Contribution to deterioration modeling and residual life estimation based on condition monitoring data

Thanh Trung Le

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Automatique - Productique**

Arrêté ministériel : 7 août 2006

Présentée par

Thanh Trung LE

Thèse dirigée par **M. Christophe BERENGUER**
et codirigée par **M. Florent CHATELAIN**

préparée au sein du **GIPSA-lab, département Automatique**
dans **l'école doctorale Electronique, Electrotechnique, Automatique, Traitement du signal (EEATS)**

Contribution to deterioration modeling and residual life estimation based on condition monitoring data

Thèse soutenue publiquement le **08 Décembre 2015**,
devant le jury composé de :

M. Olivier GAUDOIN
Professeur, Institut Polytechnique de Grenoble, Examinateur

M. Mustapha OULADSINE
Professeur, Université d'Aix-Marseille, Examinateur

M. Kamal MEDJAHER
Maître de conférences, HDR, Ecole nationale supérieure de mécanique et des microtechniques, Institut FEMTO-ST, Rapporteur

Mme. Mitra FOULADIRAD
Maître de conférences, HDR, Université de Technologie de Troyes, Rapporteur

M. Christophe BERENGUER
Professeur, Institut Polytechnique de Grenoble, Directeur de thèse

M. Florent CHATELAIN
Maître de conférences, Institut Polytechnique de Grenoble, Co-encadrant de thèse

Mme. Sophie SIEG-ZIEBA
Docteur, Centre technique des industries mécaniques CETIM, Invitée

Acknowledgments

This manuscript represents my three-year-work as a PhD student at GIPSA-lab, Grenoble INP. I would like to express my gratitude to all those who helped me during my doctoral studies and the writing of this thesis.

Foremost, my deepest gratitude goes to my supervisors Prof. Christophe Bérenguer and Dr Florent Chatelain whose support and guidance made my thesis work possible. Christophe and Florent, thank you for trusting, encouraging, listening to, guiding, advising me and for many re-lectures. During these three years of thesis, I learned a lot from you, not only in scientific aspect but also in many other things. I have to say that I am really lucky being working under your supervision.

I would also like to thank members of my PhD committee who reviewed and provides constructive feedbacks to my work. I am grateful for their insightful comments and questions. In particularly, I would like to thank Dr Kamal Mehjaher and Dr Mitra Foulardirad for having accepted to be reviewers and for their valuable comments on my manuscript.

I wish to express my gratitude to European commission for the financial support over this period. I also want to send my thanks to all the partners of the SUPREME project for their supports and for the nice time we spent together through several meetings.

To my colleagues from Grenoble Images Speech Signal and Control laboratory (GIPSA-lab), many thanks for their help and for the excellent research atmosphere. I should not forget to acknowledge my Vietnamese friends, not only in France but also in Vietnam who have been always by my side and support me during my stay in France.

Last but not least, I offer sincere thanks to my family. Words can not express how grateful I am to my parents for their love, their encouragement and great confidence in me all through these years. A special thank to my beloved wife and my little boy for their endless love which motivated and supported me a lot.

Thanh Trung LE

Abstract —

Predictive maintenance plays a crucial role in maintaining continuous production systems since it can help to reduce unnecessary intervention actions and avoid unplanned breakdowns. Indeed, compared to the widely used condition-based maintenance (CBM), the predictive maintenance implements an additional prognostics stage. The maintenance actions are then planned based on the prediction of future deterioration states and residual life of the system. In the framework of the European FP7 project SUPREME (Sustainable PREdictive Maintenance for manufacturing Equipment), this thesis concentrates on the development of stochastic deterioration models and the associated remaining useful life (RUL) estimation methods in order to be adapted in the project application cases.

Specifically, the thesis research work is divided in two main parts. The first one gives a comprehensive review of the deterioration models and RUL estimation methods existing in the literature. By analyzing their advantages and disadvantages, an adaption of the state of the art approaches is then implemented for the problem considered in the SUPREME project and for the data acquired from a project's test bench. Some practical implementation aspects, such as the issue of delivering the proper RUL information to the maintenance decision module are also detailed in this part.

The second part is dedicated to the development of innovative contributions beyond the state-of-the-art in order to develop enhanced deterioration models and RUL estimation methods to solve original prognostics issues raised in the SUPREME project. Specifically, to overcome the co-existence problem of several deterioration modes, the concept of the "multi-branch" models is introduced. It refers to the deterioration models consisting of different branches in which each one represent a deterioration mode. In the framework of this thesis, two multi-branch model types are presented corresponding to the discrete and continuous cases of the systems' health state. In the discrete case, the so-called Multi-branch Hidden Markov Model (Mb-HMM) and the Multi-branch Hidden semi-Markov model (Mb-HsMM) are constructed based on the Markov and semi-Markov models. Concerning the continuous health state case, the Jump Markov Linear System (JMLS) is implemented. For each model, a two-phase framework is carried out for both the diagnostics and prognostics purposes. Through numerical simulations and a case study, we show that the multi-branch models can help to take into account the co-existence problem of multiple deterioration modes, and hence give better performances in RUL estimation compared to the ones obtained by standard "single branch" models.

Keywords: Deterioration modeling, remaining useful life, prognostics, diagnostics, predictive maintenance, Hidden Markov Model, HMM, Hidden semi-Markov model, HsMM, Jump Markov linear systems, JMLS.

Résumé —

La maintenance prédictive joue un rôle important dans le maintien des systèmes de production continue car elle peut aider à réduire les interventions inutiles ainsi qu'à éviter des pannes imprévues. En effet, par rapport à la maintenance conditionnelle, la maintenance prédictive met en œuvre une étape supplémentaire, appelée le pronostic. Les opérations de maintenance sont planifiées sur la base de la prédiction des états de détérioration futurs et sur l'estimation de la vie résiduelle du système. Dans le cadre du projet européen FP7 SUPREME (Sustainable PREdictive Maintenance for manufacturing Equipment en Anglais), cette thèse se concentre sur le développement des modèles de détérioration stochastiques et sur des méthodes d'estimation de la vie résiduelle (Remaining Useful Life – RUL en anglais) associées pour les adapter aux cas d'application du projet.

Plus précisément, les travaux présentés dans ce manuscrit sont divisés en deux parties principales. La première donne une étude détaillée des modèles de détérioration et des méthodes d'estimation de la RUL existant dans la littérature. En analysant leurs avantages et leurs inconvénients, une adaptation d'une approche de l'état de l'art est mise en œuvre sur des cas d'études issus du projet SUPREME et avec les données acquises à partir d'un banc d'essai développé pour le projet. Certains aspects pratiques de l'implémentation, à savoir la question de l'échange d'informations entre les partenaires du projet, sont également détaillées dans cette première partie.

La deuxième partie est consacrée au développement de nouveaux modèles de détérioration et les méthodes d'estimation de la RUL qui permettent d'apporter des éléments de solutions aux problèmes de modélisation de détérioration et de prédiction de RUL soulevés dans le projet SUPREME. Plus précisément, pour surmonter le problème de la coexistence de plusieurs modes de détérioration, le concept des modèles « multi-branche » est proposé. Dans le cadre de cette thèse, deux catégories des modèles de type multi-branche sont présentées correspondant aux deux grands types de modélisation de l'état de santé des systèmes, discret ou continu. Dans le cas discret, en se basant sur des modèles markoviens, deux modèles nommés Mb-HMM and Mb-HsMM (Multi-branch Hidden (semi-)Markov Model en anglais) sont présentés. Alors que dans le cas des états continus, les systèmes linéaires à sauts markoviens (JMLS) sont mis en œuvre. Pour chaque modèle, un cadre à deux phases est implémenté pour accomplir à la fois les tâches de diagnostic et de pronostic. A travers des simulations numériques, nous montrons que les modèles de type multi-branche peuvent donner des meilleures performances pour l'estimation de la RUL par rapport à celles obtenues par des modèles standards mais « mono-branche ».

Mots clés : Deterioration, vie résiduelle, pronostic, diagnostic, maintenance prédictive, modèle de Markov caché, systèmes linéaires à sauts Markoviens.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Industrial maintenance: from corrective to predictive

Maintenance costs are a major part of the total operating costs of all manufacturing or production plants. Depending on the specific industry, maintenance costs can represent between 15 and 60 percent of the cost of goods produced [104]. For a lot of sectors, maintenance is a major issue, with a ratio between maintenance costs and added value higher than 25% [139].

The European standard NF EN 13306 X60-319 defines the maintenance as "all of technical, administrative and management activities throughout the life cycle of an asset in order to maintain or restore it to the state where it can perform the intended functions" [113]. Intuitively, "maintain" refers to maintenance activities implemented on a working system in order to prevent it from a breakdown. While "restore" implies the "correction" activities that are conducted once the system failed already. Accordingly to these two activity types, maintenance methods can be classified into two main groups: corrective and preventive, see Figure 1.1.

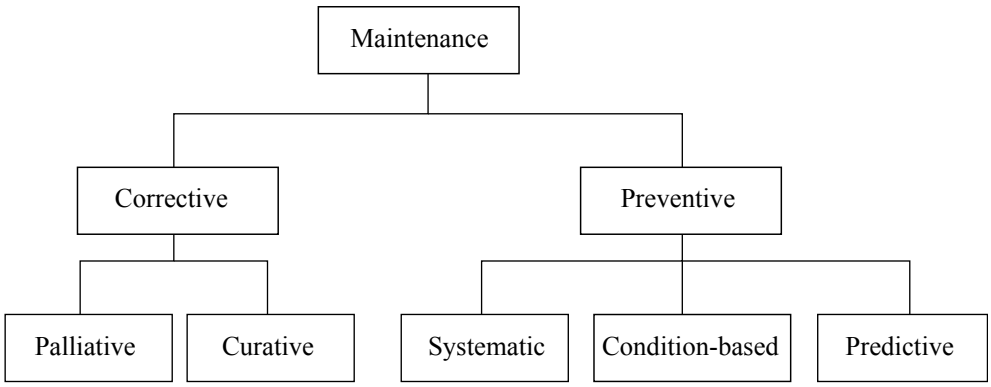


Figure 1.1: Maintenance methods classification

1.1.1 Corrective maintenance

Corrective maintenance is defined as "maintenance carried out after failure detection and is aimed at restoring an asset to a condition in which it can perform its intended function" [113]. It can be palliative or curative.

1.1.1.1 Palliative maintenance

Palliative maintenance is commonly called troubleshooting. It refers to actions that allow a failed component or equipment to accomplish temporarily all or a part of its requested functions. Palliative maintenance is mainly carried out while waiting for curative ones.

1.1.1.2 Curative maintenance

The curative maintenance seeks the underlying causes of failure and repairs the failed component(s) for the aim of bringing it to the initial state. In contrast to the palliative ones, the result of the curative activities should have a permanent effect.

1.1.2 Preventive maintenance

According to the European standard NF EN 13306 X60-319, preventive maintenance is defined as "maintenance performed at predetermined intervals or according to prescribed criteria and intended to reduce the probability of failure or degradation of the operation of an asset" [113]. The preventive maintenance aims to increase the availability and the reliability of the system of interest. In addition, preventive maintenance methods can also help to reduce the maintenance cost, especially by limiting unplanned breakdowns. It consists of three types: systematic, condition-based and predictive.

1.1.2.1 Systematic maintenance

Systematic maintenance is the preventive maintenance that is performed at predetermined intervals of time or a defined number of units of use but without control of the condition of the property. If the maintenance activities are planned at specified calendar times, it is called clock-based or calendar-based maintenance. Otherwise, if they are implemented at specified age of the item, we have the age-based maintenance [131]. By intervening at fixed times in the future, this type of maintenance can prevent some catastrophic failures from occurring. However, this could also lead to some unnecessary repairs when machine or system still remain in a good condition. In addition, some criti-

cal failures may still happen before the planned dates of intervention. In such cases, the machine must be repaired using the corrective methods.

1.1.2.2 Condition-base maintenance (CBM)

CBM is a preventive maintenance program that is widely studied in the literature. It is based on using real-time data to prioritize and optimize maintenance resources. The state of the system is observed by condition monitoring system. Such a system will determine the equipment's health, and act only when one or more indicators show that equipment is going to fail or that equipment performance is deteriorating [67]. Compared to the systematic one, the CBM attempts to maintain the correct equipment at the right time. The CBM can be therefore shortly described as "maintenance when need arises". However, its implementation is more complex and expensive than the corrective and the systematic maintenance methods.

1.1.2.3 Predictive maintenance

Predictive maintenance is the most advanced maintenance program that is currently developed in the literature [104]. Within a predictive maintenance program, maintenance actions are planned based on the prediction of system's health state in the future and/or the time at which the system will fail. This is where the word "predictive" comes from. Intuitively, by knowing the future state of the system, the predictive maintenance could bring higher maintenance performance than the CBM in terms of reducing unnecessary intervention actions, planing the purchase of spare parts and hence reducing the maintenance cost. Moreover, by predicting the time to failure of the monitored system, the predictive maintenance could also avoid unplanned breakdowns.

In the literature, the predictive maintenance can be considered as a new development of the CBM method.

Figure 1.2 summarizes the evolution and the development of the maintenance methods in the literature. Compared to the other ones, the predictive maintenance can give more benefits (in terms of the availability of system, the save of maintenance costs, ect.). However, the price to pay for achieving such benefits is also the highest.

For continuous production industries such as food, cement or paper, it is well-known that unscheduled downtime has a direct impact on energy consumption and can affect the quality of the production by generating "off-specs" products. Ensuring continuity of production is therefore one of the key requirements in order to improve the productivity of a manufacturing plant. Predictive maintenance is hence a critical issue since it can help to anticipate the failure and give an optimal planning of maintenance operation. If being correctly implemented, the predictive maintenance could help to reduce drastically the maintenance costs and ensure an optimal use of each equipment, including up-time,

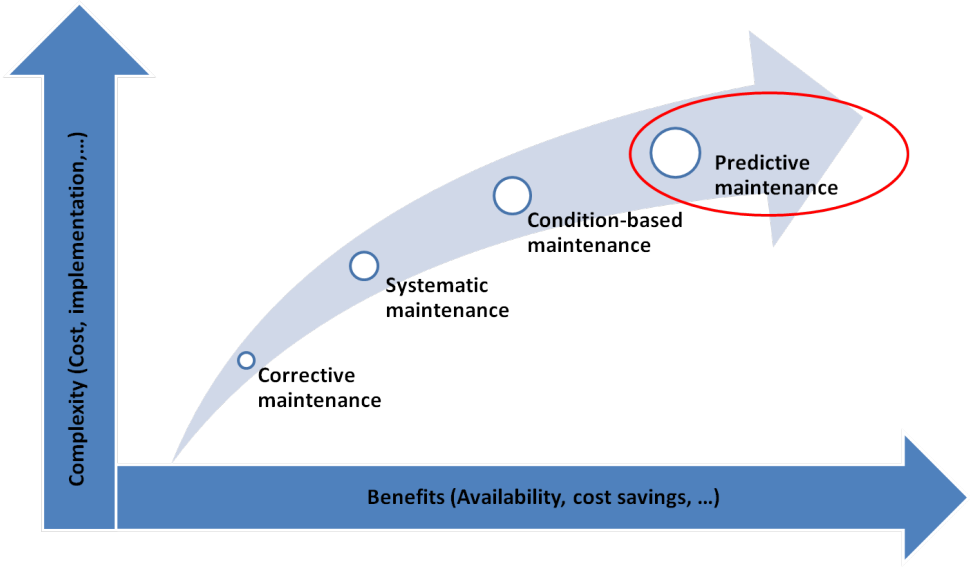


Figure 1.2: Maintenance methods evolution

production quality and energy consumption.

1.2 SUstainable PREdictive Maintenance for manufacturing Equipement - SUPREME project

The SUPREME project is one of the European projects funded under the Framework Program - FP7 and was officially launched on 01/09/2012. The overall objective of the project is to provide new tools for predictive maintenance, which will also reduce energy consumption, and can then be qualified as sustainable predictive maintenance. Specifically, SUPREME provides new tools to adapt dynamically the maintenance and operation strategies to the current condition of the critical components in production equipment. It develops also a reference model to achieve an integrated approach to optimal energy consumption by means of predictive maintenance tools.

SUPREME is a collaborative project of ten partners coming from different countries in Europe. The partners are specialists in different domains and come from both industrial and academical areas, among which Grenoble Institute of Technology - Grenoble INP, Table 1.1. With the presence of a partner coming from the paper mill Condat in France, the proposed tools and the developed methods are firstly tested and applied in paper industry. However, they could be also useful in other sectors with similar problems (i.e. with continuous production requirements) such as power generation, chemical, food industry, and cement.

Table 1.1: SUPREME project partners

Participant organization name	Short name	**Country**
Centre Technique des Industries Mecaniques	CETIM	France
Loy & Hutz AG	L& H	Germany
Fraunhofer-Gesellschaft Zur Foerderung Der Angewandten Forschung E.V	FhG-IPA	Germany
Optimitive SL	Optimitive	Spain
EC Systems SPZOO	ECS	Poland
Institut Polytechnique de Grenoble	Grenoble INP	France
Orloga SA	Orloga	Spain
Condat	Condat	France
ENDEL SAS	ENDEL	France
Ceske Vysoke Uceni Technicke V Praze	CVUT	Czech Republic

1.2.1 Concept and objectives of the project

1.2.1.1 SUPREME concept

As already presented, the SUPREME project aims to develop new tools for predictive maintenance, which will also reduce energy consumption, and can then be qualified of sustainable predictive maintenance. In order to better understand the concept of the project, the predictive maintenance implementation should firstly be represented. Figure 1.3 summarizes the five main stages of a predictive maintenance program.

The five stages include: data pre-processing, feature extraction, diagnostics, prognostics and decision making. The data (information) from targeted physical system is collected by sensors and stored through a process called data acquisition. In practice, however, such collected data are rarely useful in their raw form because they may contain the measurement noises and errors. Thus the data must be pre-processed, i,e. filtered, compressed, correlated, etc., in order to remove artifacts and reduce noise levels as well as the volume of data to be processed subsequently. Once the pre-processing module confirms that the sensor data are "clean" and formatted appropriately, features or signatures of normal or faulty conditions can be extracted [144]. The procedure of extracting useful information is called feature extraction. At this stage, various signal processing techniques such as spectral analysis, Wavelet Packet Decomposition (WPD) algorithm, etc. could be implemented. The extracted-features are the essential inputs to fault diagnostics algorithms.

The next stage is the diagnostics in which faults could be detected and isolated. This stage, together with the data pre-processing and feature extraction stages, can be implemented in a Condition Monitoring System (CMS). Based on the extracted features and/or

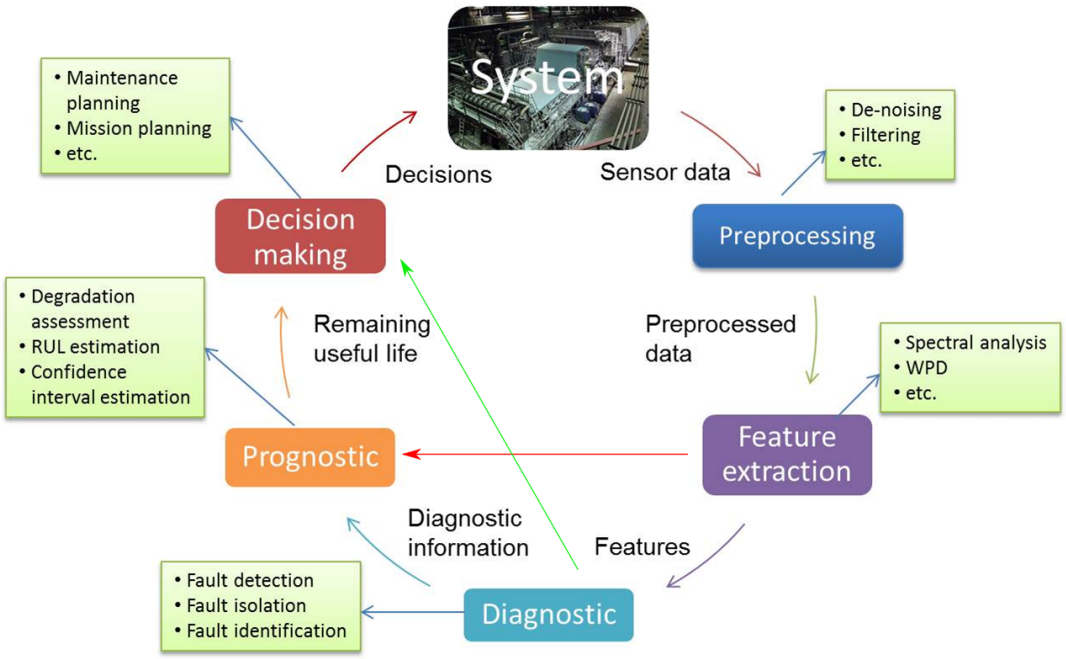


Figure 1.3: Illusatration of a predictive maintenance program

the diagnostics information, the prognostics could be implemented. The main objective of this stage is to estimate the remaining useful life (RUL) of the monitored system. For this end, a possible solution is to construct a deterioration model that represent the temporal evolution of a default. Then basing on this model, one can assess the actual level of deterioration of the system as well as predict its future evolution for the purpose of estimating the system's RUL. As the prediction is always uncertain, uncertainties should be characterized in this stage, i.e. with the help of confidence intervals. It is the prognostics stage that makes the predictive maintenance be more advanced in comparison with the condition-based maintenance one.

Decision making is the last stage for the implementation of a predictive maintenance program. In this stage, the maintenance decision, i.e. repair, overhaul, etc., can be made by basing on both the diagnostics information about the default of system and the RUL estimation result. This will help to avoid catastrophic failures as well as unnecessary interventions, or in other words, to save the maintenance cost and improve the availability of systems.

The concept of the SUPREME project is constructed from specific technical tasks of the five above stages. It is represented in Figure 1.4.

The realization of the SUPREME predictive maintenance concept requires the developments of new solutions based on affordable (components costs consistent with installations costs and/or down-time costs...) and intelligent technologies (smart sensors, embedded signal processing...) in relation with:

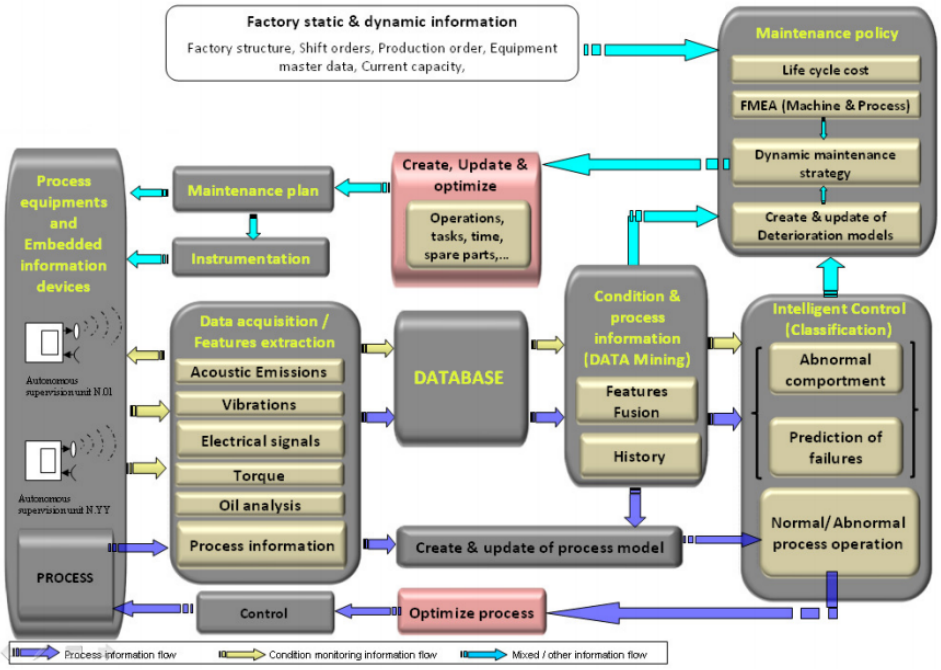


Figure 1.4: SUPREME project concept

- **Embedded Condition Monitoring Modules**: including sensor integration in mechanical components, robust wireless communication, generic detection technology suitable to non-stationary and noisy environment, robust data validation, advanced signal processing, automated configuration of the monitoring systems, interface to the automation system and simple but communicative user interface.
- **Reliability and Maintainability Modules**: including risk-based identification of critical components, condition and risk based maintenance planning, remaining useful life prediction, dynamic adaptation of maintenance intervals and tasks.
- **Intelligent control and data mining Modules**: taking into account process state and process condition data integration, knowledge-based assistance for optimal operation and maintenance, Artificial Intelligence and Data Mining for intelligent failure-mode data analysis.

