 2012.

[113] M. Fox, J. Gough, and D. Long, "Detecting execution failures using learned action models," in *Proc. Assoc. Adv. Artif. Intell.*, 2007, pp. 968–973.

[114] A. J. Brown, V. M. Catterson, M. Fox, D. Long, and S. D. J. McArthur, "Learning models of plant behavior for anomaly detection and condition monitoring," in *Proc. Int. Conf. Intell. Syst. Appl. Power Syst.*, 2007, pp. 1–6.

[115] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Comput. Surv.*, vol. 41, no.3 , pp. 1–72, 2009.

[116] M. Markou and S. Singh, "Novelty detection: A review - Part 1: Statistical approaches," *Signal Process.*, vol. 83, no. 12, pp. 2481–2497, 2003.

[117] M. Basseville and I. V. Nikiforov, *Detection of abrupt changes: theory and application*, vol. 104. Englewood Cliffs: Prentice Hall, 1993.

[118] R. MacKay Altman, "Assessing the goodness-of-fit of hidden Markov models," *Biometrics*, vol. 60, no. 2, pp. 444–450, 2004.

[119] F. LeGland and L. Mevel, "Fault detection in hidden Markov models : a local asymptotic approach," *Proc. 39th IEEE Conf. Decis. Control* (Cat. No.00CH37187), vol. 5, pp. 4686–4690, 2000.

[120] B. Chen and P. Willett, "Detection of hidden Markov model transient signals," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 36, no. 4, 2000.

[121] W. H. Woodall and D. C. Montgomery, "Research issues and ideas in statistical process control," *J. of Qual. Technol.*, vol. 31, no. 4. p. 11, 1999.

[122] W. Woodall and D. Montgomery, "Some current directions in the theory and application of statistical process monitoring," *J. Qual. Technol.*, vol. 46, no. 1, pp. 78–94, 2014.

[123] J. Borwanker, G. Kallianpur, and B. L. S. P. Rao, "The Bernstein-von Mises theorem for Markov processes," *Ann. Math. Statist.*, vol. 42, no. 4. pp. 1241–1253, 1971.

[124] C. C. Heyde and M. Johnstone, "On asymptotic posterior normality for stochastic processes," *J. R. Statist. Soc. B*, vol. 41, no. 2, pp. 184–189, 1979.

[125] E. Vernet, "Posterior consistency for nonparametric hidden Markov models with finite state space," *Electron. J. Statist.,* vol. 9, pp. 717-752, 2015.

[126] L. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite state Markov chains,” *Ann. Math. Statist.*, vol. 37, pp.1559-1563, 1966.

[127] T. Petrie, “Probabilistic functions of finite state Markov chains,” *Ann. Math. Statist.*, vol. 40, no. 1, pp. 97–115, 1969.

[128] T. Rydén, "Consistent and asymptotically normal parameter estimates for hidden Markov models," *Ann. Statist.,* vol. 22, pp. 1884-1895, 1994.

[129] F. LeGland and L. Mevel, “Exponential forgetting and geometric ergodicity in hidden Markov models,” *Proc. 36th IEEE Conf. Decis. Control*, vol. 1, 1997.

[130] F. LeGland and L. Mevel, “Asymptotic properties of the MLE in hidden Markov models,” *Control Conference (ECC)*, European, pp. 3440 - 3445, 1997.

[131] L. Mevel and L. Finesso, “Convergence rates of the maximum likelihood estimator of hidden Markov models,” *Proc. 39th IEEE Conf. Decis. Control*, vol. 5, 2000.

[132] L. Mevel and L. Finesso, “Asymptotical statistics of misspecified hidden Markov models,” *IEEE Trans. Automat. Contr.*, vol. 49, no. 7, pp. 1123–1132, 2004.

[133] L. Bordes and P. Vandekerkhove, “Statistical inference for partially hidden Markov models,” *Commun. Statist. Theory Methods*, vol. 34, no. 5, pp. 1081–1104, May 2005.

[134] P. Ailliot and F. Pène, "Consistency of the maximum likelihood estimate for non-homogeneous Markov–switching models," *ESAIM: Probability and Statistics,* vol. 19, pp. 268-292, 2015.

[135] D. Pouzo, Z. Psaradakis, and M. Sola, "Maximum likelihood estimation in possibly misspecified dynamic models with time inhomogeneous Markov regimes," https://arxiv.org/abs/1612.04932, Dec. 2016.

[136] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[137] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*. New York-Berlin:Springer, 2006.

[138] R. E. Kass and L. Wasserman, “The selection of prior distributions by formal rules,” *J. Amer. Statist. Assoc.*, vol. 91, no. 435, pp. 1343–1370, 1996.

[139] A. Agresti and D. B. Hitchcock, “Bayesian inference for categorical data analysis,” *Statist. Methods Appl.*, vol. 14, no. 3, pp. 297–330, 2005.

[140] A. Jasra, C. C. Holmes, and D. A. Stephens, “Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling,” *Statist. Sci.*, vol. 20, no. 1, pp. 50–67, 2005.

[141] E. Rodríguez and S. G. Walker, “Label switching in Bayesian mixture models: deterministic relabeling strategies,” *J. Comput. Graph. Statist.*, vol. 23, no. 1, pp. 25–45, 2014.