These three modules and their specific tasks correspond to the three work packages 3, 4 and 5 of the project respectively and will be presented in more detailed in Section 1.2.2. They cover all the technical aspects of the SUPREME project and their relation with the project concept is shown in Figure 1.5.

1.2.1.2 Scientific and technology objectives

In order to reach a big step in the development of predictive maintenance, the SUPREME project aims to integrate and interface different tools (existing and to be developed), from

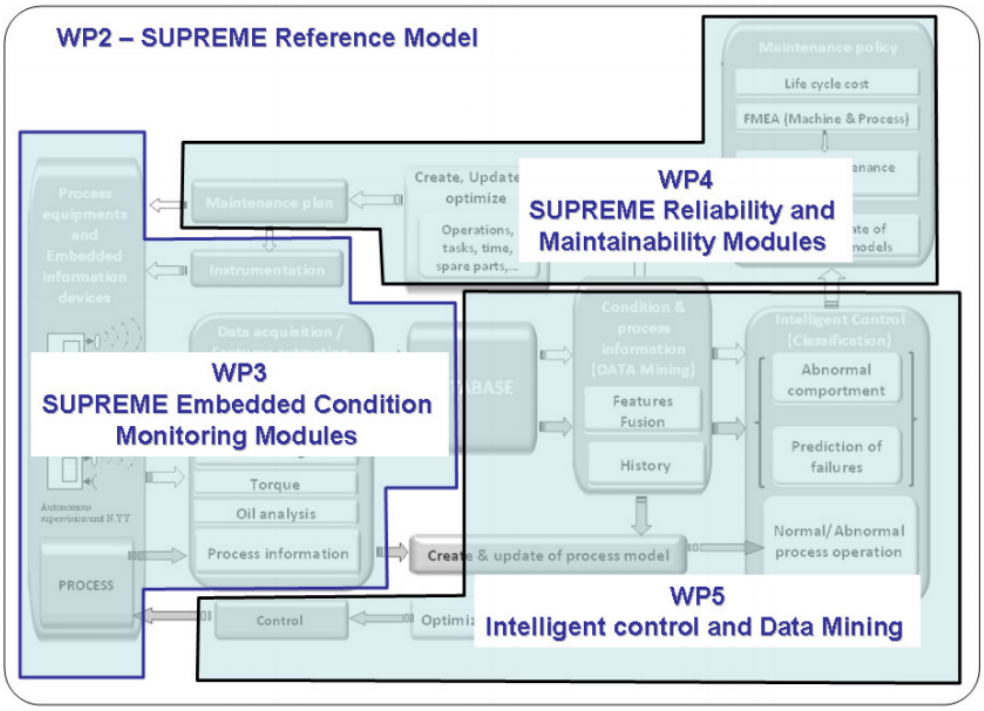


Figure 1.5: SUPREME project concept and related work packages

different technical fields, which is not currently done. The first key objective is to establish a new modular architecture with generic interfaces to link maintainability modules and local embedded condition monitoring modules, both integrating intelligent data capacities. The second objective is to solve specific challenges in each technical module.

The SUPREME project also provides innovative answers to achieve a new generation of predictive maintenance (the sustainable predictive maintenance), knowledge based systems based on smart supervision devices, which can:

- Take into account maintenance at the design stage.
- Monitor the real state/behavior of functioning of the different components or subsystem that are involved in the system and to have embedded signal processing in order to deliver ready to use information.
- Automated configuration procedures, creating much more accurate fit of the system to the monitored machine with greatly reduced human involvement.
- Embedded advanced data analysis algorithms, reducing the data flow only to events and integrated machinery quality indicators.
- Validate all the innovative achievements by presenting the project impact on industrial sector through a Research and Technological Development (RTD) platform.

- Integrate process condition and process state data to provide energy optimization capabilities by means of Intelligent IT supported plant operation under both normal and abnormal situations.
- Automatically learn failure mode patterns and to apply them for further process diagnostics.
- Optimize maintenance planning decisions based upon failure prediction and cost/benefit strategies.

The proposed predictive maintenance system enables to contribute to:

- Higher productivity under varying conditions,
- Smaller losses of operational time due to improved maintenance planning,
- Lower energy consumption (optimal mode of production, avoidance of lost production due to unacceptable quality...),
- Guaranteed production quality (optimal mode of production, detection of malfunctioning...),
- Improve safety of operators (especially by reducing unplanned maintenance interventions, by increasing the safety level of the components and control systems...),
- Protection against risk caused by failures of components (explosion, machine destruction, pollution...).

To demonstrate the potential of the achieved predictive maintenance system, SUPREME implements its achievements on an industrial field, i.e. in a paper manufacturer. The platform is provided by the industrial partners of the project (Orloga and Condat). In this context, the project expected overall achievements are to reduce by 20% the lost time due to machinery failures and process problem, and thus increase the total efficiency of the machine, while reducing energy costs as well as environmental cost. These objectives could be achieved through combined new technological developments in equipment monitoring, intelligent data processing and the dynamic adaptation of the maintenance strategy over the equipment's life-cycle.

1.2.2 Work packages

In order to reach all of its objectives as well as to facilitate the following of the project progress, the SUPREME project is divided into eight smaller parts, called the work packages and abbreviated by WP. Each WP deals with a specific technical or management aspect of the project. The organization of these work packages are represented in Figure 1.6.

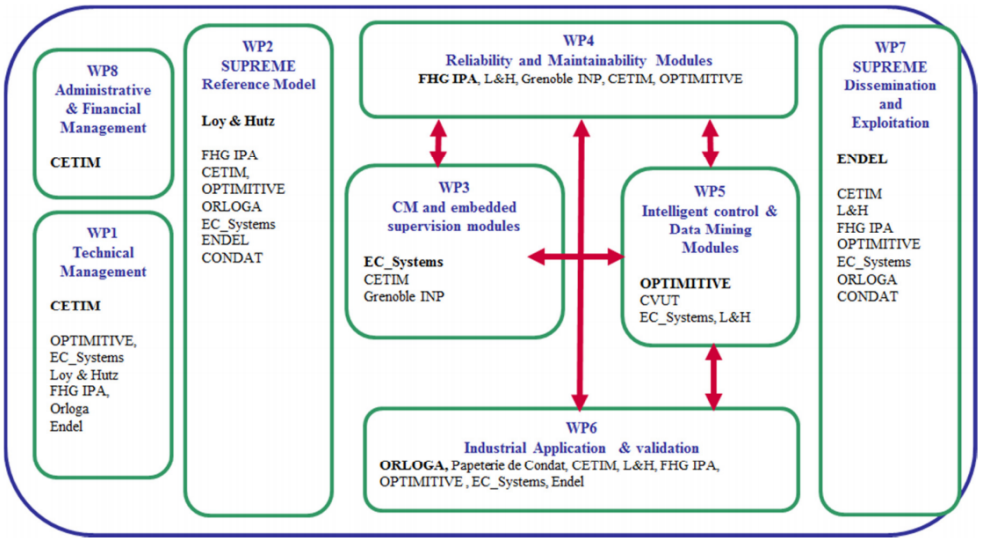


Figure 1.6: SUPREME work plan (work packages)

1.2.2.1 WP1: Technical and scientific management

This WP covers the technical and scientific management of the project, including:

- the organization of the project meetings,
- the scientific coordination and monitoring of other work-packages and tasks.
- the supervision of the project progress
- the scientific review of the work performed by the project partners, i.e. through deliverables.
- research risk and quality management in the project.

1.2.2.2 WP2: SUPREME Reference Model

The WP2 deals with the development of the SUPREME Reference Model, in order to ensure a successful industry-driven implementation of the SUPREME concepts and technologies for a sustainable predictive maintenance. For this end, the following works are carried out:

- Requirements analysis in order to specify the inter-dependencies and interfaces for the SUPREME Modules as a basis for development of the required tools in WP3, WP4, WP5 and their future integration,
- Continuously capturing and updating of the state-of-the-art and beyond the state-of-the-art.

- Conception and realization of the SUPREME Reference Model, by identifying, collecting, specifying and modeling of the SUPREME Modules as logic entities. The SUPREME Reference Model includes phases (e.g. Equipment and component risk assessment, maintenance strategy, etc.), sub-processes the underlying methods and models, the required input and output information and their inter-dependencies.

- Identification and mapping of the required generic tools (e.g. condition monitoring system) to the already existing commercial or in-house developed existing tools/systems in order to perform the functionality of each module and phase of the reference model.

1.2.2.3 WP3: SUPREME Embedded Condition Monitoring Modules

The WP3 is focused on the design and the development of the embedded condition monitoring system (ECMS). It consists of following key components:

- Design and development of ECMS: including the analysis of the embedded condition monitoring requirements, the design of its embedded CM hardware and software.

- Development of advanced signal processing tools that have the capacity for: automatic detection of the characteristics of the signal (peaks, modulations, ...), signalization of the system failures and more robust fault detection.

- Development of multi-sensor approach

- Embedded CM prototype realization and test: including Laboratory verification, on site test and implementation of custom algorithms

1.2.2.4 WP4: SUPREME Reliability and Maintainability Module

The overall objective of this WP is to develop the SUPREME Reliability and Maintainability Module aiming at the implementation of a new life cycle orientated condition and risk-based maintenance planning. The tasks included within this WP are: risk-based identification of critical components, condition and risk based maintenance planning, remaining useful life prediction, dynamic adaptation of maintenance intervals and tasks. Compared to the predictive maintenance implementation in Figure 1.3, this WP4 corresponds to the two last stages: prognostics and decision making. The main works of this thesis are involved in the development of the reliability sub-module. For a clarification of the problem statement and the works to be done within this thesis, the WP4 will be presented in more detailed in Section 1.2.3.

1.2.2.5 WP5: Intelligent Control and Data Mining Modules

The objective of this WP is to perform the intelligent processing of real-time data coming from sensors, including data fusion, feature extraction and matching, failure prediction and operation optimization targeted towards energy saving, both in normal and abnormal working conditions. Specifically, data mining techniques are performed on failure modes starting from historical data. Then knowledge will be generated from this data mining task, which could be exploited for failure prediction. Moreover, such knowledge is also used for energy optimization by means of enhanced operation of the plant. In this context, hybrid approaches could be used to perform data fusion as required for pre-processing, and adequate process identification and optimization are carried out according to nature and complexity of the data.

1.2.2.6 WP6: Industrial Application and Validation

The objectives of the WP6 are to:

- Test and validate the SUPREME modules developed,
- Integrate and interface the SUPREME modules, to develop the SUPREME predictive maintenance tool and validate it on the demonstrator case, a paper machine.
- Compare the results with other maintenance approaches in order to quantify the reduction of operational loss time and the decrease of energy consumption.

1.2.2.7 WP7: Continuous Dissemination and Exploitation

WP7 Dissemination of the SUPREME results towards industry and particularly small and medium enterprises (SMEs) is an essential activity of the project. SMEs are the main target of the dissemination task. Two ways are identified to spread new knowledge of SUPREME results, including: i) direct exploitation, mainly by the consortium members including SMEs and ii) special tasks of training towards SMEs by e-learning modules.

1.2.2.8 WP8: Administrative and financial management

This work package covers the day to day administrative management of the project as well as administrative aspects of project progress and deliverable reporting. The overall project management is operated by the Project Management Board (PMB). Its main tasks are to follow the project's progress and tackle unexpected issues.

1.2.3 Reliability and Maintainability module

As mentioned above, the overall objective of the WP4 - Reliability and Maintainability (R&M) module is to improve the maintenance activities aiming at the implementation of a new life cycle orientated condition and risk-based maintenance planning. To achieve this objective, the R&M module is divided into four sub-modules:

- high level data management providing the relevant information of the following methods and tools,
- risk assessment and management of components relevant for value stream,
- deterioration models to predict the residual useful lifetime for high risk rated components and
- dynamic adaption of the maintenance strategy and plan over the time.

The structure and the relationship between these sub-modules is shown in Figure 1.7 .

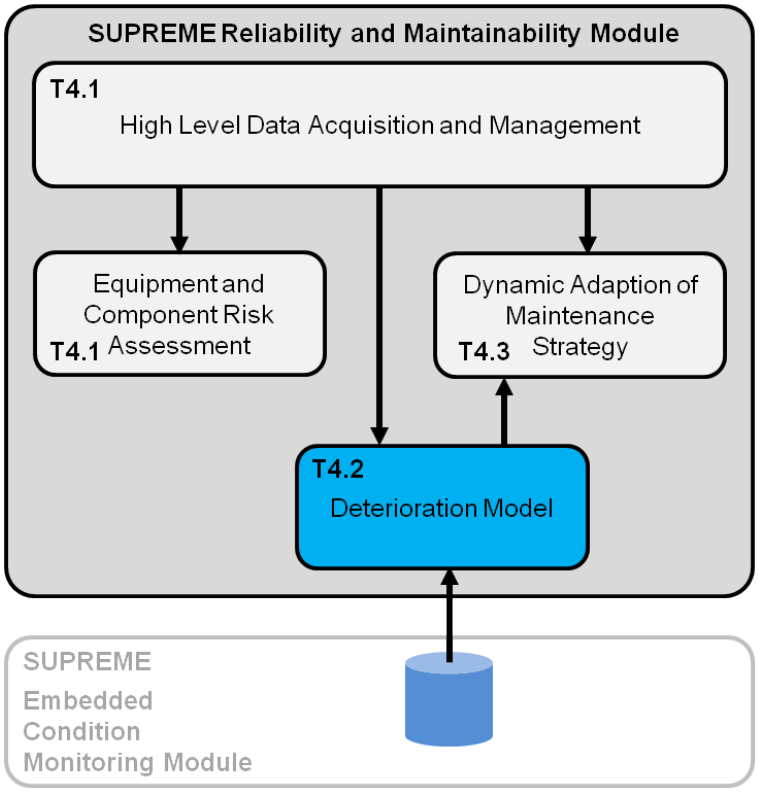


Figure 1.7: WP4 sub-modules structure

1.2.3.1 High level data acquisition and management sub-module

This task concerns the development of a tool for high level data management in order to manage production system and maintenance system data for the WP4. The main purpose is to manage the high level features, production and maintenance related data of the R&M module, i.e. production master data, production orders, maintenance master data or maintenance orders. This sub-module provides the necessary data for the others sub-modules of the WP4.

1.2.3.2 Equipment and component risk assessment sub-module

This sub-module is dedicated for the enhancement of a method and tool for systematic identification of critical components of an equipment relevant for the value chain. In order to increase the availability of production systems, one key aspect is to identify critical machines and components and to define corresponding failure risk measures. A holistic risk management can be constructed for addressing this problem. Such a risk management system is carried out and implemented by the two partners: IPA-FhG and L&H.

1.2.3.3 Deterioration model sub-module

In order to make the transition from classical preventive maintenance policies (i.e. time-based or age-based) to more dynamic predictive maintenance policies, it is necessary to make the transition from classical reliability modeling to deterioration-based reliability modeling and prediction of residual useful life. Within this sub-module, the main developments are related to deterioration modeling, damage prognostics and residual life prediction of the identified critical component (c.f. Figure 1.8).

This sub-module can be considered as an implementation of the prognostics stage within a predictive maintenance program (c.f. Figure 1.3). The works of this thesis are found within this sub-module. The output of this sub-module is then the basis for the dynamic adaptation of maintenance strategy sub-module.

1.2.3.4 Dynamic adaptation of maintenance strategy sub-module

Once deterioration models are constructed and the prediction of the RUL are made, the next and crucial stage in implementing the predictive maintenance is to make the maintenance decision. This sub-module is dedicated to the dynamical adaptation and improvement of maintenance strategies and plans, considering the current situation within the production.

By taking into account the current maintenance work orders, production orders, person

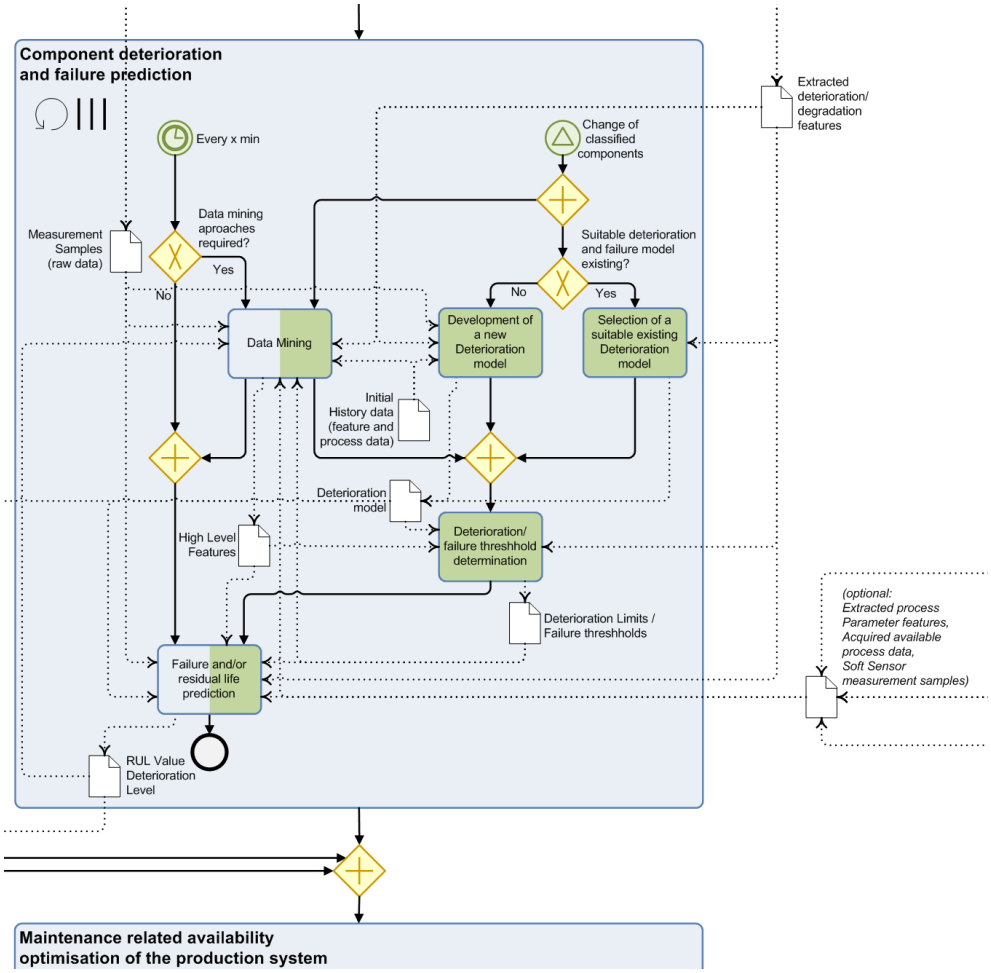


Figure 1.8: Reference model of the deterioration sub-module [53]

capacities, component risk and stop costs, this sub-module expects to achieve a minimized costs for the monitored components by minimizing unplanned downtime of a machine.

1.3 Problems statement and contributions of the thesis

Involved in the deterioration sub-module of the WP4, the main works of the thesis concentrate on the development of deterioration models and the residual life prediction methods with the applicability for the potential SUPREME application cases. This damage prognostics and the residual life prediction process consist of two main steps, each of them corresponding to specific technical difficulties:

- Assessing the current deterioration level of the item, by analyzing and processing the condition monitoring information gathered online (vibration, current,...) and developing a deterioration model for the item: this model should be able to describe the temporal evolution of the item wear and will be the basis for the prediction

of the future evolution of the system, given its present actual (or estimated) state. Stochastic processes (with covariates) can be used as efficient tools in order to capture the item-to-item and temporal variability.

- Developing the residual life prediction based on the deterioration model, integrating the online information available on the item both on its state and on its environment and operational conditions.

In this thesis, a complete review of the existing deterioration models and the associated residual life estimation methods is firstly conducted and presented in Chapter 2. This step is essential for the development of new deterioration models for the SUPREME project. Indeed, it helps to have a comprehensive overview of the models and methods that have been used in the literature as well as to analyze their advantages and their limitations. With the development of the condition monitoring system within the WP3, a large amount of CM data could be acquired within the SUPREME project. The literature review concentrates hence on the data-driven approaches. Moreover, as the deterioration processes are often stochastic in practice and the prediction is always uncertain, the uncertainties management should also be taken into consideration. An efficient way to do this is to model the deterioration phenomena by stochastic processes as well as characterize the residual life by probabilistic distributions. A special attention is therefore paid to the stochastic deterioration models in this first study.

Based on the literature review results, the main contributions of this thesis can be summarized in the two following points:

- Adaptation of the state-of-the-art of stochastic deterioration models and associated RUL estimation methods to the SUPREME application case (Chapter 3).
- Advancement of the state-of-the-art: development of several advanced deterioration models and associated RUL prediction methods to overcome the identified limitations and to have ready-to-use deterioration models when the required data is available (Chapters 4, 5 and 6).

Specifically, to simulate the real deterioration processes, a test-bench was constructed in the framework of the SUPREME project and is presented in Chapter 3. The health indicators are extracted from vibration signals acquired from accelerometers located on the housing of the main bearing of the test-bench. Investigating the temporal evolution of this indicator, two different stages are realized: During the first stage, the deterioration indicators are small and stay relatively constant while they begin to propagate quickly in the second one. A so-called incubation/propagation is hence adapted and the RUL estimation procedure is also presented. In addition, in the SUPREME framework, the output of the dynamic adaptation of maintenance strategy sub-module is to modify the production orders of the machine in order to have a best adaptation of the load applied on the deteriorated component so that its lifetime could be extended, at least it could

operate until the next planned intervention action. To model the impacts of different loads on the deterioration processes, a load-dependent deterioration modeling study is implemented. Furthermore, some practical implementation issues are also addressed in this chapter.

Concerning the second point, we are interested in this thesis the co-existing problem of several deterioration modes, even within a component. Indeed, such deterioration phenomenon has not been carefully taken into consideration in the literature. To tackle this problem, the concept of the multi-branch model is proposed and developed in the second part of this thesis. Depending on the nature (discrete or continuous) of the health state of the component, two types of multi-branch models are introduced. For the discrete-state case, the multi-branch Hidden Markov Model is presented in Chapter 4 and the multi-branch Hidden semi-Markov Model in Chapter 5. Chapter 6 is dedicated to represent the multi-branch model for the continuous-state case. For each model, a two-phase framework is implemented for both the diagnostics and prognostics purposes. Through numerical as well as case studies, we show that the multi-branch models, by taking into account the co-existence of multiple deterioration modes, can give better performances in RUL estimation compared to the ones obtained by standard "mono-branch" models.

1.3.1 List of publications

The main contributions of this thesis are the principal subject of the following publications:

1.3.1.1 International journals

[1] **T.T. Le**, F. Chatelain, C. Bérenguer, "Multi-branch Hidden Markov Models for remaining useful life estimation of systems under multiple deterioration modes". Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, (1st revision under review) 2015

1.3.1.2 International conference papers with proceedings

[1] **T.T. Le**, F. Chatelain, C. Bérenguer, Hidden Markov models for diagnostics and prognostics of systems under multiple deterioration modes, in Safety and Reliability: Methodology and Applications - Proc. of the European Safety and Reliability Conference - ESREL 2014 - 14-18 September, 2014 - Wroclaw, Poland - T. Nowakowski, M. Mlynczak, A. Jodejko-Pietruczuk & S. Werbinska-Wojciechowska (eds) - London : Taylor & Francis (CRC Press/Balkema), 2015 - ISBN : 978-113802681-0, pages 1197-1204 [77]

[2] **T.T. Le**, C. Bérenguer, F. Chatelain, Multi-Branch Hidden Semi-Markov Modeling

for RUL Prognosis, in Proc. Annual Reliability and Maintainability Symposium - RAMS 2015 - January 26-29, 2015 - Orlando, FL, USA - Los Alamitos, CA, USA : IEEE - ISBN : 978-1-4799-6702-5 [CD-ROM] - ISSN : 0149-144X, pages 122-128, paper#03D1 - doi:10.1109/RAMS.2015.7105132 [75]

[3] **T.T. Le**, C. Bérenguer, F. Chatelain, Prognosis based on Multi-branch Hidden semi- Markov Models : A case study, in Proc. 9th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes - SAFEPROCESS 2015 - September 2-4, 2015. Arts et Métiers ParisTech, Paris, France - IFAC, 2015, pages 91-96 [76]

[4] **T.T. Le**, F. Chatelain, C. Bérenguer, Jump Markov Linear Systems for deterioration modeling and Remaining Useful Life estimation, in Safety and Reliability of Complex Engineered Systems - Proc. of the 25th European Safety and Reliability Conference - ESREL 2015 - 7-10 September, 2015 - Zurïch, Switzerland - L. Podofillini, B. Sudret, B. Stoja- dinovic, E. Zio & W. Kröger (eds) - London : Taylor & Francis (CRC Press/Balkema), 2015 - ISBN : 978-1-138-02879-1, pages 1287-1295 [78]

1.3.1.3 European project deliverable

[1] R. Haug, M. Eltabach, J. Christien, C. Johann, **T.T. Le**, C. Bérenguer, P. Granjon, T. Atienza, D. Lucke, T. Adolf, J. Klema, R. Cernoch. Deliverable 2.12: State of the Art and Beyond the State of the Art (WP2 - FP7 Project SUPREME - SUstainable PREdictive Maintenance for manufacturing Equipment - Grant Agreement 314311), 09/2013, 66 pages [57]

[2] D. Lucke, T. Adolf, R. Haug, **T.T. Le**, C. Bérenguer, F. Chatelain, J. Christien. Deliverable 4.1 Enhancement of Existing Competencies, Methodologies and Tools (WP4 - SUPREME Reliability and Maintainability Module - FP7 Project SUPREME – Sustainable PREdictive Maintenance for Manufacturing Equipment - Grant Agreement 314311), 07/2014, 51 pages [54]

[3] D. Lucke, T. Adolf, R. Haug, **T.T. Le**, C. Bérenguer, F. Chatelain, J. Christien. Deliverable 4.2 Development of the new Required Competencies, Methodologies and Tools (WP4 - SUPREME Reliability and Maintainability Module - FP7 project SUPREME - SUstainable PREdictive Maintenance for manufacturing Equipment - Grant Agreement 314311), 07/2014, 42 pages [55]

[4] D. Lucke, T. Adolf, R. Haug, **T.T. Le**, C. Bérenguer, F. Chatelain, J. Christien. Deliverable 4.3 Realisation of SUPREME Reliability and Maintainability Modules in LivingLabs (WP4 - SUPREME Reliability and Maintainability Module - FP7 project SUPREME - SUstainable PREdictive Maintenance for manufacturing Equipment - Grant Agreement 314311), 01/2015, 65 pages [56]

Part I

Deterioration modeling based on condition monitoring data and remaining useful life estimation: background, state of the art and adaptation to the SUPREME project

Introduction of the first part

This part is dedicated to present the research work conducted during the first stage of our participation to the SUPREME project. Involved in the deterioration models sub-module of the WP4, the first essential task is to carry out a literature review about the existing deterioration models as well as the remaining useful life estimation methods. Such a review is indispensable and is the basis for the next stage of the thesis' works.

This part contains two chapters: Chapter 2 and Chapter 3. In Chapter 2, we firstly introduce some background about the deterioration modeling as well as the RUL estimation with the requirement of uncertainties management. The existing approaches and their inherent advantages and limitations are then analyzed. With the development of an embedded condition monitoring system within the SUPREME project, several types of CM data could be made available. The data-driven prognostics approaches are hence selected for a more detailed study in this chapter.

Based on the presented literature study, Chapter 3 is dedicated to the investigation of an adaptation of the state-of-the-art deterioration models and RUL estimation method to a SUPREME application case. Specifically, the so-called incubation/propagation model is applied for the data acquired from a test-bench that is constructed in the framework of the SUPREME project for modeling the deterioration processes in practice. The numerical results show that such model can well represent the dynamic in evolution of the real deterioration processes. The RUL estimation method is also addressed for this two-phase deterioration model.

Chapter 3 also addresses the impacts of environmental factors on the deterioration dynamics. Indeed, in the SUPREME framework, the estimated RUL is provided to the dynamic adaptation of maintenance strategy sub-module whose output is to modify the production orders of the machine, i.e. the load applied on the deteriorated component so that its lifetime could be extended. Such the load changes will affect back to the deterioration evolution. To model such impacts, a load-dependent deterioration study is presented. Furthermore, some aspects of practical implementation are also addressed. For example, the problem of information exchange between the project partners is presented in this chapter.

Chapter 2

Deterioration modeling and remaining useful life estimation: state-of-the-art

2.1 Preliminary

Prognostics plays an important role in implementing a predictive maintenance program. It is this stage that distinguishes between the predictive maintenance and the condition-based maintenance. Indeed, prognostics helps to predict accurately and precisely the remaining useful life (RUL) of a failing component or system. This estimation is an essential source of information so that maintenance personnel can schedule and optimize maintenance actions in order to reduce maintenance costs and avoid unscheduled downtime.

There exists several definitions of prognostics in the literature. For example, Heng et al. defined the prognostics as the forecast of an asset's remaining operational life, future condition, or risk to completion [59]. Engel considered the prognostics as "the capability to provide early detecting of the precursor and/or incipient fault condition of a component and to have the technology and means to manage and predict the progression of this fault condition to component failure" [35]. In the industrial and manufacturing areas, prognostics is interpreted as the methodology to predict the remaining useful lifetime (RUL) of a machine or a component once an impending failure condition is detected, isolated, and identified [144]. The RUL here is the time left for a system to accomplish its required function, i.e. before a completed failure, given the past, the current and/or the future operation profiles of the system.

From the above definitions, one can realize that estimating the RUL plays the center role in prognostics implementation. Since the prediction is always uncertain, uncertainties management becomes an important issue in prognostics [144]. There are several sources of uncertainties that can affect the RUL estimation results, such as modeling uncertainties, measurement noises or operating environment uncertainties [141]. Characterizing the RUL by a probabilistic distribution is an effective way to take into account such uncertainties. For example, in [67], Jardine *et al.* define the RUL as a conditional random variable:

$$RUL = T_f - t_0 \mid T_f > t_0, Z(t_0)$$

where T_f is a random variable representing the time to failure, t_0 is the current age and

$Z(t_0)$ is the past condition profile (including environment data) up to the current time and/or existing knowledge on the future use of the system.

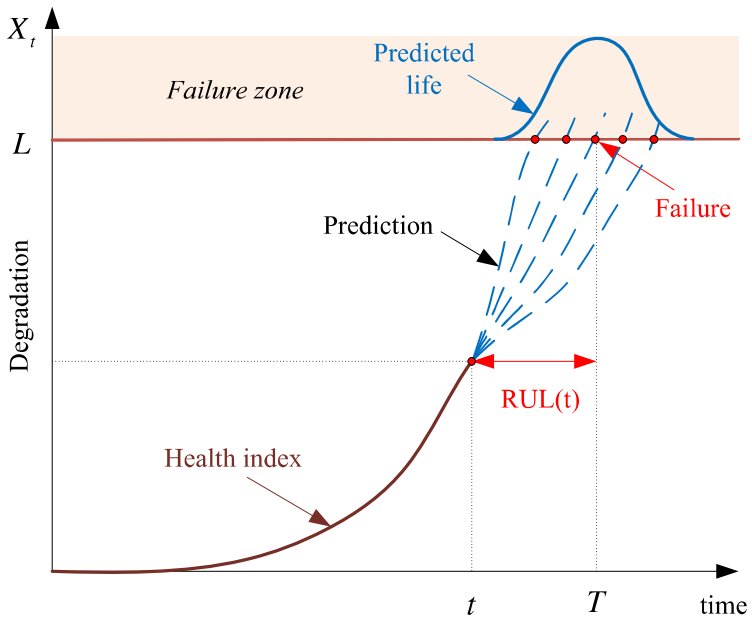


Figure 2.1: Deterioration and RUL estimation illustration

Figure 2.1 clarifies the above definition of the RUL and gives an example about how the RUL distribution can be estimated. In this figure, t indicates the current age of a component. Based on the features and the information obtained from the diagnostics stage, an indicator representing the health state of component from the beginning of life until time t is constructed. To estimate the RUL, the failure need to be defined. In the literature, a widely used assumption about the failure is based on the "first hitting time" principle: the system is supposed to be failed once its health indicator reaches for the first time a predetermined threshold [83]. By propagating the temporal evolution in the future of the health index in this example until it reaches a threshold L, the RUL distribution is obtained.

It should be noted that the RUL of a component shares the same probability law as its conditional time to failure; that is survival function of the RUL coincides with the conditional reliability, with only a time translation difference:

$$\mathbb{P}\left(RUL > u\right) = \mathbb{P}\left(T_f > t + u \mid T_f > t, Z\left(t\right)\right)$$

This leads the estimation of the RUL of a component being equivalent to the estimation of its conditional reliability. In the case of deterioration-based reliability, the conditional part integrates all the knowledge on the deterioration level delivered by the monitoring information.

In the next section, we firstly review the state-of-the-art of the prognostics approaches that exist in the literature. Then in Section 2.3.2 and Section 2.3.1, the deterioration

models as well as the associated RUL estimation methods are studied in detail. The conclusion is given in Section 2.4.

2.2 Prognostics approaches classification

Depending on the type of used data and techniques, the existing prognostics approaches can be classified into three main categories [17, 144]:

- Experience-based prognostics
- Evolutionary or trending models-based
- Physical model-based prognostics

The range of prognostics approaches as a function of the applicability to various systems and their relative implementation cost are summarized in Figure 2.2.

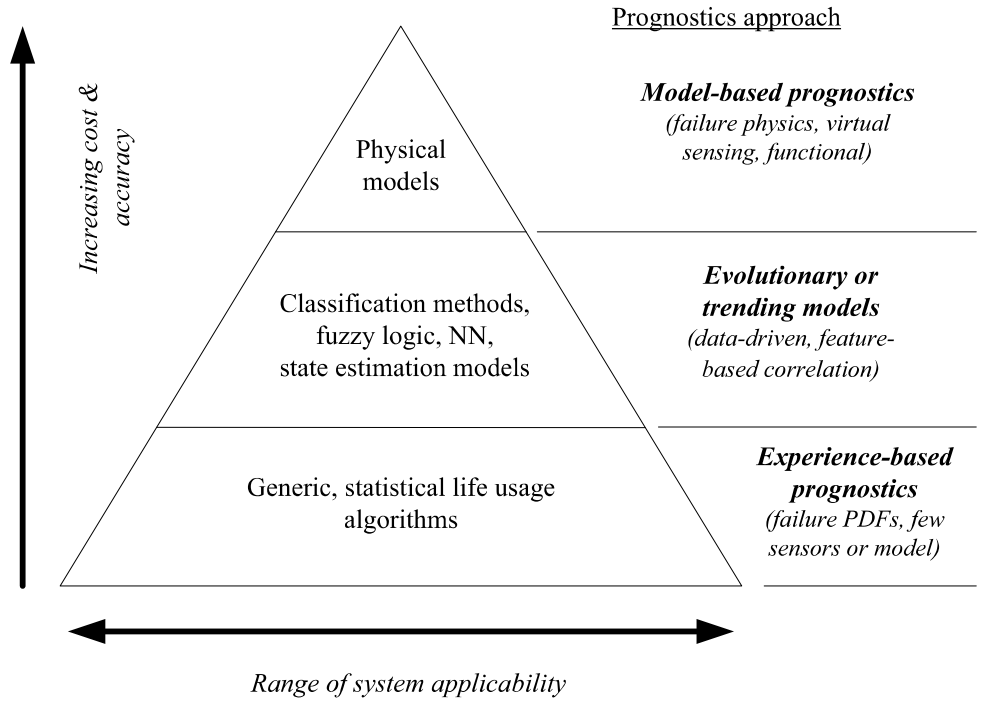


Figure 2.2: Classification of prognostics approaches [17]

2.2.1 Experience-based prognostics

This approach is the traditional reliability estimation method, which can be called as reliability-based modeling approach in the literature. It involves analyzing the event

data (e.g. time to failure data) of a population of identical units and determining a probability density function and related hazard function for the population. This density function is not representative of a single fault progression from incipience to final failure, it just provides information about when failures are typically expected to occur. Various parametric failure models have been used to model failure data, such as Exponential, Lognormal, Gaussian or Weibull distribution. The most commonly used among them is the Weibull distribution because of its ability to describe many types of behavior, including infant mortality in the "bath-tub" curve [137].

The advantage of the experience-based prognostics approaches is that the time-to-failure distributions are derived from observed statistical data specifically relevant to the equipment of interest. Such data could be easily extracted from a company's existing computerized maintenance management systems [137]. In other words, it does not require the condition monitoring data hence the installation of sensors is not necessary. In addition, this approach is relatively straightforward for being analyzed by reliability engineers. This property gives it a wide range of system applicability.

Despite its merits, experience-based prognostics approaches possess some limitations. Firstly, they may have little value in making maintenance's decisions since they only provide general overall estimate for the entire population of identical components. Obviously, one would be more interested in the ongoing reliability information of a particular component that is currently running in a machine, rather than in the mean time to failure or mean residual life (MRL) of the whole population [59]. Lastly, estimating the parameters of the lifetime distribution requires sufficient historical failure data, which may be difficult to achieved in real-life situation, such as in the case of the paper machine as found in the SUPREME project.

2.2.2 Data-driven approaches

As indicated by its name, data-driven approaches assess the health state and estimate the RUL of a machine basing on the available data. According to Jardine *et al.*, the data used for prognostics can be categorized into two main types: event data and condition monitoring (CM) data [67]. The event data include the information on what happened and/or what was done (i.e. installation, breakdown, minor repair, etc.) while the condition monitoring data are the measurements related to the health condition/state of the physical system of interest.

The data-driven approaches can be further divided into two main classes: Evolutionary or trending-based and Artificial intelligence-based.

2.2.2.1 Evolutionary or trending prognostics

The fundamental principle of the evolutionary or trending-based approaches is that the health of monitored equipment is represented by some key CM features called the health indicators. The RUL is then estimated by propagating these indicators toward a predefined failure threshold to predict the "first hitting time" [83]. For this purpose, several regression techniques were used in the literature, such as Auto-Regressive (AR) models [162, 160], support vector regression approaches [124, 18], general path model [95], Lévy process [145, 30, 158], etc. Among them, the two most popular ones in the literature are the general path model and the Lévy process.

2.2.2.2 Artificial learning techniques

Artificial Neural Network (ANN) is currently the most commonly used data-driven technique in the prognostics literature. The idea of using ANN for RUL estimation is that the ANN computes an estimated output for the remaining useful life of a component, directly or indirectly, from a mathematical representation of the component derived from observation data rather than a physical understanding of the failure processes. This method is usually called the black box method in the literature.

A typical ANN consists of a layer of input nodes, one or more layers of hidden nodes, one layer of output nodes and connecting weights as illustrated in Figure 2.3. It can use a number of different types of data as the inputs including process variables, condition monitoring indicators, asset characteristics (e.g. age, operating hours) and maintenance history features (e.g. time since last overhaul). The RUL of a system can be obtained at the output.

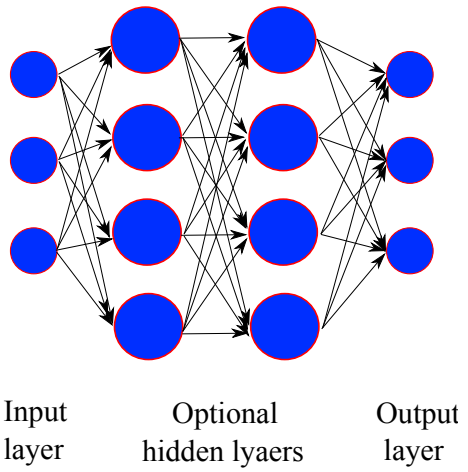


Figure 2.3: An neural network example

The principle of a neural network is that it learns the unknown function by adjusting its weights with repetitive observations of inputs and outputs. An activation function is

associated with each node that defines if and how information is transmitted to subsequent nodes. Calculated values of each node's activation function are then used as inputs to any subsequent nodes. The same activation function is generally used for all nodes in a layer; however, other layers in the same network may use different functions. As processing (computing the activation function) can be performed by the nodes in parallel, neural networks are computationally very efficient. The global behavior of a particular network is determined by its architecture (nodal arrangement), synaptic weights and parameters of the nodal activation function. The behavior of one node in the network is shown in Figure 2.4.

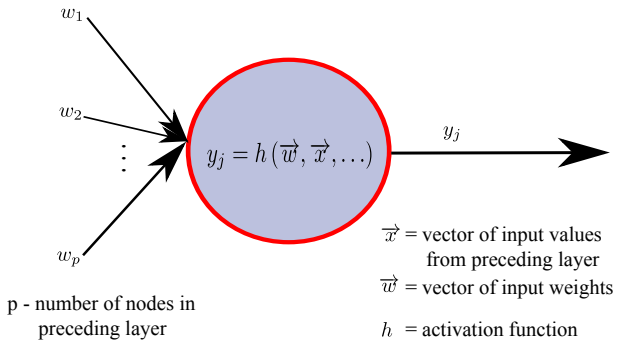


Figure 2.4: Behavior of a single ANN node

Neural network architectures used for RUL estimation can be classified as feed-forward (static) neural networks (FFNN as shown in Figure 2.3) or dynamic networks. In the first architecture, the inputs to a particular layer depend only on the outputs of the preceding layer while in the second one; the inputs to a particular layer are dependent on the outputs of the preceding nodes as well as preceding iterations of the network itself. Most prognostics RUL models developed to date have been based on static networks, in which the popular architectures are Multilayer Perceptron (MLP), Radial Basis Function (RBF) network and General Regression Neural Network (GRNN). Among these, multilayer perceptron is the most popular architecture [137].

The traditional MLP approach was used by Huang et al. in [62]. In their paper, they dealt with a new scheme for the prediction of a ball bearing's RUL based on Self-Organizing Map (SOM) and Back Propagation (BP) neural network methods. Similarly, Gebraeel et al. have applied the MLP networks for the data acquired from accelerated bearings test [42]. They also proposed another model that grouped the bearings into several clusters and then used a GRNN to compute a single regression function for the degradation signals of the bearings from the same clusters.

Traditional neural networks are good at mapping non-linear, numerical information. Unfortunately, much of the practical data for describing failures is linguistic, ambiguous and/or approximate. Thus, a number of researchers have integrated fuzzy logic into MLP networks to be able to capture this additional information. For example, in [156], Wang et al. used a neural-fuzzy (NF) network to predict spur gear condition value one step ahead. The fuzzy interference structure is determined by experts, whereas the fuzzy membership

functions are trained by the neural network. Bonissone and Goebel proposed a Hybrid Soft Computing Model based on Adaptive Network-Based Fuzzy Inference System (ANFIS) to predict the time to break of papers in the wet-end section of paper machine [14]. Yam et al. used a Recurrent Neural Network (RNN) to perform one-step ahead prediction of CM data from a planetary gear train [161]. Wang and Vachtsevanos in [153] developed a Recurrent Wavelet Neural Network (RWNN) to predict rolling element bearing crack propagation.

The advantages of using ANN for RUL estimation are:

- It can model complex, multidimensional, non-linear relationship between the RUL at the output and the CM data at the inputs.
- It can use many different types of input data such as sensor readings, fuzzy inferences, user inputs, etc.
- There are many user-friendly software packages available for developing neural network because most commercial prognostics modeling system uses ANN for RUL prediction.

Despite of their merits, the biggest problem of applying ANN for RUL estimation is that they do not naturally provide confidence limits for their prediction. Other disadvantages of using this approach are that they require a large data set for training the model and that it is difficult to interpret the obtained result (no physical meaning). Without physical meaning and confidence of the estimation, it seems to be very difficult to schedule an effective maintenance plan.

2.2.3 Physical model-based approaches

Physical model based prognostics techniques use analytic and/or physics-based models of the degradation phenomenon to predict the dynamics and degradation in system behavior. These approaches apply to situations where accurate mathematical models of system behavior can be constructed from first principles [17]. Model-based methods assume that the system is represented as a physical-based model.

A common physics-based approach in the literature is crack growth modeling which can be fallen into two major categories: deterministic and stochastic models. Fatigue crack growth in machinery components such as bearings, gears, shafts, and aircraft wings is affected by a variety of factors, including stress states, material properties, temperature, lubrication, and other environmental effects. To date, almost the available empirical and deterministic fatigue crack propagation models are based on Paris' formula [144]. Li et al [93, 91, 92] modeled rolling element bearing defect growth using a variation of Paris' law:

$$\dot{D} = \frac{dD}{dt} = C_0 \left(D\right)^n$$

which states that the rate of defect growth is related to the instantaneous defect area D under a constant operating condition. The parameters C_0 and n are material constants that need to be determined experimentally. Since these parameters often vary with factors other than the instantaneous defect size, an adaptive scheme was developed to predict the defect propagation rate and defect area size by Li et al. [91, 92].