[142] F. J. Massey, Jr, “The Kolmogorov-Smirnov test for goodness of fit,” *J. Amer. Statist.*, vol. 46, no. 253, pp. 68-78, 1951.

[143] Silverman, *Density Estimation for Statistics and Data Analysis*. London-New York: Chapman and Hall, 1986.

[144] M. Basseville, “Distance measures for signal processing and pattern recognition,” *Signal Process.*, vol. 18, no. 4, pp. 349–369, 1989.

[145] *Semiconductor Manufacturing Handbook*. New York: McGraw-Hill, 2004.

[146] H. Thomas, G. Morfill, V. Demmel, J. Goree, B. Feuerbacher, and D. Möhlmann, “Plasma crystal: Coulomb crystallization in a dusty plasma,” *Phys. Rev. Lett.*, vol. 73, no. 5, pp. 652–655, 1994.

[147] A. Bleakie and D. Djurdjanovic, “Feature extraction, condition monitoring, and fault modeling in semiconductor manufacturing systems,” *Comput. Ind.*, vol. 64, no. 3, pp. 203-213, 2013.

[148] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, 1936.

[149] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, “Engineering applications of the self-organizing map,” *Proc. IEEE*, vol. 84, no. 10, pp. 1358-1383, 1996.

[150] J. F. MacGregor and T. Kourti, “Statistical process control of multivariate processes,” *Control Eng. Pract.*, vol. 3, no. 3, pp. 403–414, 1995.

[151] P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

[152] D. Zhang, A. D. Bailey III, and D. Djurdjanovic, "Bayesian identification of hidden Markov models and their use for condition-based monitoring," *IEEE Trans. Rel.,* vol. 65, no. 3, pp. 1471-1482, Jun. 2016.

[153] H. Pham and H. Wang, "Imperfect maintenance," *European journal of operational research*, vol. 94, pp. 425-438, 1996.

[154] W. Kern, "The evolution of silicon wafer cleaning technology," *Journal of the Electrochemical Society,* vol. 137, pp. 1887-1892, 1990.

[155] V. Wong, B. C. Richardson, A. Lui, and S. Baldwin, "End point determination of process residues in wafer-less auto clean process using optical emission spectroscopy," 2004.

[156] L. Wang, M. G. Mehrabi, and E. Kannatey-Asibu, “Hidden Markov model-based tool wear monitoring in turning,” *J. Manuf. Sci. Eng.*, vol. 124, no. 3, pp. 651–658, 2002.

[157] H. Ocak and K. A. Loparo, “HMM-based fault detection and diagnosis scheme for rolling element bearings,” *J. Vib. Acoust.*, vol. 127, no. 4, pp. 299–306, 2005.

[158] G. D. Forney, "The Viterbi algorithm," *Proc. of the IEEE*, vol. 61, pp. 268-278, 1973.

[159] A. H. Tai, W.-K. Ching, and L. Y. Chan, "Detection of machine failure: Hidden Markov Model approach," *Computers & Industrial Engineering*, vol. 57, pp. 608-619, Sep 2009.

[160] A. Giantomassi, F. Ferracuti, A. Benini, G. Ippoliti, S. Longhi, and A. Petrucci, "Hidden Markov Model for Health Estimation and Prognosis of Turbofan Engines," *International Design Engineering Technical Conferences & Computers and Information in Engineering Conference,* pp. 681-689, 2011.

[161] D. Hernando, V. Crespi, and G. Cybenko, "Efficient computation of the hidden Markov model entropy for a given observation sequence," *IEEE Transactions on Information Theory,* vol. 51, pp. 2681-2685, 2005.

[162] S. L. Scott, "Bayesian methods for hidden Markov models: Recursive computing in the 21st century," *J. Amer. Statist. Assoc.,* vol. 97, no. 457, pp. 337-351, Mar. 2002.

[163] M. Yuwono, Y. Qin, J. Zhou, Y. Guo, B. G. Celler, and S. W. Su, "Automatic bearing fault diagnosis using particle swarm clustering and Hidden Markov Model," *Engineering Applications of Artificial Intelligence,* vol. 47, pp. 88-100, 2016.

[164] M. Celen, "Joint maintenance and production operations decision making in flexible manufacturing systems", Ph.D. Thesis, The University of Texas at Austin, Austin, TX, 2016.

[165] E. L. Lehmann and G. Casella, *Theory of Point Estimation*: Springer Science & Business Media, 1998.

[166] M. J. Schervish, *Theory of Statistics*. New York: Springer-Verlag, 1995.

[167] H. Teicher, "Identifiability of mixtures of product measures," *The Annals of Mathematical Statistics,* vol. 38, pp. 1300-1302, 1967.

[168] R. I. Jennrich, "Asymptotic properties of non-linear least squares estimators," *The Annals of Mathematical Statistic,* vol. 40, pp. 633-643, 1969.

[169] C. J. Adcock, "Sample size determination: a review," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 46, pp. 261-283, 1997.

[170] H. Lebesgue, "Intégrale, longueur, aire." *Annali di Matematica Pura ed Applicata (1898-1922)* 7.1 (1902): 231-359.

[171] P. Billingsley, *Probability and Measure*, 3rd ed. New York: J. Wiley & Sons, 1995.

[172] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.

[173] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA:Thomson Learning, 2002.