With the above model, the defect area D can be deterministically decided at any time given the parameters C_0, n and an initial defect size. However, the defect propagation is stochastic in nature. It means that identical bearings under the same operating conditions can have significantly different crack propagation paths. Taking this problem into account, Li et al. [93] developed a stochastic bearing fatigue defect-propagation model by introducing a log-normal random variable, $e^{Z(t)}$ to the deterministic model:

$$\dot{D} = \frac{dD}{dt} = C_0 \left(D\right)^n e^{Z(t)}$$

The log-normal random variable is used to characterize the amount of uncertainty in material properties or environmental factors. Since the resulting growth equation is a stochastic differential equation, we can use the Monte Carlo simulation of probabilistic neural network to estimate its parameters distribution [144].

Recently, the crack growth state model have been integrated with the particle filter by Orchard and Vachtsevanos to estimate the crack length on the plate of a UH-60 planetary gearbox [116]. They have compared their methodology with an EKF-based approach for long-term prediction. The particle filter-based approach was demonstrated to be suitable for on-line fault diagnostics and failure prognostics.

Another physical models that were used in the literature for bearing prognostics is the fatigue spall initiation and progression model and stiffness-based prognostics model. In [117], Orsagh et al. used a stochastic version of the Yu-Harris bearing life equation [168] to predict spall initiation and the Kotzalas-Harris progression model to estimate the time to failure. Qiu et al. [127] constructed a vibration model that based on the sitffness and damping coefficient. By using that model, the authors established the equations relating the bearing lifetime with the natural frequency of vibration and its amplitude. The recursive least square (RLS) parameter estimation algorithm has been used for on-line application purpose.

The main advantage of using physical models is the ability to directly incorporate the existing understanding of the physical mechanisms of failure. Once such model is available and accurate, it can give the better performance in RUL estimation than other methods [137]. Additionally, the changes in model outputs as described by the residuals tend to have a direct (or easily translatable) physical meaning. Understanding the significance of model outputs is hence straightforward and easy to action.

However, such accurate physical model is usually difficult to establish for complex systems in reality because it requires a deep understanding of how the failure mechanisms

behave under the range of relevant operating conditions. Another disadvantage of this approach is that it is only suitable for the use in some particular cases. In the complex working environment as in a paper machine, it seems to be impossible to establish accurate physical models.

2.3 Stochastic deterioration models and RUL estimation methods

A deterioration model, as indicated by its name, represents the dynamic time evolution of deterioration processes. It can also be used to link between the observed symptoms (e.g. measures) and the actual health states of a system. The purpose of modeling deterioration processes is to i) help the maintenance persons to assess actual machine deterioration level given the monitoring information, ii) predict or extrapolate the future evolution of machine health and iii) estimate residual useful life of the machine.

Figure 2.5 shows a classification of deterioration models existing in the literature. According to the deterioration nature of item, the deterioration processes can be classified as discrete or continuous processes. The former assumes that the deterioration state of a system changes directly from one level to another whilst the latter assumes that system health varies continuously in time domain.

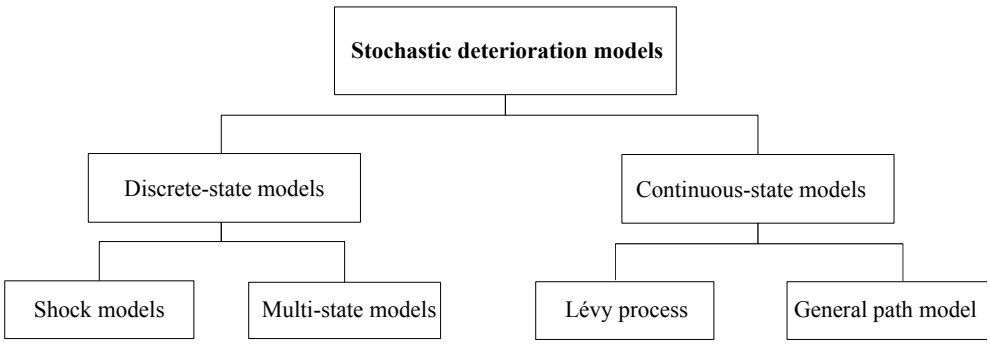


Figure 2.5: Deterioration models classification

2.3.1 Discrete-state deterioration models

2.3.1.1 Shock models

Shock models have been studied by several authors and provide a realistic formulation for modeling certain reliability systems situated in random environment. These models are usually physically motivated. For instance, the extreme and cumulative shock models may be appropriate descriptions for the fracture of brittle materials, such as glass, and for the damage due to the earthquakes or volcanic activity, respectively [99].

Shock models describe systems subject to shocks that occur at random time with random magnitude. There are several types of shock models existing in the literature, such as cumulative shock models, extreme shock models [109]. The former supposes that systems break down when the cumulative shock magnitude exceeds a given threshold while the latter is based on the assumption that one single large shock exceeding a given threshold can lead to systems failure. For the purpose of modeling deterioration processes, in this section, we consider only the cumulative shock model. An illustration of this model is depicted in Figure 2.6.

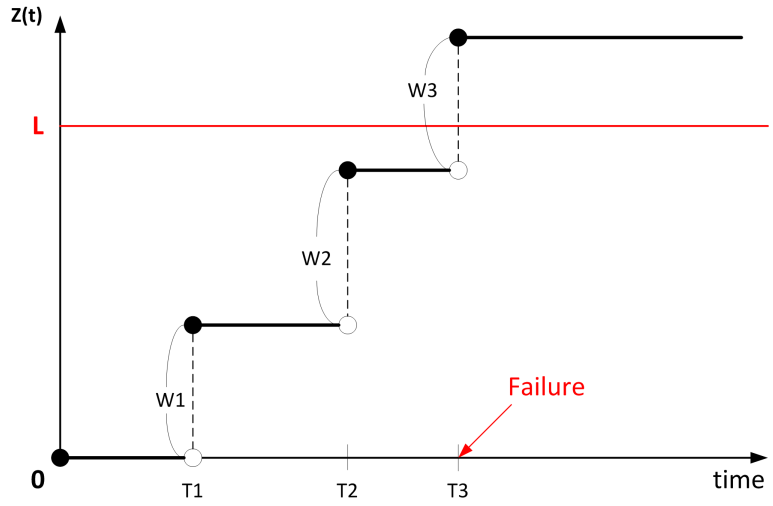


Figure 2.6: Illustration of cumulative shock models

In this standard cumulative model, the paths of the degradation process $Z(t)$ are monotone increasing step functions. It is commonly assumed that arrivals of shocks on the item follow a non-homogeneous Poisson process and random damage is accumulated to the previous damage level of the item which remains unchanged between shocks [110]. The item is said to be failed when the total amount of damage exceeds a failure threshold L which is usually assumed to be constant and pre-determined in the literature.

Let W_i, $(i = 1, 2, \ldots)$ be random variables representing the damage due to i-th shock, $N(t)$ be the total number of shocks received up to time t, $t > 0$, the total damage at time t can be determined by:

$$Z(t) = \sum_{i=1}^{N(t)} W_i \quad (N(t) = 1, 2, \ldots)$$

From this formula and based on the renewal theory, several reliability quantities such as the reliability function, the mean time to failure or the hazard rate, etc. can be calculated [109]. Several stochastic process have been proposed and investigated in the literature in order to represent the occurrence of the damage. We can name here some typical processes, such as the birth process [126], the counting process [122], etc.

2.3.1.2 Multi-state models

Multi-state models simplify a system behavior into a stochastic process with a finite number of state values. Markov chain and semi-Markov chain are the two most popular processes to model this type of deterioration. Markov chain assumes that the system can, at any instant, be in only one state. As the time passes, the system will either remain in the same state or jump to another one with some probabilities. An example of using Markov chain for modeling the deterioration of a bearing is illustrated in Figure 2.7. In this example, bearing health conditions are described by five discrete states: normal, nick, scratches, more nicks and failure. Since the bearing's degradation process is usually irreversible in practice, there is no probability of returning to the previous state from current state. The bearing is said to be failed when its health state reaches for the first time the final (failure) state.

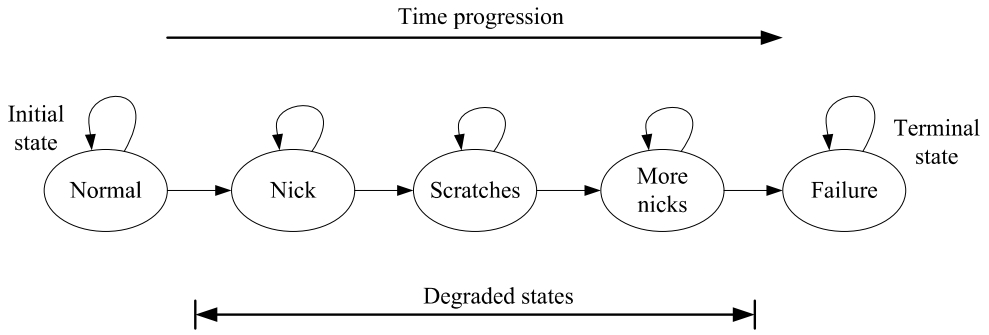


Figure 2.7: A Markov chain representing bearing deterioration process

The principle of RUL estimation using Markov-based models is to compute the first passage time (FPT) which is the amount of time the process will take to transit from the current state to the absorbing state (e.g. Failure state) for the first time. This principle is easy to understand and closer to what is used in industry, therefore it leads the Markov-based models to have been widely applied for RUL estimation and to maintenance decision making support. However, there are two limitations which limit the capacity of applying this model to bearing prognostics framework. Firstly, it is assumed that the system health state sojourn time is exponentially distributed to have the constant rate of transition between states. This assumption is inappropriate in practice and can be resolved by semi-Markov models [69]. The second problem is that the system's state is supposed to be observed directly. Such assumption, however, is usually difficult to achieve especially for bearings in reality. The Hidden Markov Model can be used to overcome this problem. So, in what follows, we will concentrate on the use of Hidden Markov Model (HMM) and Hidden semi-Markov Model (HsMM) for bearings prognostics in the literature.

Hidden Markov Model composes of two stochastic processes: a Markov chain which is unobservable representing the real health state of the component, and an observable process which could be the observed CM information or the extracted features. An example of a five states HMM for bearings is illustrated in Figure 2.8.

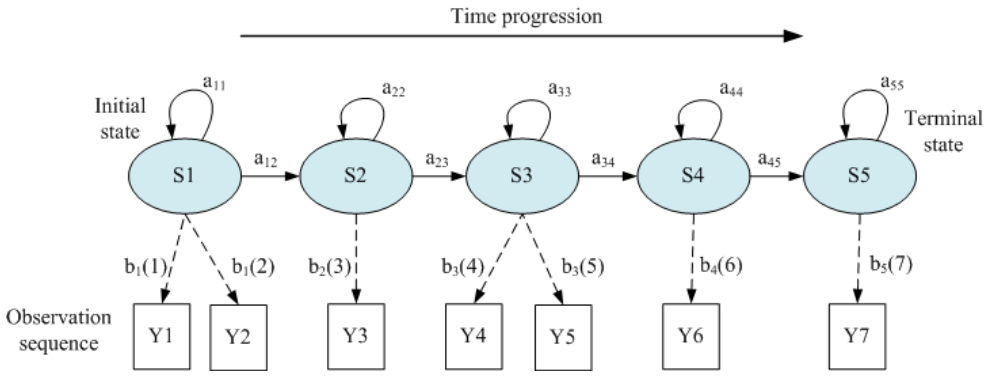


Figure 2.8: An example of an 5-state HMM model

In [169], Zhang et at. proposed an adaptive stochastic fault prediction model base on HMM. In their work, the authors used the principal component analysis (PCA) technique for feature extraction from raw vibration data. The extracted features are then used to train the HMM. One of the key advantage of their algorithm is its capability for on-line learning. However, for the purpose of RUL estimation, the HMM was trained to predict the evolution of actual health indicator, so that it still requires a predefined failure threshold. Along this line of work, Yu in [164] presented a HMM based framework for offline modeling and online health monitoring and assessment, as depicted in Figure 2.9. In his paper, an HMM-based Mahalanobis distance (MD) is proposed for quantifying the deviation degree of the current state with the healthy state space of bearings. Ocak et al. also used the HMM for tracking the severity of bearing fault but they implemented the wavelet packet decomposition (WPD) for feature extraction purpose [115]. Recently, Medjaher et al. [100] and Tobon-Mejia et al. [142] developed the Mixture of Gaussians Hidden Markov Models (MoG-HMMs) represented by Dynamic Bayesian Networks to represent the evolution of the component's health condition from which its RUL can be calculated.

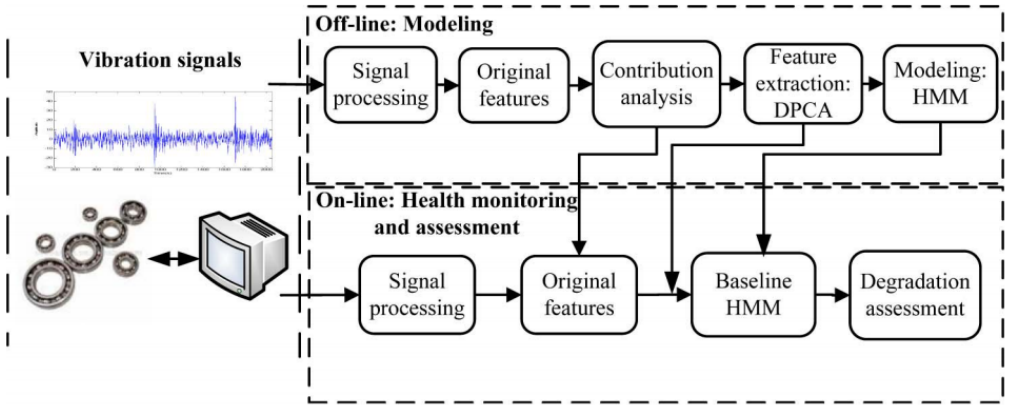


Figure 2.9: HMM based framework for bearing health degradation monitoring [164]

An inherent limitation of HMM technology is that its state duration follows an ex-

ponential distribution. In other words, HMM does not provide adequate representation of temporal structure. So that, for a more powerful prognostics tool, HsMM has been introduced in the literature. As a generalization of HMM, HsMMs are useful in many engineering applications such as signal estimation, diagnostics and prognostics. It is traditionally defined by allowing the unobserved state process to be a semi-Markov chain. In other word, it relax the inherent Markovian property of the Markov-based models and allows modeling the state duration by any arbitrary distributions such as Gamma, Weibull or Gaussian. Thus, HsMM possess both the flexibility of hidden Markov chain for approximating complex probability distributions and the flexibility of semi-Markov chains for representing temporal structures.

The HsMM has four basic problems that must be solved to be of use as a prognostics tool. Dynamic programming techniques have been developed to efficiently solve these problems. A detailed description is given by [32] and [165]. It should be noted that, although it has been successfully applied in speed recognition, the literature of HsMM in RUL estimation is very limited. It has just attracted some attentions in health state prognostics since 2006. For example, Dong *et al.* showed a case study of a HsMM based method on health diagnostics of an UH-60A Blackhawk main transmission and RUL estimation of pumps [33]. Through the experimental studies, the authors argued that HsMM offers several significant advantages over HMM in equipment health prognostics. Further, Dong has incorporated an Auto-regressive structure in a HsMM (AR-HsMM) to avoid the Markov chain's memory-less property [31]. However, this model used only partial history information and was not completely free from the memoryless assumption. Bechhoefer *et al.* also used HsMM for health diagnostics and prognostics, applied for generator shaft of a helicopter [7].

2.3.2 Continuous-state deterioration modeling

In practice, many industrial systems are under continuous physical degradation due to erosion [24], corrosion [114], crack propagation [147], or wear [22], etc... These phenomena can be described by continuous processes. This section reviews the models that consider the deterioration as a continuous process.

The continuous deterioration modeling can be divided into three main categories: Physic-of-failure models, general path models and stochastic processes.

2.3.2.1 General path model

General path model (also called random coefficient regression model) was first introduced by Lu and Meeker to present a general path of a population of units [95]. It is commonly used in industry as well as in academic domain to describe a phenomenon of continuous degradation [93, 101]. The idea is to adjust degradation measures by a regres-

sion model with random and/or fix coefficients. Given the observed sample degradation $S(t)$ at time t, the general degradation model can be represented as:

$$S(t) = g(t, \lambda, \theta) + \epsilon \tag{2.1}$$

where $S(t)$ denotes the deterioration measure at time t, λ and θ are the vectors of coefficients of the random and the fixed effect respectively, g is a deterministic continuous function representing the underlying degradation and ϵ corresponds to measurement errors. The different forms of g (i.e. linear, concave, convex, etc.) allow to model different mechanisms of degradation, so that this model can be easily adapted to several deterioration categories in practice. One advantage of the general path models is that they are relatively simple to construct and the model parameters can be estimated efficiently by classical statistical methods [101]. Another advantage is the direct application of measures of degradation in the statistical analysis. The reliability function or the remaining useful lifetime distribution can be often obtained in an analytic form.

From the first model of Lu and Meeker, many applications of the general trajectory model have been implemented in the literature [96, 132, 173], and its theory is becoming more sophisticated. Indeed, the fundamental assumptions of random-effect model are listed by [155]. Haghighi et al. synthesized in [50] various approaches (parametric, semi-parametric and non-parametric) to estimate the survival function based on the general path model. The review article of Si et al. [136] and the series of work of Gebraeel et al. [44, 40, 43, 41] summarize several models and random effects regression models to predict the residual life of a system.

The analysis of previous studies shows that the general path model is relatively simple to construct and can be easily adapted to different categories of degradation by modifying its underlying function g. The model parameters can be estimated efficiently by conventional statistical methods [101]. Another advantage of this model is the direct application of degradation measures in the statistical analysis. The reliability or residual life calculated from the model is often obtained in analytical form.

2.3.2.2 Lévy processes

The second category of continuous deterioration models is the Lévy processes. The class of Lévy processes is a widely used tool for modeling the evolution of the degradation of systems. A stochastic process $\{x_t\}_{t\geq 0}$ is called a Lévy process if:

- $x_0 = 0$
- It has independent increments, that is for any $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$: $x_{t_2} - x_{t_1}, x_{t_3} - x_{t_2}, \cdots, x_{t_n} - x_{t_{n-1}}$ are independent

- It has the property of stochastic continuity:

$$\lim_{h \to 0} \mathbb{P}\left(|x_{t+h} - x_t| \geq \epsilon\right) = 0 \quad \forall \epsilon > 0$$

If the increments are stationary, it is called homogeneous Lévy process, otherwise it is non-homogeneous [2].

Note that the third property of the Lévy process does not imply that it is always a continuous process. In fact, contrary to the general path model, the Lévy process gives a degradation process from an infinite number of independent microscopic jumps. This makes the modeling of component deterioration more flexible and appropriate, especially in a dynamic operating environment since it can take into account the temporal uncertainties. The drawback of using the Lévy process for deterioration modeling is that the theoretical calculation may be sometime very difficult. In the literature, the choice of modeling a continuous degradation phenomenon with the class of Lévy processes is often based on a Gamma or a Wiener process.

2.3.2.3 Gamma process

A Gamma process is a stochastic process with independent, non-negative increments following a Gamma distribution with an identical scale parameter. It was first proposed by Abdel-Hameed as a proper model for degradation occurring random in time [1]. One advantage of using this process for deterioration modeling is that the required mathematical calculations are relatively straightforward. The RUL can be hence obtained in an analytical form. The Gamma process is suitable to model gradual damage monotonically accumulating over time (Figure 2.10) in a sequence of tiny increments, such as wear, fatigue, corrosion, crack growth, erosion, consumption, creep, swell, degrading health index, etc. [145].

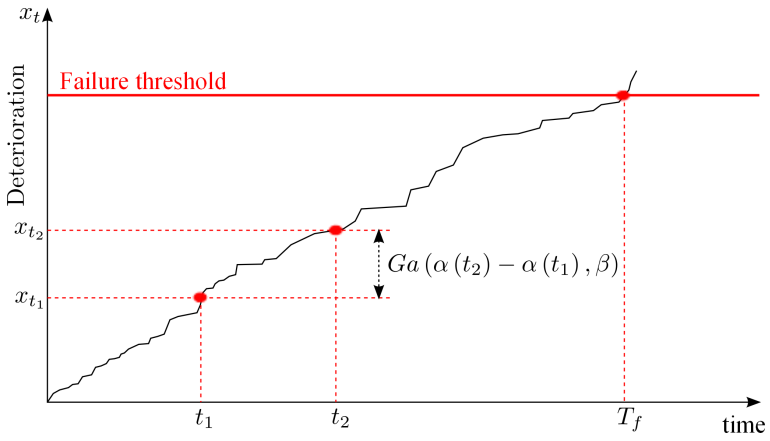


Figure 2.10: Deterioration evolution modeled by a Gamma process

If $\{x_t\}_{t \geq 0}$ is a Gamma process with shape parameter $\alpha(t)$ and scale parameter β then:

a. $x_0 = 0$ with a probability equal to 1.

b. $x_{t_2} - x_{t_1} \sim Ga\left(\alpha(t_2) - \alpha(t_1), \beta\right)$

c. x_t has independent increments

where $Ga\left(\alpha(t), \beta\right)$ denotes the Gamma distribution with shape parameter $\alpha(t)$ and scale parameter β. Its probability density function is given by:

$$f\left(u\right) = \frac{\left(\frac{1}{\beta}\right)^{\alpha(t)}}{\Gamma\left(\alpha\left(t\right)\right)} u^{\alpha(t)-1} \exp\left(-\frac{u}{\beta}\right) \tag{2.2}$$

with $u > 0$ and $\Gamma\left(\cdot\right)$ signifies the Gamma function:

$$\Gamma\left(y\right) = \int_0^\infty u^{y-1} e^{-u} du \quad \texttt{for } y > 0 \tag{2.3}$$

The cumulative distribution function of the Gamma distribution is also given by:

$$F\left(u\right) = 1 - \frac{\Gamma\left(\alpha\left(t\right), \frac{u}{\beta}\right)}{\Gamma\left(\alpha\left(t\right)\right)} \tag{2.4}$$

where $\Gamma\left(a, z\right) = \int_z^\infty u^{a-1} e^{-u} du$ for $z \geq 0$, $a \geq 0$ corresponds to the incomplete Gamma function.

Over a time interval of length t, the average degradation speed-rate for a Gamma process is $\alpha\left(t\right) \cdot \beta$ and its variance $\alpha\left(t\right) \cdot \beta^2$. The choice of α and β allows to model various degradation behaviors from almost-deterministic to very chaotic. Given the degradation data, these parameters can be estimated using classical statistical methods such as maximum likelihood method, moment method, Bayesian statistics method, etc. [145].

If $\alpha(t)$ is a linear function of time i.e. $\alpha(t) = \alpha \cdot t$, the process is called stationary or "homogeneous" Gamma process. In this case, the expected deterioration is linear over time. The stationary of the gamma process basically follows from the property that increments are independent and have the same distribution type as their sum. In mathematical terms, this property is covered by the so-called infinite divisibility i.e. a random variable following a Gamma law with parameters α and β can be written as the sum of n independent variables following a Gamma law with parameters $\frac{\alpha}{n}$ and β.

A good survey of applications of the Gamma process and its properties can be found in [145].

Since Gamma process is suitable to model gradual damage monotonically accumulating over time in sequence such as wear, fatigue, corrosion, crack growth, creep and

degradation health index (e.g. extracted from vibration data), it has been widely used in the literature for bearings deterioration modeling and RUL estimation. For example, Park and Padgett considered Gamma processes in accelerated degradation modeling using the fatigue crack growth data and then inferred the lifetime distribution at the use condition [119]. Lawless and Crowder incorporated the gamma process with random effects to model unit-to-unit differences in the degradation signals among a population of similar components. The proposed method was also verified using crack-growth data [74]. Zhou et al. proposed a Gamma-based state space model to predict engineering asset life when multiple degradation indicators are involved and the failure threshold on these indicators are uncertain. A case study using the vibration data of bearings from LNG pumps has been conducted [172]. Zhang et al. [170] also used the Gamma-based state space model for RUL estimation of a gearbox by taking a few wear monitoring data as direct condition information and taking the plentiful vibration data as indirect condition information. They combined the Expectation-Maximization (EM) algorithm with the particle filter to estimate the real states as well as the parameters of the model. For more specific details of Gamma process and its application in the context of maintenance, one can refer to the excellent review of [145].

The use of Gamma process to model the evolution of deterioration indicators has a nice property: the sum of Gamma distributed increments is again a Gamma variable. The mathematical calculations for modeling degradation through Gamma process are hence relatively straightforward [48]. Additionally, using Gamma process-based model, the distribution of the RUL can be obtained. Another advantage of this model is that its physical meaning is easy to understand. In contrast to the random coefficient regression methods, Gamma process-based degradation models can take the temporal variability into account as argued by Pandey et al. in [118]. However, we should note that the Gamma process seems only appropriate to represent degradation by strictly monotonic process. This property is usually held for the direct health indicators such as the crack depth on a gear. But this type of indicator is very difficult to measure on-line in reality, especially in the case of a paper machine. When applying the Gamma process for indirect indicators (e.g. features extracted from vibration signals), the non-linear relationship between these indicators and the actual health of equipment should be taken into account.

2.3.2.4 Wiener process

The Wiener process $\{X(t)\}_{t\geq 0}$, also called Gaussian process or Brownian motion with drift, is a continuous-time stochastic process with drift parameter $\mu(t)$ and variance parameter $\sigma^2\,\sigma > 0$:

$$X\left(t\right) = \mu\left(t\right) + \sigma \cdot B\left(t\right) \tag{2.5}$$

where $B(t)$ is Brownian motion, i.e. it is normally distributed variable with mean 0 and variance t, known also as the white noise in several fields.

Similar to the Gamma process, if the drift parameter is a linear function of time $\mu(t) = \mu \cdot t$, the Wiener process is called homogeneous and has the following properties:

- $X(0) = 0$ with a probability equal to 1.
- The increments are independent normally distributed: for any $t_i < t_j$, $X(t_i) - X(t_j)$ is normally distributed with mean $\mu \cdot (t_j - t_i)$ and variance $\sigma^2 \cdot (t_j - t_i)$.
- In particular, for any time $t \geq 0$ the mean value of $X(t)$ is $\mu \cdot t$ and the variance is $\sigma^2 \cdot t$.

A characteristic feature of this process, in the context of deterioration modeling, is that a system degradation alternately increases and decreases (c.f. Figure 2.11), similar to the exchange value of a share [145]. For this reason, unlike the Gamma one, the Wiener process is inadequate in modeling monotonous deterioration processes. Instead, it has been widely used for modeling the non-monotonous degradation.

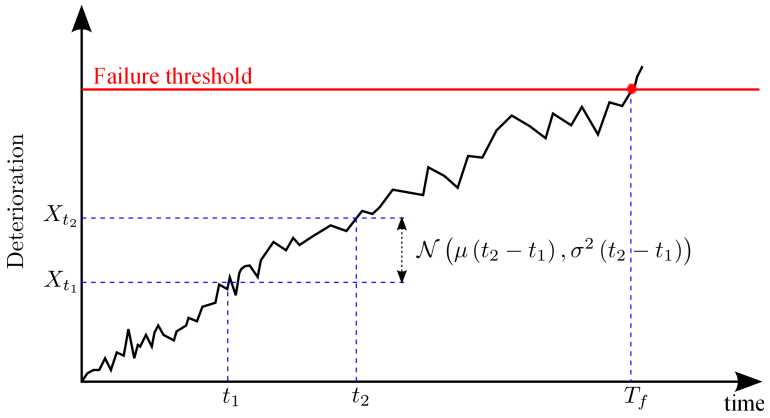


Figure 2.11: Homogeneous Wiener process. Evolution of the degradation to failure

The parameters of the Wiener process can also be estimated using classical statistical methods such as maximum likelihood method, moments method, Bayesian method, etc. given the degradation data.

2.3.3 Environment effects in deterioration modeling

The deterioration models analyzed above can model a wide range of degradation processes. However, they suffer an important limitation that is these models don't take into account the variation of dynamic environment in which the system operates. In reality, changes of environmental factors, such as changes of temperatures, humidity, etc. may significantly affect the deterioration processes. Several models have been proposed in the literature to model the effects of environmental factors in component deterioration processes. They will be reviewed in this section.

In the literature, evolution of environmental factors (usually called covariates) is often represented by a stochastic process and their impacts on the continuous degradation process can be divided in four principal types [64]: punctual impact and temporal impact on degradation process, impact on failure threshold and impact on failure modes (c.f. Figure 2.12).

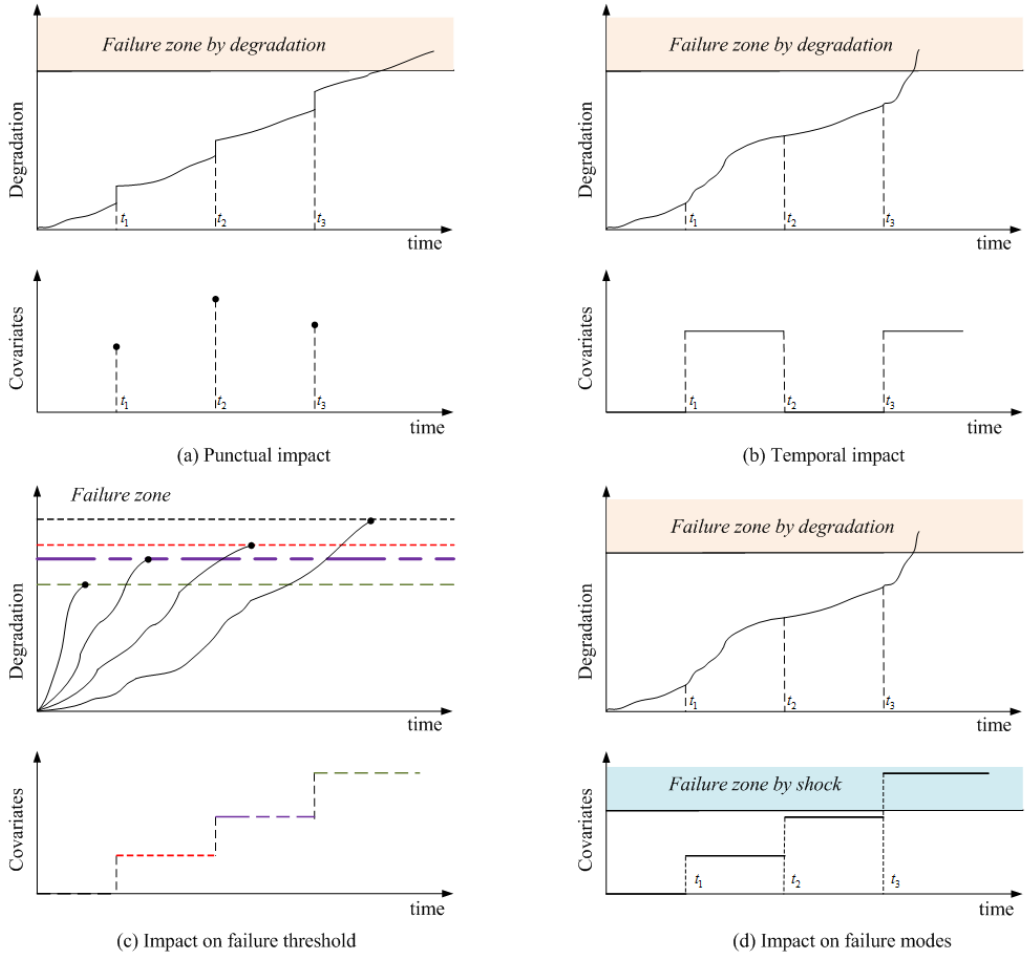


Figure 2.12: Covariates impacts on the deterioration model [64]

- Punctual impact on degradation

  For this impact type, the evolution of covariates can generate a point impact on the degradation. It means that an abrupt change in covariates states causes a sudden increase in the level of degradation of the system (c.f. Figure 2.12(a)). This type of impact is considered in [27, 28, 123, 157].

- Temporal impact on degradation

  For this second type, the covariates have a temporal impact on the degradation process. Very often, in this case, the speed and / or the variance of the degradation process are driven by covariates (c.f. Figure 2.12(b)). In the literature, the proportional hazards model of Cox [26] and its derivations (model covariates proportional

pattern explicit hazards, etc... [49]) are the most used models to express this relationship. Bagdonavicius and Nikulin [5] and [74] proposed to incorporate covariates in a Gamma process through its shape or scale parameter. Other examples can be found in [85, 125, 138].

- Impact on failure threshold

  The third type of impact does not change the state of degradation of the system, but its failure threshold (c.f. Figure 2.12(c)). In this case, the failure threshold is no longer fixed, but represented as a random variable depending on covariate distribution. As noted by Lu and Meeker in [95], the most commonly used distributions are uniform, normal or lognormal. For example, Wang and Coit developed in [152] a degradation model whose failure threshold is distributed according to a normal distribution, while Feng et al. proposed in [36] a log-normal distribution to model the threshold.

- Impact on failure modes

  The last type of impact considers that covariates can lead to a further failure mode of the system (c.f. Figure 2.12(d)). Since the failure of a system can be divided into shock and failure by degradation, we are interested in models representing both these two failure modes. Such models are called Degradation- Threshold -Shock model (DTS) [85]. We consider a system fails as soon as the level of degradation exceeds a critical threshold or due to shock occurs. This model was proposed for the first time by Lemoine and Wenocur [86]. After that, a large number of extensions and applications are developed [28, 63, 90, 123, 146]. DTS models with covariates can be considered as a general model that combines the previous types of impacts.

Covariate based hazard models

One of the most popular covariate-based models existing in the literature is the Proportional Hazard Model (PHM). The PHM model is firstly proposed by Cox in 1972 [26]. Due to its generality and flexibility, PHM was quickly and widely applied in different areas of lifetime analysis such as the biomedical, reliability and economics. Almost all covariate models are based on PHM theory. For this reason, we will study the use of PHM model on the estimation of the RUL in this section.

The conventional PHM assumes that the Hazard rate of a system at time t consists of two multiplicative factors, a baseline hazard function and a function of covariates, that is

$$h\left(t|z(t)\right) = h_0\left(t\right)\psi\left(\beta z\left(t\right)\right)$$

where $h_0\left(t\right)$ is the unspecified baseline hazard function which is dependent on time only and without influence of covariates. The positive functional term $\psi\left(\beta z\left(t\right)\right)$ is dependent on the effects of different factors, which have multiplicative effect on the baseline hazard function, where β is a column vector consisting of the corresponding regression coefficients

and $z(t)$ is the covariates vector. These coefficients are generally estimated by maximizing the partial likelihood function without specifying the baseline hazard function $h_0(t)$.

In this PHM model, the baseline hazard function $h_0(t)$ can be either parametric or non parametric. In the case that this baseline function is non parametric, it can be estimated using the failure event data and censored data, such as the Kaplan-Meier estimator [131]. When it is specified to a parametric distribution, the Weibull distribution is widely used as the baseline function for PHM [136]. In the following, we will investigate how the RUL can be calculated from the hazard rate of a system.

We know that the reliability of a system can be deduced from the hazard rate by the formula

$$R(t|Z(t)) = P(T > t|Z(t)) = \exp\left\{-\int_0^t h(s|z(s))\,ds\right\}$$

where $Z(t) = \{z(s), 0 \leq s \leq t\}$ denotes the entire covariate information history up to time t and $z(s)$ is the information obtained at time s.

Based on the PHM model, we can define the RUL as $X_t = \{x_t : T - t \mid T > t, Z(t)\}$ where T is the lifetime. The power density function of the RUL at time t can be then formulated as follows

$$f_{X_t}(x_t|Z(t)) = \frac{f(t + x_t|Z(t))}{R(t|Z(t))} = h(t + x_t|Z(t))\,\frac{R(t + x_t|Z(t))}{R(t|Z(t))}$$

From the above equation, we can see that a PHM needs the event data such as failure and censored data as well as CM information to estimate the RUL without an exact failure threshold. For example, Ghasemi and Hodkiewicz developed in [47] a prognostics model to estimate the Mean Residual Life of Rail Wagon Bearings within certain confidence intervals. The prognostics model was constructed using a PHM approach, informed by imperfect data from a bearing acoustic monitoring system, and a failure database. Van-Tung et al. combined the PHM with the support vector machine (SVM) to assess performance degradation and to predict the RUL of a machine [143]. However, they also used the ARMA model to model the degradation path, so that they still needed define an unacceptable level or a failure threshold. You et al. [163] developed a two-zone PHM to predict equipment's RUL, in which the life-cycle was divided into two zones, i.e. a stable zone representing the normal operation and a degradation zone representing the degraded states of equipment. Based on the detection of the change-point from the stable zone to the degradation one, the RUL of equipment can be accurately estimated.

In the framework of the covariate-based models, there are some other variants based on the basic PHM to achieve hazard modeling, such as proportional intensities model [149], accelerated failure time model, additive hazard model, proportional intensities model, etc. For a more intensive study of these models, one can refer to the review of Gorjian et al. in [48].

The main problems of using covariate-based hazard models for RUL estimation are

- The models mix the casual relationship of different covariates. In reality, some covariates may impact on the hazard rate and some others may be influenced by the hazard. So they should be modeled differently.
- When the evolution of covariates is stochastic, we have to use an other process (e.g. Markov chain) to describe the covariate process. This is an added burden to the model.
- From the definition of PHM, we have $\frac{h(t|z_1(t))}{h(t|z_2(t))} = \frac{\psi(\beta z_1(t))}{\psi(\beta z_2(t))}$, which is known as the proportionality assumption. It means that the rate of hazard function is proportional to the difference of covariates. This assumption, however, is not always true in practice.
- The estimation of coefficients requires sufficient failure event data and associated CM information which may not always be available in reality.

2.4 Conclusion

Prognostics plays an important role in predictive maintenance implementation. It makes the predictive maintenance being more advanced than the condition-based one thanks to its ability to predict the temporal evolution of system's health as well as estimate the residual lifetime of the system.

This chapter reviews several prognostics approaches existing in the literature. The advantages and the disadvantages of each approach were also analyzed. In the framework of the SUPREME project, the WP3 is dedicated for the development of the embedded condition monitoring system (ECMS). A large volume of condition monitoring data could be made available during the project execution. The data-driven prognostics approach appears to be suitable for further investigation within the project. Moreover, as there are several sources of uncertainties that can affect the RUL estimation results, a solution for uncertainties management is to consider the RUL as a random variable and to characterize the RUL by a probabilistic distribution. For this reason, we are interested in the followings of this manuscript in the stochastic deterioration models with the RUL estimation based on the evolutionary or trending techniques.

The impacts of the covariates on the deterioration processes were also reviewed and analyzed in this chapter. Such analysis is essential since in practice a machine rarely operates under a static environment. In the context of the WP4, such impacts can be described by the affect of maintenance strategies change, i.e. change in production orders, on the deterioration dynamic of the machine. This problem will be addressed in the next chapter. Moreover, based on the review and the analysis presented in this chapter, an adaption of the state-of-the-art deterioration models and RUL estimation methods was

carried out and implemented for the data acquired from a test-bench constructed in the framework of the SUPREME project (c.f. Chapter 3).

Beside that, the analysis of the state-of-the-art helps also to identify the limitations of the existing methods, basing on which more advanced deterioration models can be developed for the SUPREME project. The main works presented in this chapter have been reported in the deliverable D2.12 entitled "State-of-the-Art and Beyond the State-of-the-Art" of the SUPREME project [57].

Chapter 3

State-of-the-art adaptation and application to the SUPREME project

3.1 Introduction

After having studied the state-of-the-art of deterioration models and the associated RUL estimation methods, the second stage of this thesis was dedicated to investigate their applications in the framework of the SUPREME project. This chapter presents some works realized as well as the obtained results in applying the state-of-the-art with both simulated and real data.

When applying a deterioration model for a real system, the first step is to learn the model, i.e. to estimate its parameters, from the available data set. A large amount of data is hence often required for this purpose. It should be noted that the data here must be the one that represents the deterioration phenomenon of the system, from the initiation of the defect until the system failure. Unfortunately, such deterioration data is quite difficult to be acquired, especially for a big and complex system like the paper machine. In effect, due to its critical role, the paper machine cannot be allowed to deteriorate until the complete failure. In order to overcome this difficulty, a test bench has been constructed to simulate the real deterioration phenomena of critical components that could be found in real production systems. Based on this test bench, deterioration data set could be created, allowing the test and the validation of the developed approaches in the framework of the SUPREME project. For this reason, in this chapter, we first apply the state-of-the-art deterioration models for the data acquired from this test bench. More specifically, in Section 3.2 the deterioration of a bearing, a component that is commonly used within a paper machine is considered. By investigating the temporal evolution of the health indicator constructed from the acquired vibration signal, a so-called incubation/propagation model is adapted. The RUL estimation method for this model is also discussed.

Moreover, as already presented in Chapter 1, the output of the "reliability" sub-module, i.e. the deterioration models and the RUL estimation results, is used as input for the "maintenance" one in order to obtain a better dynamic adaptation of maintenance strategies for the monitored component. Generally, a maintenance strategy concerns the modification of the intervention plan, i.e. the inter-inspection intervals, so that an optimal long-run expected maintenance cost rate can be obtained [64]. However, regarding the continuous production system such as the paper machine targeted by the SUPREME

project, a shutdown of the machine could lead to very expensive losses. Therefore, the output of the maintainability sub-module in this case is aimed to modify the production orders of the machine. The purpose is to have a best adaption of the load applied on the deteriorated component so that its lifetime could be extended, at least it could operate until the next planned intervention action. This leads to the fact that the machine could operate under different loads and the changes of the applied load are not stochastic but deterministic or planned. By supposing that the loads could affect the deterioration dynamic of the monitored component, a load-dependent deterioration process is applied and introduced in Section 3.3.

Furthermore, within the WP4, the fact that the reliability sub-module is developed by Grenoble-INP and CETIM in France while the maintainability one is conducted by IPA in Germany leads to the requirement of exchanging the information between these partners. For example, the estimated RUL must be defined and its format must be unified before it can be sent to IPA for the dynamic maintenance adaptation. These problems are addressed in Section 3.4.

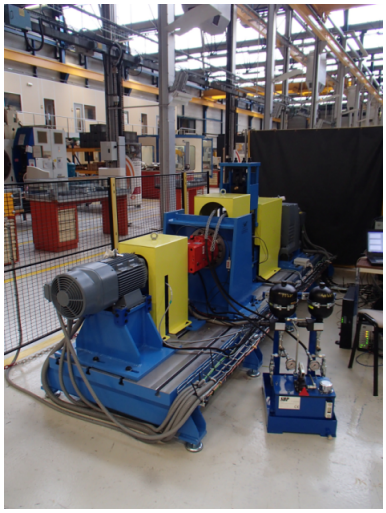
3.2 Test bench at CETIM

3.2.1 The test bench

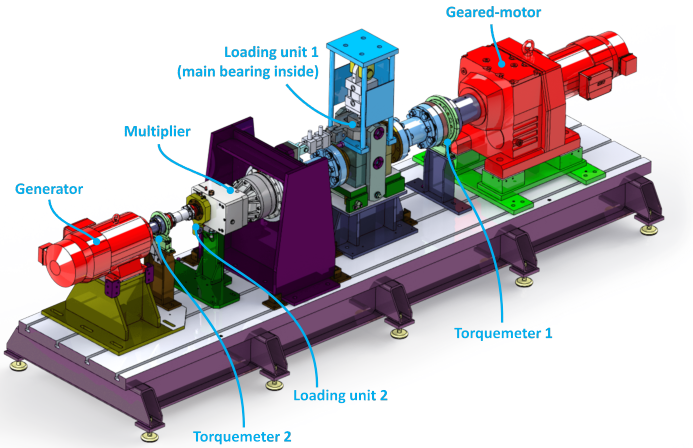
The test bench is installed at CETIM in Senlis, France. It is designed to represent a wind turbine architecture but at a smaller scale (c.f. Figure 3.1a). In effect, at the input of the bench, a geared-motor generates the rotation and simulates the wind power in reality. The main shaft bearing is loaded in axial and radial directions by two hydraulic actuators (in the loading unit). This loading unit represents the forces of the wind and the weight of blades. The multiplier gearbox has a ratio of $100.75 : 1$. It multiplies the rotational speed of the main shaft so that the generator can operate between 500 and 3000 revolutions per minute (rpm) which is usual speed on the real wind turbines. A generator is located at the output of the test bench in order to transform mechanical energy into electricity.

One of the main objectives of the test bench in the SUPREME project is to simulate real deterioration processes in practice. Thanks to the two additional loading units, defects could be generated by fatigue, i.e. by applying artificial loads on the bearings and the gearboxes. These components are selected for the deterioration generation since they play an important role in many practical production systems. In effect, a break of one of these components could lead to a complete failure of the whole system.

Figure 3.2 summarizes the eight endurance test programs that have been planned to be carried out in the framework of the SUPREME project. These tests represent different defect scenarios that can take place for the test bench. However, at the time of writing this thesis, there are only three among them that have been completed while the others



(a) An overall view



(b) Main components

Figure 3.1: The test bench installed at CETIM

are still in progress. Therefore, only the data obtained from these completed tests will be considered in this chapter.

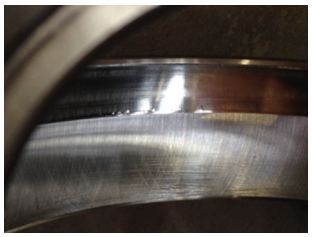
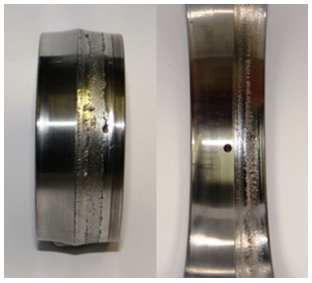
Test	Aim of the test	Progress	Start date
I. 1	**Deterioration of the main bearing**		
I. 1.1	Up to the first detected defect on the main bearing	✓ 100%	16/12/2013
I. 1.2	Medium degradation on the main bearing	✓ 100%	23/09/2013
I. 1.3	High degradation on the main bearing		
I. 1.4	High degradation on the main bearing (other brand)	✓ 100%	05/03/2014
I. 2	**Planetary gear deterioration**	15%	12/05/2014
I. 3	**Parallel gear deterioration**		
I. 4	**Planetary + parallel gears deterioration**		
I. 5	**Deterioration of the output bearing**		
I. 6	**Parallel gear + output bearing deterioration**		
I. 7	**All damaged components (health monitoring only)**		
I. 8	**Parallel gears deterioration with emergency stops and high acceleration**		

Figure 3.2: Summary of tests for the test bench

The completed tests, denoted by I.1.1, I.1.2 and I.1.4, represent 3 different defect levels of the main bearing, ranging from a very early damage (i.e. start of spalling) to a very advanced deterioration with spalling on inner race, outer race and rollers. Table 3.1 summaries these tests with some photos illustrating the defects at the end of each test.

It is worth noting that the three tests I.1.1, I.1.2 and I.1.4 are conducted independently. In effect, after having finished one test, the defected bearing is removed and changed by a new one before starting the other tests. The bearings are selected to have the same

Table 3.1: Summary of tests on the main bearing

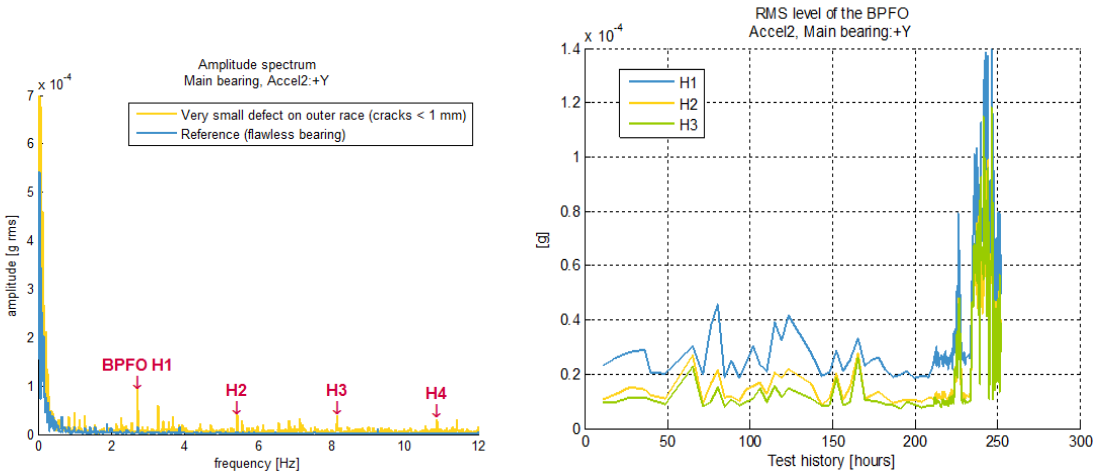
Test	Duration (hours)	Defect level	Photo of defects
I.1.1	258	**Very low damage:** Cracks and start of spalling on the outer race Defect size: 1mm	
I.1.2	215	**Medium damage:** Some spalling on the outer race Defect size: 2mm	
I.1.4	194	**High damage:** Spallings spread over complete the inner and the outer races, some spallings on the rollers Defect size: all the races	

type, i.e. the spherical roller bearings with two rows of rollers. Furthermore, the tests are conducted under the same operating conditions. For these reasons, we can consider the data acquired from these tests are mutually independent and one deterioration model can hence be developed for representing the deterioration of these tests.

3.2.2 Health indicator construction based on vibration signal analysis

The defects on the bearing are often unobservable, especially when the system is operating. However, it could be detected through the condition monitoring data, i.e. by the vibration signals acquired by accelerometers located on the bearing's house. For example, Figure 3.3a represents the spectrum analysis on the main bearing at the end of the test I.1.1 (very small defect) compared to the beginning (flawless bearing) of the test. Abnormal high amplitudes can be observed at the Ball Pass Frequency of Outer race (BPFO) frequencies, which imply the appearance of the defect on the outer race of the bearing. In this figure, H1 represents the BPFO frequency (H1 = 2.721 Hz) while H2, H3 and H4 are the harmonics corresponding to the values of 2xH1, 3xH1 and 4xH1 respectively. If we record the values of the amplitudes at the frequencies H1, H2 and H3

for the whole test, we obtain the evolution in time of these values as shown in Figure 3.3b.



(a) Frequency spectrum (b) Temporal evolution of amplitudes

Figure 3.3: Test I.1.1 - very small defect

In this figure, the amplitudes at H1, H2 and H3 frequencies are relatively small and stable for the period from the time zero until about 225$[h]$ of operation. After that, a considerable increment can be noticed. It can be deduced that the bearing stays in a normal condition from the beginning of test until time $t = 225[h]$. Then, the bearing moves to the defect state. Therefore, the amplitudes at the BPFO frequency can be considered as the indicator representing the deterioration level of the bearing under the test I.1.1.

The same analysis can be performed for the medium and high level of defect tests. Since the defects also appeared on the outer race of the bearing under the test I.1.2, we consider the amplitude at the BPFO frequency as the deterioration indicator for this test as well. Regarding the test I.1.4, since the defects appear not only on the outer race, but also on the inner race and the rollers, a further investigation needs to be conducted. From the frequency spectrum of the bearing at the end of this test, we realize that the amplitude at the BPFI frequency (Ball Pass Frequency of Inner race) is much higher than the one at the BPFO frequency (c.f. Figure 3.4). In other words, the defects are exhibited more clearly at the BPFI frequency than the other one. For this reason, we use the amplitude at the BPFI frequency to be the deterioration indicator for the test I.1.4.

In summary, the temporal evolution of the deterioration indicators for all the tests are shown in Figure 3.5. Compared to the test I.1.1, the tests I.1.2 and I.1.4 show clearer trends in the defected period. This is reasonable since these two tests are designed to produce higher levels of defect than the first one.

Moreover, two stages can be noticed in the temporal evolution of the indicators: one stage in which the indicators values are relatively small and stable and another stage in which they vary significantly in time. The first stage can be considered as an incubation

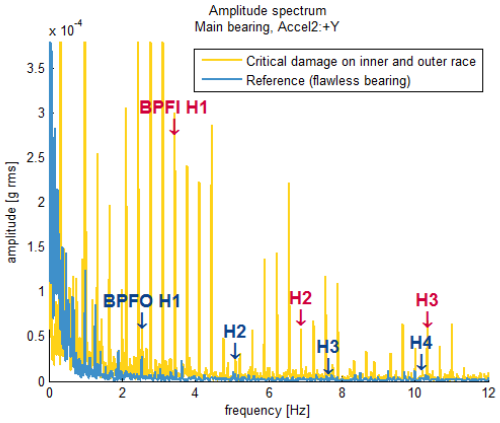


Figure 3.4: Test I.1.4 - Critical defect, spectrum [0 - 12] Hz

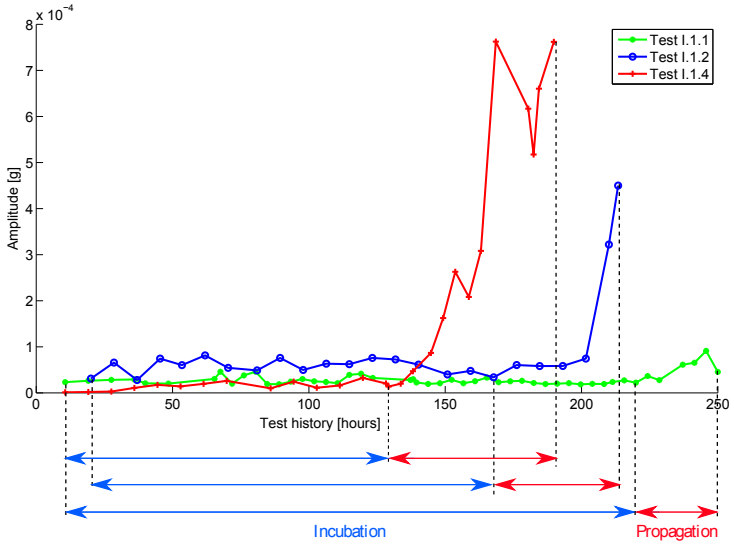


Figure 3.5: Evolution of deterioration indicators

period while the second one is a propagation one. In order to take into account such two-stage deterioration phenomenon, in the next section the so-called incubation/propagation model will be presented and adapted to the deterioration data of the main bearings.

3.2.3 Incubation/propagation model

3.2.3.1 Introduction

As indicated by its name, the incubation/propagation model is used to represent the deterioration processes that progress through two different stages: incubation and propagation. This type of model is motivated from the fact that, once a defect is initiated, there may exist a period at the beginning in which the defect does not expose clear evidences about its appearance. It means that, if we are looking at the health indicators

extracted from raw data, we do not realize any abnormal signs but the defect has already initiated: the indicators are relatively stable in this period. This period is hence called the incubation period. Then, after some period of time, the defect begins to progressively propagate and shows clearer trends in its temporal evolution. This second stage is called the propagation stage. Figure 3.6 illustrates an example of such two-stage deterioration process.

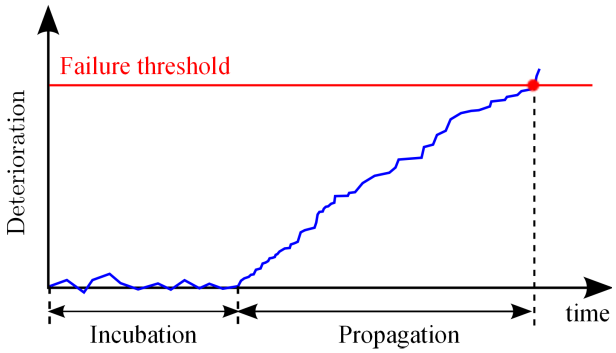


Figure 3.6: Example of the incubation/propagation process

As the deterioration indicator value in the incubation period is relatively small and stable, we are more interested in its duration than in the level of the deterioration within this period. One common way to model the incubation duration is to consider it as a random variable that follows a parametric distribution, such as the Weibull, the Normal or the Poisson distributions. Regarding the propagation period, the stochastic processes studied in Chapter 2, i.e. the Gamma or the Wiener processes, could be used for representing the temporal evolution of the deterioration indicator in this period.

3.2.3.2 Adaption to the test bench case

In the case of the test bench, due to the limitation of the available data (there are only three data sequences corresponding to the three completed tests), the distribution that has the least free parameters could be a good choice from the statistical learning point of view. Moreover, as the vibration signals are acquired periodically by the condition monitoring system, the time can be assumed to be discrete. Consequently, the duration of the incubation period will take the integer values. Compared to others distributions such as Gaussian, Gamma or Weibull, the Poisson distribution has only one parameter appearing to be a suitable one. It is hence chosen for representing the incubation duration in the case of the test-bench.

A discrete random variable x is said to have a Poisson distribution with parameter $\lambda > 0$, if, for $k = 0, 1, 2, \ldots$, the probability mass function of x is given by:

$$f(k;\lambda) = \mathbb{P}(x = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.1}$$

where $k!$ denotes the factorial of k.

Regarding the propagation period, the Brownian motion with drift (i.e. the Wiener process) is suitable for representing the evolution of the deterioration indicators in the test bench case due to its non-monotonic increments. The Wiener process is assumed to be homogeneous with two parameters μ and σ, that is:

$$X(t) = \mu \cdot t + \sigma \cdot B(t) \tag{3.2}$$

where $B(t)$ is the standard Brownian motion, i.e. it is normally distributed variable with mean 0 and variance t.

The parameters of the incubation/propagation model can be estimated through some standard statistic techniques such as maximum likelihood estimator (MLE). In effect, given a set of n samples k_i, for $i = 1, \cdots, n$, the maximum likelihood estimation for the parameter λ of the Poisson distribution is the sample mean:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} k_i$$

In the test bench case, 3 values of the incubation duration in the three tests are recorded, which are 135h, 200h and 230h respectively. We have:

$$\hat{\lambda} = \frac{1}{3}(135 + 200 + 230) = 188[h]$$

In the following, we use the superscript i to imply the test I.1.i. Regarding the propagation period, let x_j^i denote the indicator value at time t_j^i for $j = 1, \ldots, N^i$ where N^i is the number of data points with respect to the test I.1.i. We assume that the deterioration increments, defined by $\delta x_j^i = x_{j+1}^i - x_j^i$, are independent and identical distributed, i.e. they follow a Normal distribution with mean $\mu \cdot \left(t_{j+1}^i - t_j^i\right)$ and variance σ^2:

$$f_{\mu,\sigma}\left(\delta x_j^i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\left[\delta x_j^i - \mu \cdot \left(t_{j+1}^i - t_j^i\right)\right]^2}{2\sigma^2}\right)$$

Denote $\delta X^i = \left(\delta x_1^i, \ldots, \delta x_{N^i-1}^i\right)$, the joint log-likelihood function is given by:

$$\log f\left(\delta X^1, \delta X^2, \delta X^3\right) = \sum_{i=1}^{3} \left(\left(N^i - 1\right) \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{i=1}^{N^i-1} \frac{-\left[\delta x_j^i - \mu \cdot \left(t_{j+1}^i - t_j^i\right)\right]^2}{2\sigma^2} \right)$$

The parameters of the Wiener process can be estimated by maximizing this log-likelihood function. In effect, by letting the partial differential of the function with respect

to μ and σ equal to zero respectively, we obtain:

$$\hat{\mu} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{N^i-1} \delta x_j^i \left(t_{j+1}^i - t_j^i\right)}{\sum_{i=1}^{3}\sum_{j=1}^{N^i-1} \left(t_{j+1}^i - t_j^i\right)^2} \tag{3.3}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{3}\sum_{j=1}^{N^i-1} \left(\delta x_j^i - \hat{\mu}\left(t_{j+1}^i - t_j^i\right)\right)^2}{\sum_{i=1}^{3}\left(N^i - 1\right)} \tag{3.4}$$

Firstly, we estimate the parameter μ by Equation (3.3) and then use this value to estimate σ by Equation (3.4). In this case, we obtain:

$$\begin{cases} \hat{\mu} = 6.47 \cdot 10^{-6} \\ \hat{\sigma} = 1.2 \cdot 10^{-4} \end{cases}$$

3.2.3.3 Deterioration regeneration

This section is dedicated to demonstrate the capacity of the incubation/propagation model in representing the dynamic in the evolution of the defects found in the test-bench case. A possible way to do that is to regenerate some trajectories of "artificial" (i.e. simulated) health indicators from the model learned in the previous section.

Each trajectory of indicator is regenerated in two steps. Firstly, the incubation duration is randomly generated from the Poisson distribution with the parameter $\hat{\lambda}$. The health indicators are fixed at a small value and are kept constant during this period. Then, the temporal evolution of the indicators in the propagation period is simulated from Equation (3.2). Figure 3.7 shows both the simulated trajectories and the real ones of the health indicators.

The real health indicator trajectories corresponding to the three tests I.1.1, I.1.2 and I.1.4 show the variety in the incubation times and represent the three different level of defect at the end of each test. However, as earlier mentioned, these three tests are conducted in the same operating conditions with the same type of component (i.e. the main bearing). The different between the last values of the indicators is due to the fact that these tests are stopped at different moments. Therefore, we can intuitively assume that these curves express the same dynamic in temporal evolution within the propagation period.

Figure 3.7 shows that the "artificial" indicators generated from the learned incuba-

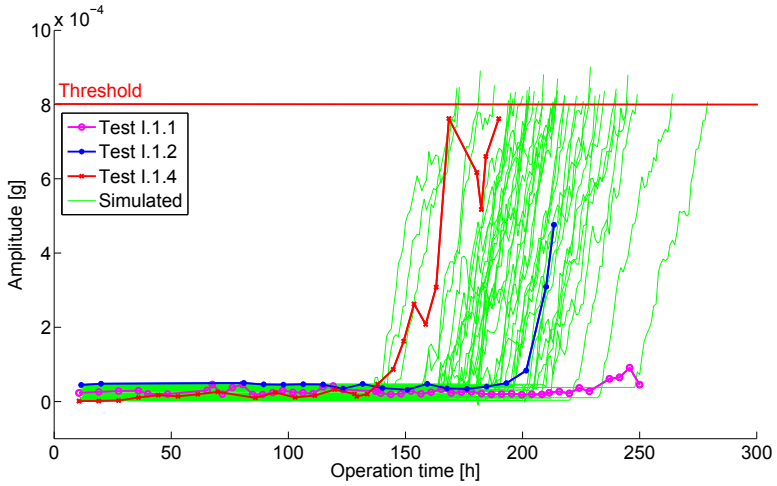


Figure 3.7: Real vs simulated indicators

tion/progpagation well represent the temporal evolution of the real ones. In effect, the simulated incubation times ranges from $140[h]$ to $230[h]$ demonstrating that the variety of the real ones is well taken into account. Furthermore, the dynamic of the real indicators in the propagation period is also well represented. It can be deduced that the learned incubation/propagation model has well represented the real deterioration phenomenon of the bearings in the case of the test bench.

3.2.3.4 RUL estimation

In this section, we demonstrate the RUL estimation procedure for the adapted incubation/propagation model in case that a failure threshold is available.

Supposing that the bearing fails once its health indicators reach for the first time the threshold L. Let x_{t_0} denote the deterioration level at the current time t_0. According to the first hitting time principle (c.f. Chapter 2), the RUL of the bearing at this time can be calculated by taking into account the stage, i.e. incubation or propagation, in which the bearing is belonging to:

$$\begin{aligned}\mathbb{P}\left(RUL < t \mid x_{t_0} < L\right) = \mathbb{P}\left(RUL < t \mid x_{t_0} < L, \text{ the equipment is in stage 1 at } t_0\right) \\ + \mathbb{P}\left(RUL < t \mid x_{t_0} < L, \text{ the equipment is in stage 2 at } t_0\right) \quad (3.5)\end{aligned}$$

We consider two following scenarios:

- Scenario 1: The bearing is still in the incubation period at t_0.

  In this scenario, the RUL can be calculated as the sum of the two residual times: the

remaining time of being in the incubation stage and the duration of the propagation period. We have:

$$RUL = RUL_1 + RUL_2 = \tau - t_0 + T \tag{3.6}$$

where $RUL_1 = \tau - t_0$ is a random variable representing the remaining time in the incubation period and $RUL_2 = T$ represents the propagation period duration. Supposing that these two variables are independent, the probability density function (pdf) of the RUL can be calculated by:

$$f_{RUL}(t) = \left(f_{RUL_1} * f_{RUL_2}\right)(t) \tag{3.7}$$

where the symbol $*$ denotes the convolution product.

If the incubation duration follows a Poisson distribution, the conditional remaining incubation time given the deterioration level at time t_0 follows a truncated Poisson distribution. That is:

$$f_{RUL_1}(t) = f_\tau(t \mid \tau > t_0) = \frac{f_\tau}{1 - F_\tau(t_0)} \tag{3.8}$$

where $F_\tau(\cdot)$ denotes the cumulative mass function of the Poisson distribution.

Regarding the second pdf f_T, since the bearing is currently in the incubation period, its deterioration level is relatively small. We can suppose that the bearing will start the propagation stage at $x_{t_0+RUL_1} \approx 0$. The pdf f_T hence coincides with the pdf of the first hitting time (FHT), i.e. the time to exceed the threshold L for the first time from $x = 0$, given the stochastic process used for modeling the temporal evolution of the indicators in this period. If the homogeneous Wiener process with drift is used (c.f. Equation (3.2)), the FHT will follow the inverse Gaussian distribution with two parameters η and ξ, i.e. $FHT \sim IG(\eta, \xi)$ whose pdf is given by:

$$f_{RUL_2}(x) = \sqrt{\frac{\xi}{2\pi x^3}} \cdot \exp\left[-\frac{\xi}{2\eta^2}\frac{(x-\eta)^2}{x}\right] \tag{3.9}$$

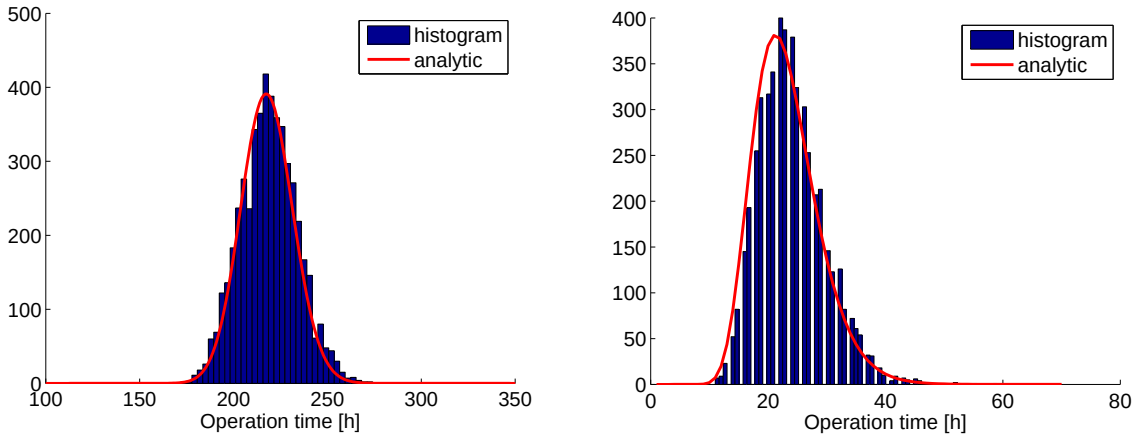
where

$$\begin{cases} \eta = \dfrac{L}{\mu} \\ \xi = \dfrac{L^2}{\sigma^2} \end{cases}$$

From Equations (3.8) and (3.9), the RUL pdf can be evaluated, i.e. numerically, by Equation (3.7).

- Scenario 2: The bearing is in the propagation period at time t_0

  Suppose that the deterioration indicator x_{t_0} at time t_0 could be perfectly estimated or determined. Given x_{t_0}, the bearing will fail at time t_f if the deterioration accumulation between t_0 and t_f exceeds for the first time the level $L - x_{t_0}$. In other



(a) Scenario 1 (b) Scenario 2

Figure 3.8: RUL estimation results for the two scenarios

words, the bearing's RUL in this case shares the same pdf form with the variable representing the first hitting time to reach the threshold $L - x_{t_0}$ from 0. In case that the evolution of the deterioration is modeled by a Homogeneous Wiener process, the pdf of the RUL can be given by Equation (3.9) where the parameters η and ξ are now become:

$$\begin{cases} \eta = \dfrac{L - x_{t_0}}{\mu} \\ \xi = \dfrac{(L - x_{t_0})^2}{\sigma^2} \end{cases}$$

To verify the exactitude of the above RUL calculation, a numerical investigation is conducted. The analytic RUL pdf is compared to the histogram obtained by the Monte Carlo simulation method. Figure 3.8 shows the RUL estimation results in the two scenarios for the incubation/propagation model.

3.3 Load-dependent deterioration processes

3.3.1 Introduction

As discussed in Section 3.1, the aim of the dynamic adaptation of maintenance strategies sub-module is to modify the production orders applied to the machine so that it can operate in the most suitable load conditions. This leads to the requirement of the development of the deterioration models based on the stochastic processes that are able to take into account such changes in the loads. In this section, we suppose that the loads have temporal impacts on the deterioration processes. Specifically, the rate or the variance of the deterioration process is assumed to be dependent on the operation conditions [64]. To

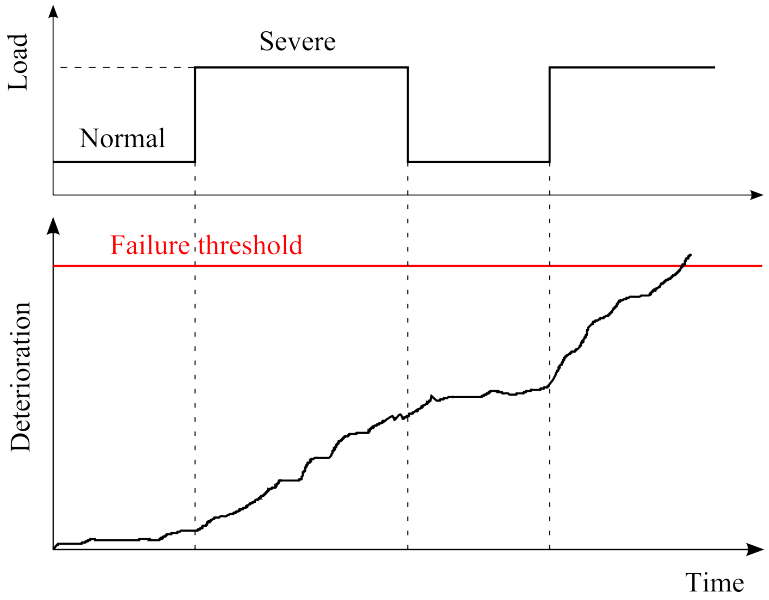


Figure 3.9: Example of load-dependent deterioration process

model such phenomenon, a simple and efficient solution is to consider the parameters of the deterioration process as a function of the machine's loads.

An example of the load-dependent deterioration process is illustrated in Figure 3.9

In this figure, two production orders corresponding respectively to the "normal" and "severe" load conditions are considered. Under the "normal" load, the deterioration process evolves more slowly than under the "severe" one. In real systems, the machine often operates in a dynamic environment, leading to the random changes in operating conditions. However, in the framework of the SUPREME project, the production orders are optimized beforehand. Therefore, the loads can be considered to be deterministic in this case. From the RUL estimation point of view, this helps to reduce the variance in the RUL results.

In the test bench case, the Wiener process has been selected to model the deterioration process due to the non-monotone in temporal propagation of health indicators. Nevertheless, in several practical situations, the deterioration could be a monotonic and gradual deterioration process. For modeling such deterioration phenomena, the Gamma process is the most appropriate one [145]. Therefore, we assume in this section that the deterioration of the equipment of interest follows a Gamma process. For the sake of simplicity, the process is supposed to be homogeneous with the shape parameter $\alpha(t) = \alpha \cdot t$ and the scale parameter β. If we denote α_1, β_1 and α_2, β_2 the couples of parameters of the Gamma process corresponding respectively to the normal and severe conditions of load, we have $\alpha_1 \cdot \beta_1 < \alpha_2 \cdot \beta_2$, which implies that under the normal load condition the deterioration rate is smaller than under the severe one.

3.3.2 RUL estimation

Consider the normal and severe loads in the above example and suppose that the parameters of the Gamma process under these two load conditions can be perfectly estimated from the available training data. Given the load evolution in the future, the RUL of the equipment can be estimated as follows.

Denote t_0 the instant at which we want to estimate the RUL and $x_0 = x(t_0)$ the deterioration level of the equipment at that time. The equipment is assumed to be failed once its deterioration level reaches for the first time a predetermined threshold L. To estimate the RUL, it is necessary to calculate firstly the probability density function (pdf) of the total increments of deterioration from the initial instant t_0. Given the information up to t, the RUL shares the same distribution with the conditional reliability:

$$R(t \mid x_0) = \mathbb{P}(x_t - x_0 < L - x_0) = \int_0^{L-x_0} f(x)\, dx \tag{3.10}$$

where $f(x)$ is the pdf of the total increments in the period $[t_0, t]$.

By knowing beforehand the loads dynamic, we know how many times the load will change as well as its state after each transition. Suppose that the load will change n times and denote d_i, $i = 1, 2, \ldots, n$ the time at which the nth change occurs. We can divide therefore the interval $[t_0, t]_{t>t_0}$ into $(n+1)$ sub-intervals $[t_0, d_1]$, $[d_1, d_2]$,...,$[d_{n-1}, d_n]$, $[d_n, t]$. Let $\alpha^{(i)}$ and $\beta^{(i)}$ denote the parameters of the Gamma process corresponding to the load state in the ith interval. Moreover, denote u_i the total increments in the ith interval, the pdf of the total increments can be expressed as the convolution of the increments in different regimes of loads:

$$\begin{aligned} f(x) = (f_1 * f_2 * \ldots * f_{n+1})(x) = \int_0^{L-x_0} \int_0^{L-x_0-u_1} \cdots \int_0^{L-x_0-u_1-\ldots-u_n} & f_{n+1} \\ & \left(L - x_0 - u_1 - \ldots - u_n, \alpha^{(n+1)} \cdot (t - d_n), \beta^{(n+1)}\right) \cdot \\ & f_n\left(u_n, \alpha^{(n)} \cdot (d_n - d_{n-1}), \beta^{(n)}\right) \cdot \\ & \qquad \vdots \\ & f_2\left(u_2, \alpha^{(2)} \cdot (d_2 - d_1), \beta^{(2)}\right) \cdot \\ & f_1\left(u_1, \alpha^{(1)} \cdot (d_1 - t_0), \beta^{(1)}\right) \cdot \\ & \mathrm{d}u_{n+1}\mathrm{d}u_n \ldots \mathrm{d}u_2\mathrm{d}u_1 \end{aligned}$$

where $f_1, f_2, \ldots, f_{n+1}$ are the pdfs of the corresponding increments; $f(x, \alpha, \beta)$ the density function with respect to the variable x of a Gamma distribution with two parameters α and β and $*$ denote the convolution product.

The above expression of the conditional reliability is, however, too complex to be calculated even by numerical techniques. In this case, the Monte Carlo method can be adapted to overcome the difficulty. In effect, given the current deterioration level and

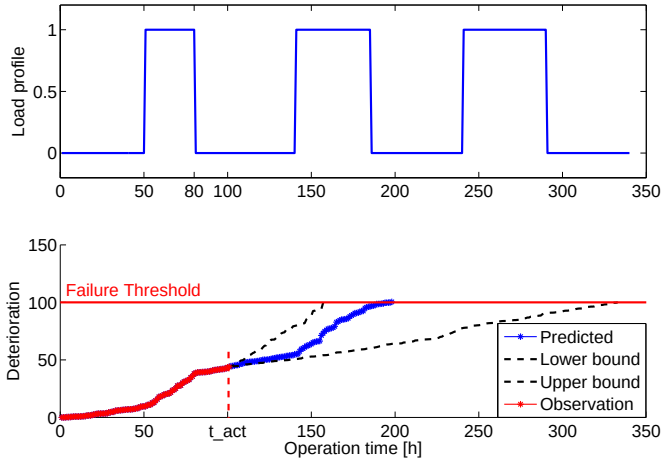


Figure 3.10: Prediction of deterioration evolution under different load

the future load changes, the temporal deterioration evolution can be simulated. Once it reaches for the first time the failure threshold, a RUL value can be determined. Figure 3.10 illustrates the deterioration prediction given the load profile.

In this figure, the load profile value 0 signifies the normal condition and the value 1 represents the severe one. The blue curve at the bottom figure represents a predicted evolution in the future of the deterioration indicator. In addition, the upper and the lower bounds of the RUL represented by the black curves are also predicted by considering the worst case and best case scenarios. In these scenarios, the load are supposed to be unchanged (i.e. it remains either in the normal or in the severe condition) for the prediction horizon.

By repeating this simulation a large number of times (i.e. 10000 times), the histogram of the RUL is obtained and shown in Figure 3.11.

Regarding the boundaries of the RUL, since the load do not change under the worst case of best case scenario, the RUL probability density function can be analytically computed as the case of the homogeneous Gamma process presented in Section 2.2.2. These boundaries are represented by the black and red curves in Figure 3.11. It can be noticed that the histogram of the RUL under varied load scenario is the mixture of the ones calculated under the two worst and best cases.

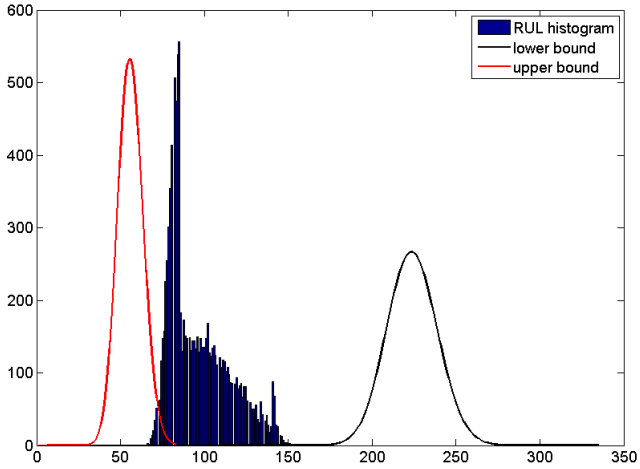


Figure 3.11: RUL estimation result

3.4 Information exchange between SUPREME partners

3.4.1 Database access

The previous sections demonstrate the RUL estimation for the different cases: one for the main bearing of the test-bench at CETIM and the other for the case of component whose deterioration process is load-dependent. In the framework of the SUPREME project, for practical implementation purpose of the deterioration modeling as well as the estimation of the RUL, the RUL calculation conducted at Grenoble-INP can be triggered in two modes: manual or automatic. In the first mode, the RUL is calculated once Grenoble-INP receives the demand from the dynamic adaptation of maintenance strategy sub-module (i.e. from IPA) while in the second one, the RUL is calculated periodically, i.e. after every 6 hours.

Figure 3.12 illustrates an installation scheme of the sub-modules of the WP4 in the framework of the SUPREME project. For the applications of the predictive maintenance tools on a paper machine, the ECMS developed in the WP3 is implemented on site at the paper mill located in Condat, France. Once the RUL prediction is triggered, the deterioration model sub-module implemented by Grenoble-INP must access to this database in order to acquire the data relevant to the component of interest. However, this database was inaccessible from outside of the paper mill due to some IT difficulties. One solution to provide required data for the WP4 was given by ECS: all the database is duplicated and stored in another server located at ECS in Krakow, Poland so that this server can be accessed by all the others partners of the project.

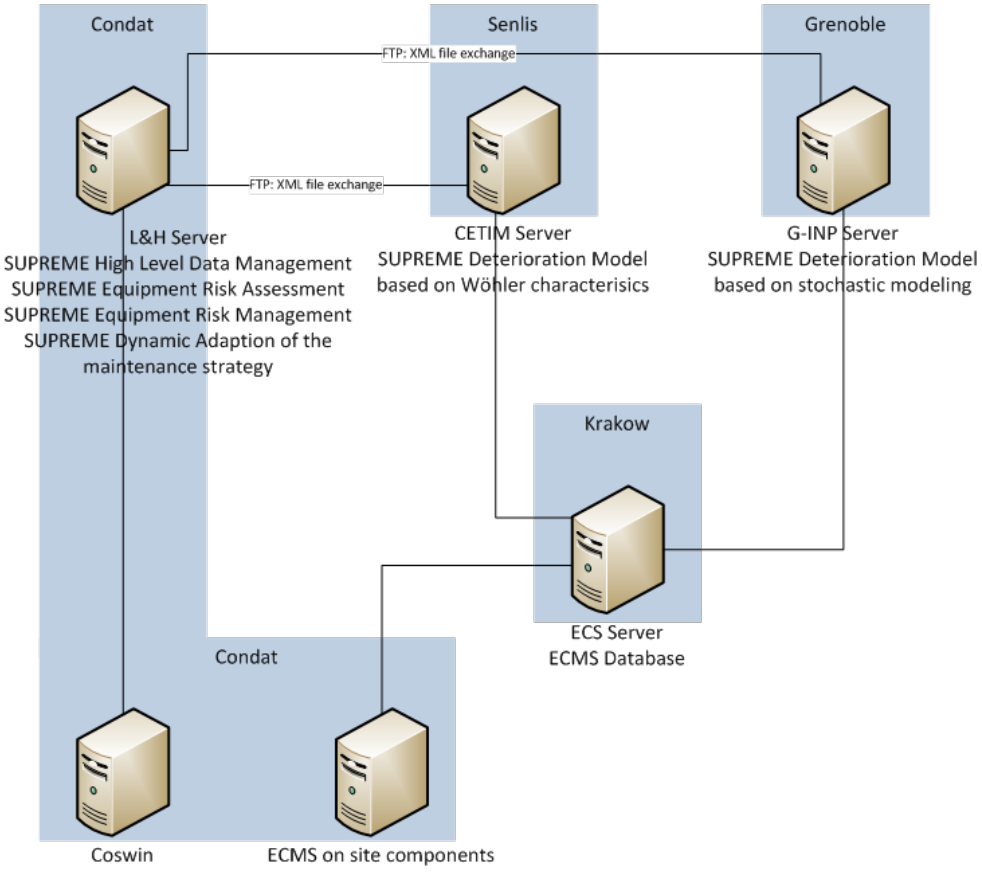


Figure 3.12: Installation scheme of the SUPREME R&M Module

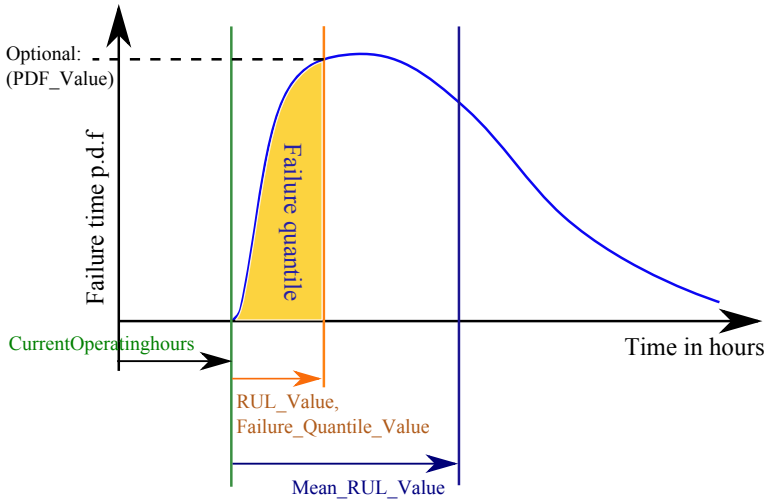


Figure 3.13: RUL exchange format

3.4.2 RUL information exchange

After having been estimated, the RUL, i.e. in the format of a probability density function, needs to be sent to the sub-module "dynamic adaptation of maintenance strategies". In other words, it needs to be sent to IPA partner located in Germany. For this end, one solution is to define the extensible markup language (xml) files that contain all the RUL information and then put them on an ftp (File Transfer Protocol) server in common for all the WP4 partners (c.f. Figure 3.12).

Figure 3.13 represents RUL format with different elements that are contained within an xml file for the data exchange.

An .xml file is structured by a root element containing the identified information of the requested component such as name, component ID, deterioration model, etc. and several RUL nodes in which each node represents RUL information at one timestamp and contain the following elements:

- RUL_ID: each RUL node should have a unique ID for the identification purpose.
- Date time stamp: The moment that the RUL is required to be calculated.
- Current Operating hours: The number of operating hours of the component until the requested time.
- Mean RUL: The mean value of the RUL
- RUL_quantile nodes: A simple way to exchange the distribution of the RUL is to store it by several different couples of failure quantile values and corresponding RUL values. As shown in Figure 3.13, a RUL_Value is a value that indicates the remaining time before a component will fail with a certain probability which is indicated by a Failure_Quantile_value.

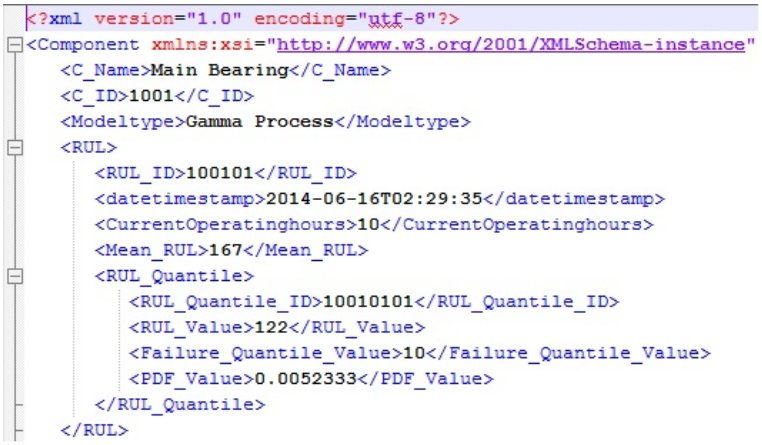


Figure 3.14: An example of the exported .xml file

To assure that the xml file could be used by the WP4 partners, an xsd (XML Schema Definition) file has been provided by IPA. This xsd file is an xml schema definition defined by IPA in order to unify the exchanged file format. Every xml files after having been exported must be validated by the software Notepad++ using this schema definition. Figure 3.14 represents an example of the exported xml file with all required RUL information.

Finally, all the exported .xml files are put on the ftp server at Cetim in Senlis, France that is accessible for all the WP4 partners.

3.5 Simulation Software Environment Test

To facilitate the applications as well as the development of the deterioration models and its associated RUL estimation methods, a simulation software environment test has been developed and implemented (c.f. Figure 3.15).

This environment is integrated an user graphic interface, which helps to facilitate the required steps and unifies all the developments under a unique environment. In the main function panel, there are five functions which correspond to different tests and developments that were carried out in the framework of the SUPREME project. The first one, namely “load-independent test” corresponds to the trivial case where the deterioration is modeled by the stochastic processes that are independent to the operating conditions. The “multiple deterioration modes” function is reserved for the development of the novel deterioration models beyond the state-of-the-art that will be presented in the next chapters of this manuscript. The three remaining functions are the ones that have been applied and presented in this chapter. Once choosing a test, a corresponding supplementary window will appear and supply all the necessary tools to carry out the test and validation.

For the visualization purpose, two figures are integrated in the software: The figure on the top represents the deterioration indicators (for training as well as for testing purpose).

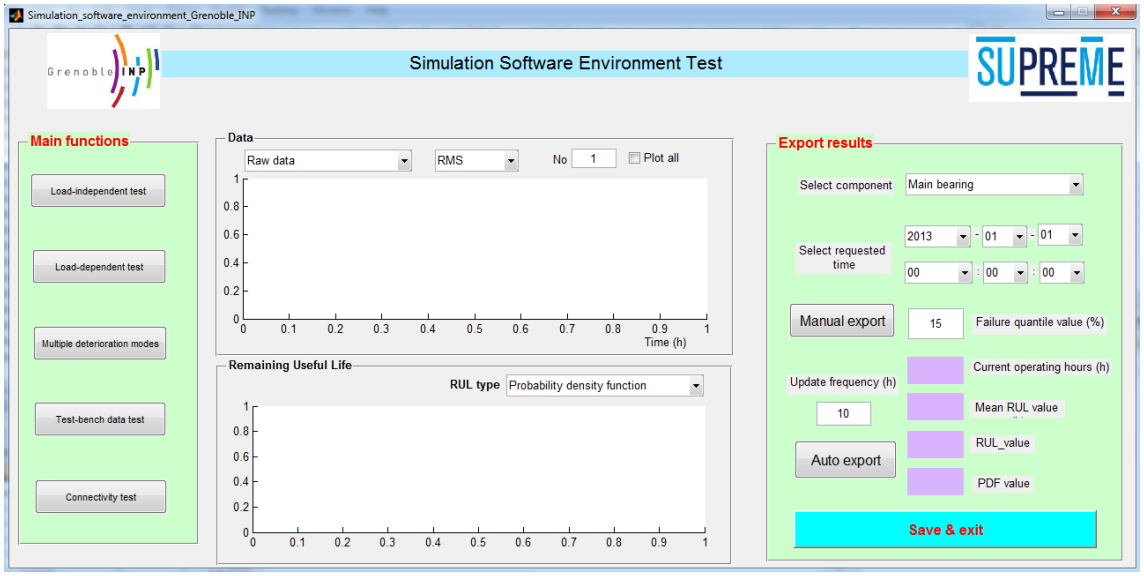


Figure 3.15: Simulation Software Environment Test developed at Grenoble INP

The second figure in the bottom is used to represent the estimated RUL results. There are two options of displaying the RUL results: The probability density function (pdf) or the cumulative distribution function (cdf).

After having calculated the RUL, the results can be exported in an xml file format for the information exchange purpose between the WP4 partners by using functions provided in the "export results" panel.

3.6 Chapter summary and conclusion

This chapter presented the applications of the state-of-the-art deterioration models and the associated RUL estimation methods as well as the works realized in the framework of the SUPREME project. Firstly, the incubation/propagation model has been applied for the vibration data acquired from the test-bench at CETIM. The numerical results showed that the learned model is able to represent the deterioration dynamic of the main bearing in a real situation.

Furthermore, a load-dependent study was also carried out in order to adapt to the case that the production order could be changed according to the output of the dynamic adaptation of maintenance strategy sub-module. Under the varied loads, the RUL of equipment was estimated thanks to the Monte Carlo method. The numerical results showed that the final RUL is the mixture of the ones estimated under the worst and the best cases. Finally, a practical installation scheme for the data exchange between the SUPREME partners was introduced.

These works terminate the first stage of this thesis. The corresponding results have

been reported in the two deliverables: D4.2 “Development of the new Required Competencies, Methodologies and Tools” [55] and D4.3 “Realisation of SUPREME Reliability and Maintainability Modules in Living Labs” [56] of the SUPREME project.

Part II

Beyond the state-of-the-art: multi-branch modeling for deterioration and prognostics

Introduction of the second part

An application of the state-of-the-art deterioration models to the test-bench data presented in the last chapter shows very promising results both in terms of deterioration modeling as well as remaining useful life estimation. However, the deterioration phenomena of practical systems such as of a paper machine could be more complex than the ones found in a test environment. Moreover, in order to validate a deterioration model, from a statistical point of view, it is necessary to have enough data for training the model, i.e. to learn its parameters. It should be noted that the data considered here is the one that represents the deterioration evolution from the beginning to the end of life of the system.

Unfortunately, such sources of data is not available at the beginning of the project. One reason is that for a quite big machine like the paper machine, it is not possible to let it deteriorating until a completed failure to obtain a full sequence of deterioration data. For this reason, it is required in the second stage of this thesis, to develop some advanced deterioration models and the associated RUL estimation methods in order to have some "ready-to-use" tool for adapting to the SUPREME project cases once the deterioration data is available. This part is dedicated to accomplish this requirement.

We are interested particularly in the co-existence problem of several deterioration modes. Indeed, in practice, multiple deterioration modes could compete and co-exist even within one component. To take into account such phenomenon, the concept of "multi-branch" model is introduced in this part. Depending on the nature of deterioration states, i.e. discrete or continuous, two classes of the multi-branches models are studied. The discrete cases are dealt in the two chapters 4 and 5 while Chapter 6 is dedicated to the continuous one.

Specifically, based on the Markov modeling method, the so-called multi-branch Hidden Markov Model (Mb-HMM) is firstly presented in Chapter 4. Through numerical studies, we show that the multi-branch model concept can give better performances in RUL estimation compared to the one obtained by standard "mono-branch" HMM model. Chapter 5 extends the Mb-HMM to the multi-branch Hidden semi-Markov Model (Mb-HsMM) in order to overcome state sojourn time problems due to the Markovian property that are inherent to the Markov-based model.

However, both the Mb-HMM and the Mb-HsMM models assume that the deterioration modes are exclusive once having initiated. Chapter 6 overcomes this limitation and applies the multi-branch model concept to the continuous-state case with the implementation of the Jump Markov Linear Systems (JMLS). The difficulties in model training are also addressed in this chapter.

Chapter 4

Multi-branch Hidden Markov Model

4.1 Introduction

This chapter is dedicated to develop novel deterioration models as well as associated RUL estimation methods that are beyond the state-of-the-art. The main purpose is to be able to adapt to more complicated practical cases that can be found in the framework of the SUPREME project.

Hidden Markov Model (HMM) have been successfully applied in temporal pattern recognition, such as speech, handwriting and gesture recognition [128] thanks to their strong mathematical basis. An HMM consists of two stochastic processes: A Markov chain with a finite number of discrete states describing an underlying state evolution mechanism and an observation process which relates to the state process by some probabilistic links. In the context of predictive maintenance, HMM models equipment's health conditions by several meaningful states, such as "good", "minor defect", "maintenance required" and "failure" and therefore can give easy-to-interpret results for maintenance personnel [136]. The features extracted from condition monitoring data can be represented by the observation process that links to the health states by some probabilistic relationships. For this reason, HMM is being more and more investigated in the recent years to be used as an efficient tool for modeling the deterioration processes as well as for the estimation of the RUL [6, 15, 19, 45, 100, 142, 151].

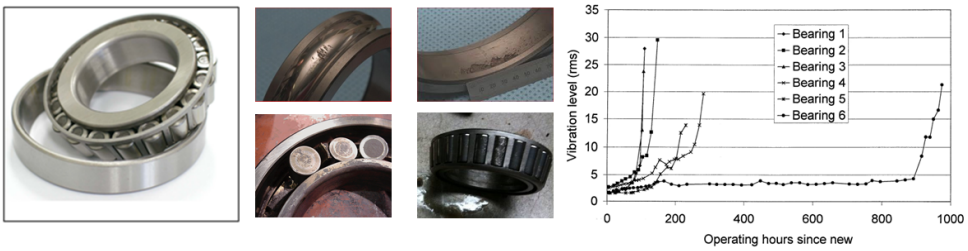


Figure 4.1: Different deterioration rates of the bearing

However, almost all of HMM studies in the literature have dealt only with the mono-mode deterioration case, meaning that the equipment is assumed to degrade following one unique mechanism during its whole life. On the contrary, in many real-life applications, several deterioration mechanisms could co-exist in competition even within a single component. As an illustration, a bearing could deteriorate at different rates depending on the

element that is being defected, such as outer race, inner race, cage, or ball (c.f. Figure 4.1). Another example is the propagation of a fatigue crack: the crack depth is affected by several factors such as loading ratio, environment, micro-structure, geometry, etc [73]. This dependency together with the dynamic of the operational conditions could lead to different propagation rates of the crack.

In this chapter, we propose a multi-branch model to deal with the co-existence problem of multiple deterioration modes. The model is based on the HMM and is called the multi-branch Hidden Markov Model (MB-HMM). In the next section, we firstly review some basic theories of the standard HMM model. Then the extension to the MB-HMM model is introduced in Section 4.1.2. Based on the MB-HMM model, a diagnostics and prognostics framework is proposed in Section 4.2. Finally, the performance of the proposed model is evaluated through numerical studies in Section 4.3.

4.1.1 Hidden Markov Model background

Hidden Markov Model is an extension of Markov chains in which the states are "hidden" or cannot be observed directly. Instead, they can be revealed through the observations which are related to the states by some probabilistic links [128]. Figure 4.2 illustrates an example of the HMM model.

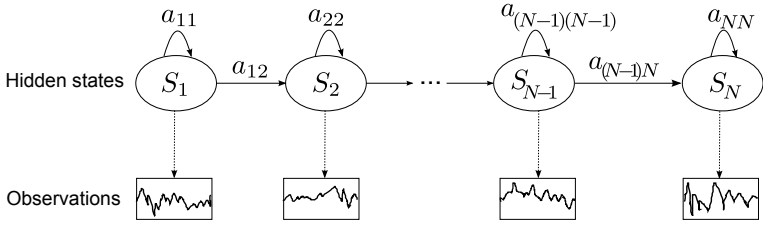


Figure 4.2: A left-right HMM model

With the development of digital technologies, the sampling step plays nowadays an important role in data acquisition system or in data processing procedure [137]. The observations to be modeled by the deterioration models are therefore often sampled at discrete time intervals. For this reason, we only consider the discrete-time case in using the HMM model for deterioration modeling purpose in this chapter. Denote q_t and o_t the hidden state and the observation at time t, $t = 1, 2, \ldots$, a discrete-time HMM is characterized by the following elements [128]:

- A finite set of hidden states, i.e., $S = \{S_1, S_2, \ldots, S_N\}$
- An initial state probability distribution, i.e., $\pi = \{\pi_i\}$ where $\pi_i = \mathbb{P}\{q_1 = S_i\}$, $1 \leq i \leq N$
- A state transition probability matrix, i.e., $A = \{a_{ij}\}$ where $a_{ij} = \mathbb{P}(q_{t+1} = S_j \mid q_t = S_i)$

- An observation model, i.e., $B = \{b_j(\cdot)\}$, where $b_j(o_t) = \mathbb{P}(o_t \mid q_t = S_j)$

In the literature, the compact notation $\lambda = (A, B, \pi)$ is often used to represent the HMM model.

Generally, the observations within the HMM model can be discrete or continuous. The models are called the discrete HMM model (DHMM) or continuous HMM model (CHMM) correspondingly. However, the CHMM model is more suitable for modeling continuous signals which are often found in the framework of industrial condition monitoring [6, 128, 151]. In the CHMM models, the distributions $b_i(\cdot)$ are typically specified using a parametric model family. A common distribution used in the literature is the mixture of Gaussian thanks to its ability of approximating closely any finite continuous density functions [89]. The probability density function of a finite mixture of Gaussian distributions is given by:

$$b_j(\mathbf{x}) = \sum_{k=1}^{K} c_{jk} \mathcal{N}(\mathbf{x}; \mu_{jk}, \Sigma_{jk}) \tag{4.1}$$

where $\mathbf{x}$ is the vector being modeled, c_{jk} is the mixture coefficient for the kth mixture in state S_j and $\mathcal{N}(\mathbf{x}; \mu_{jk}, \Sigma_{jk})$ denotes the probability density function of a Gaussian distribution with mean vector μ_{jk} and covariance matrix Σ_{jk} for the kth mixture in state S_j. The mixture coefficients c_{jk} satisfy the stochastic constraint:

$$\sum_{k=1}^{K} c_{jk} = 1, \qquad 1 \leq j \leq N \tag{4.2}$$

$$c_{jk} \geq 0, \qquad 1 \leq j \leq N, 1 \leq k \leq K \tag{4.3}$$

There are two main types of the HMM models: the ergodic or fully connected model and the left-right or the Bakis model [128]. In deterioration modeling framework, the deterioration processes are often irreversible if no intervention action is carried out. It means that a component cannot come back from current health state to a better one as time passes. The component can only stay in the current health state or move to the next one. In this context, the left-right HMM model appears to be a suitable one for modeling deterioration processes whose severity increases over time (c.f. Figure 4.2). Furthermore, one can assume that in almost of cases, the deterioration level could not jump directly more than δ states, $\delta = 2, 3, \ldots$. The state transition probabilities of the left-right HMM model have the following properties:

$$\begin{aligned} a_{ij} &= 0, \quad && j < i \\ a_{ij} &= 0, && j > i + \delta \\ a_{NN} &= 1, \end{aligned} \tag{4.4}$$

The second equation avoids the jump of more than δ states while the last one implies that the final state S_N is an absorbing state. It means that the model cannot escape

from the final state once having reached it. This state can hence be used to represent the failure state in the deterioration modeling framework. Moreover, since the model always starts from state S_1, the initial state distribution of the left-right HMM model has the property:

$$\pi_i = \begin{cases} 0 & i \neq 1 \\ 1 & i = 1 \end{cases} \tag{4.5}$$

In the subsequent sections of this chapter, we will deal with the continuous "strictly" left-right HMM in modeling the deterioration processes. Strictly means that the system can transit from the current state to only the next one. In other words: $\delta = 1$ in Equation (4.4). In the following of this manuscript, the terminology left-right will be used for implying the strictly left-right topology. To take into account the continuous aspect of the observations, an other compact notation is used to represent the left-right CHMM model:

$$\lambda = (A, C, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi) \tag{4.6}$$

where C, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ correspond to the parameters c_{jk}, $\boldsymbol{\mu}_{jk}$ and $\boldsymbol{\Sigma}_{jk}$ respectively.

4.1.1.1 Three basic problems of HMM model

In order to be used in real-world applications, there are three basic problems associated with the HMM model that must be solved [128]:

1) Evaluation problem

   Given the HMM model λ and a sequence of observations $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$, compute the probability that the observed sequence was produced by the model. This problem can be solved by the Forward-Backward algorithm [128].

2) Decoding problem

   Given the HMM model λ and a sequence of observations $\mathbf{O}$, the decoding problem deals with finding the optimal hidden state sequence $Q = \{q_1, q_2, \ldots, q_T\}$ that have most likely produced the observation sequence. There are several criteria existing for defining the optimality, but the most commonly used criterion is to find the single best state sequence (path) Q that maximizes $P(Q \mid \mathbf{O}, \lambda)$ or equivalently $P(Q, \mathbf{O} \mid \lambda)$. The problem can be solved through the Viterbi algorithm, a technique based on dynamic programming methods [38].

3) Training problem

   Given the observations $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$, find the HMM model $\lambda = (A, C, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi)$ that best describe how the observations come about. The maximum likelihood estimates of the HMM parameters can be computed by the Baum-Welch algorithm [128].

The solutions for these problems are described in detail in Appendix A.

4.1.1.2 Model topology selection

The above three problems can be solved if we already know the topology of the HMM model, i.e. left-right or ergodic, number of hidden states, etc. In some particular applications, such information could be given or determined from the experience or expert knowledge. However, in most of cases, they are not available. The model topology selection becomes hence a crucial issue since a good choice of the model topology leads to a correct modeling of the dynamic behavior of the observations. In deterioration modeling framework, by supposing that the deterioration is an irreversible process, the left-right topology is a suitable one for representing its temporal evolution. The difficult remains to find the optimal number of hidden states N and the number of mixture components M of the observation model.

There exists several criteria that have been used for model selection purpose, such as the "Akaike Information Criterion (AIC)" or the "Bayesian Information Criterion (BIC)" [11]. In general, these two criteria are represented by two terms which account for a compromise between a measure of model fitness and the model complexity. The first term is based on the likelihood of the model while the second one is measured in terms of the number of free parameters and in terms of the number of observations. For example, the BIC is formally defined as:

$$BIC = -2\log\mathcal{L}\left(\hat{\lambda}\right) + k\log(n) \tag{4.7}$$

where $\mathcal{L}\left(\hat{\lambda}\right)$ is the likelihood of the model parameters $\hat{\lambda}$ that are estimated from the maximum likelihood principle, k is the number of free parameters needed to be estimated and n is the number of data used for the estimation. If the model becomes more complex, i.e. it has more parameters, the likelihood will increase but the penalized term $k\log(n)$ will also increase. This helps to avoid the overestimation problem in determining the model structure [11].

Compared to the AIC, the BIC criterion penalizes more strongly the number of parameters or equivalently the complexity of the model [16]. Moreover, the BIC criterion has also been widely used in the literature for topology selection of the HMM models [88, 9]. For this reason, the BIC criterion will be used in this chapter for the selection of the number of states N and the number of mixture components M. Specifically, these parameters can be determined by selecting the HMM model with the smallest BIC value.

In the HMM model context, the number of model parameters p can be calculated by $p = p_s + p_o$ where p_s is the number of parameters corresponding to the hidden states layer and p_o is the one for the observation layer. Specifically, we have: $p_s = (N-1)+(N-1)\cdot N$ where $N-1$ accounts for the prior state probabilities π while $(N-1)\cdot N$ represents the state transition matrix A. Concerning p_o, if the finite mixture of Gaussian assumption is used

for the observation model, we have $p_o = [O \cdot N \cdot K] + \left[\frac{O\cdot(O+1)}{2} \cdot N \cdot K\right] + [(K-1) \cdot N]$ where each term corresponds to the mean vector $\boldsymbol{\mu}$, the covariance matrix $\boldsymbol{\Sigma}$ and the mixture coefficients C respectively and O denotes the dimension of the observations.

It should be noted that in clustering applications, using only one criterion as the BIC cannot allow to determine the two parameters N and K at the same time since the problem may be unidentifiable [52]. In effect, different models having the same number of mixture components, i.e. the same product $N \cdot K$, could be the same likely, i.e. they have the same likelihood values, in generating the observations. In such case, the only thing that make the differences in BIC criterion is the second term that penalizes the complexity of the models.

This problem is, however, solved in case of the left-right HMM model. Indeed, by implying the constraints on the transition matrix, i.e. by Equation (4.4), the likelihood values are different between the models even they have the same number of total mixture components. In addition, from the above calculation of the model parameters, when $N \cdot K$ is fixed, the model complexity penalize terms are different at different values of N. For this reason, in the following, the BIC criterion will be used for topology selection purpose.

4.1.2 Extension to multi-branch HMM model

Thanks to their strong mathematical basis, the HMM models are being more and more investigated as an efficient tool for modeling the deterioration processes as well as for RUL estimation purpose [6, 15, 151]. However, almost all of the works based on the HMM models in the literature have dealt only with the mono-mode case, meaning that only one deterioration mode has been taken into consideration. On the contrary, in real-life applications, systems could deteriorate following different modes depending on several factors such as the environmental conditions, production orders, loads, etc. In order to take into account the coexistence of such different modes, a simple and efficient solution is to use several HMM models in which each one represent one possible mode and then combine them into an unified model. This idea has been introduced and investigated in the handwriting, speech and gesture recognition literature [66, 82, 81, 72, 61, 51, 98]. For example, Iyer *et al.* combined several individual HMM models to form a so-called parallel-path HMM model and applied it for modeling trajectories of the speech. In the work of Lee *et al.*, the compound model is called the "multiple parallel-path HMM" and was applied for online handwriting recognition applications. The name "multi-path model" is also used in the works of Hämäiläinen *et al.* [51] or Lee *et al.* [81]. Nevertheless, this idea has not yet been applied for the diagnostics and prognostics domain. In this section, based on the multiple model idea, we develop the multi-branch Hidden Markov Model (MB-HMM) in order to deal with the coexistence of multiple deterioration modes of systems.

Specifically, the MB-HMM model is constructed from M parallel individual standard

left-right Markov chains (c.f. Figure 4.3). Each chain together with the observations can be considered as a standard left-right HMM model representing a deterioration mode of the system. The observations are assumed to be continuous and are modeled by a finite mixture of Gaussian distribution (c.f. Equation (4.1)).

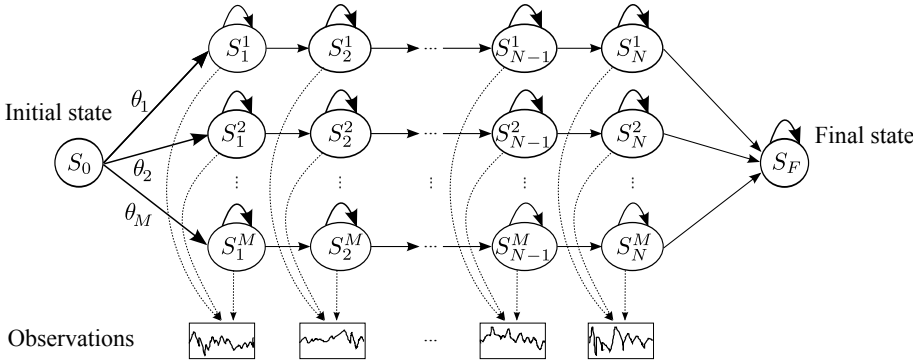


Figure 4.3: Example of left-right MB-HMM model

Apart from the branches, two extra states are added in the developed model: the initial state S_0 and the final state S_F. In the deterioration modeling context, S_0 represents the normal health condition while S_F is the failure state of the system. These states are assumed to be the two non-emitting states, i.e. they do not emit the observations. This assumption is originated from the fact that when the system is in a normal condition, the signals acquired from the CM system are often stationary and do not exhibit any evidences of defect. By supposing that the initiation and appearance of a defect can be perfectly detected thanks to diagnostics task, we are interested only in the observations obtained in the deterioration period. In addition, once entering the failure state, the system is assumed to be stopped immediately. Therefore, no more observations can be acquired and the system will stay in this failure state until it is repaired or replaced by a new one. For this reason, the final state S_F is also considered as an absorbing state.

Starting from the initial state S_0, the model can follow the branch m with a certain probability. Let $\{\theta_m\}_{m=1..M}$ denote the probabilities that the model will follow the mth branch, we have $\sum_{m=1}^{M} \theta_m = 1$. For the sake of simplicity, we further assume that once entering in a deterioration mode, the model cannot jump into another one. It means that branch switching is not allowed in this model. An example of the deterioration modes that are so exclusive is given by maintenance engineers in the SUPREME framework. In effect, they have noticed that when a crack initiates early, then it grows rapidly (such a crack is due to e.g. a defect in the material), and that when a crack initiates later, it grows more slowly (such a crack can result from fatigue).

4.2 MB-HMM based framework for diagnostics and prognostics

Generally, once implementing a mathematical model for representing the dynamic of a system, the first and crucial step is to train the model, i.e. estimate its parameters from the available data set. In the deterioration modeling framework, this step is often conducted "off-line" given the historical deterioration data and all the others condition monitoring information about the population of the equipment of interest. The learned model can then be used in an "online" phase to accomplish both the diagnostics and prognostics tasks.

Similar to the works in [164], based on the MB-HMM model, a two-phase framework for the diagnostics and prognostics is proposed in this section and is shown in Figure 4.4. It consists of the two phases: off-line (training) and on-line .

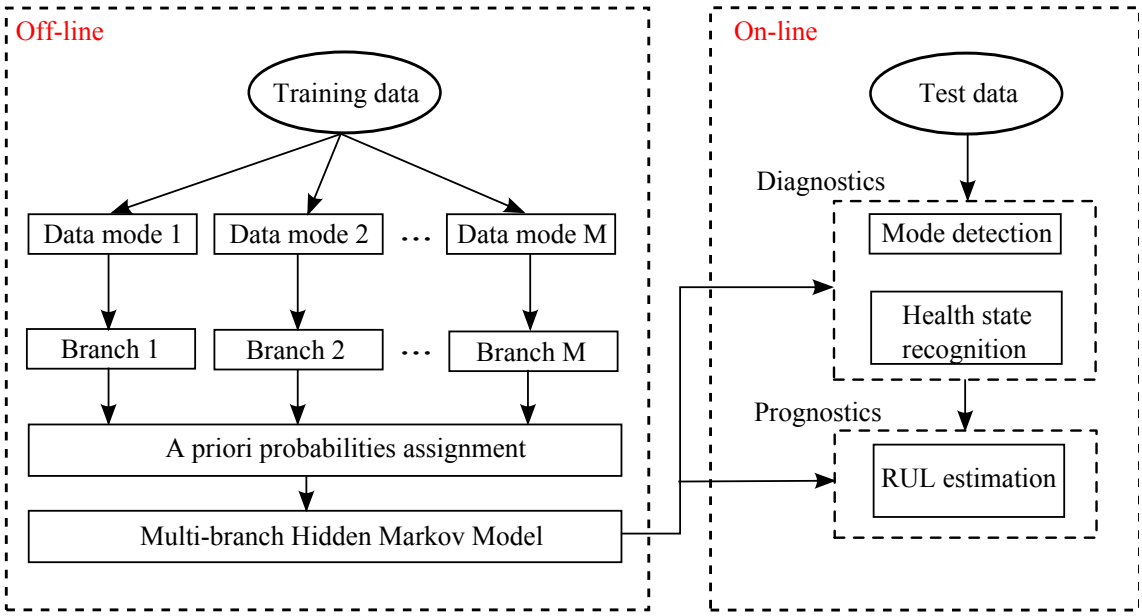


Figure 4.4: Proposed MB-HMM framework for diagnostics and prognostics

4.2.1 Offline phase

The main purpose of the offline phase is to train the MB-HMM model, i.e. to learn the model parameters from the historical deterioration data. The data can be raw signals or extracted features that represent the deterioration states of systems. We assume that enough historical deterioration data are available for training purpose. Since the model has several branches, the idea is to divide the training data set into smaller subsets in which each one corresponds to a deterioration mode and use them to train individually the branches of the model.

A crucial task in training the model is to determine the number of branches M. This task is equivalent to determine the number of deterioration modes existing in the training data set. This task can be trivial if knowledge about the system's deterioration are available. For example, if we have the judgments of engineers or experts on the deterioration processes. In the context of industrial production systems, the knowledge of the deterioration mechanisms may often be available. Therefore, we assume in this chapter that the number of branches M is already known. We further suppose that thanks to such available knowledge, the data can be grouped into M subsets. If this is not the case, clustering analysis techniques such as "k-means" can be applied with $k = M$ for grouping the training data into M "clusters" [11].

The next step is to apply the Baum-Welch algorithm to train each branch from the corresponding data subset. In our model, the system is considered to be failed once it enters the failure state S_F for the first time. For the later RUL estimation purpose, the probability of transition from the final emitting state S_N of each branch to state S_F must also be estimated. To this end, we integrate the state S_F into each branch of the compound model to form an HMM model with a non emitting state as shown in Figure 4.5.

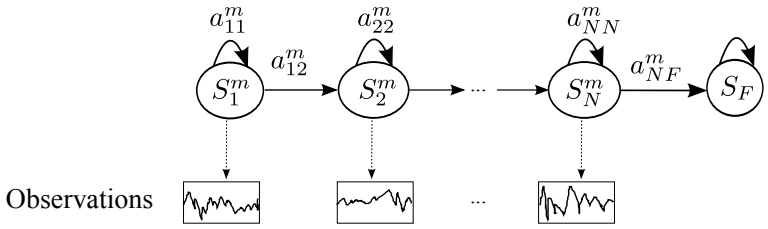


Figure 4.5: HMM with non-emitting state S_F for the mth branch

To apply the standard Baum-Welch algorithm for this model, a slight modification on the state transition matrix, the initial state distribution and the observation model must be carried out [88]. Consider the standard N-state left-right HMM model, its standard transition matrix is given by:

$$A_0 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix} \tag{4.8}$$

To take into account the existence of the non-emitting state S_F, some extra elements corresponding to this state are added and the standard matrix A_0 becomes:

$$A_n = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} & a_{1F} \\ a_{21} & a_{22} & \cdots & a_{2N} & a_{2F} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} & a_{NF} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{4.9}$$

where the final line is due to the fact the S_F is an absorbing state. Note that the stochastic constraint (i.e. $\sum_j a_{ij} = 1 \; \forall i \in \{1, 2, \ldots, N, F\}$) is still hold for this modified matrix.

Similarly, the new initial state distribution is:

$$\pi = \big[\underbrace{1 \quad 0 \quad \ldots \quad 0}_{\text{N elements}} \quad 0\big]^T \tag{4.10}$$

where the final 0 is added corresponding to the state S_F.

Another modification must be considered for the observation model. In effect, in implementing the standard forward-backward procedure, we usually use the standard observation matrix:

$$B_0 = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1T} \\ b_{21} & b_{22} & \cdots & b_{2T} \\ \vdots & \ddots & \vdots & \vdots \\ b_{N1} & b_{N2} & \cdots & b_{NT} \end{bmatrix} \tag{4.11}$$

where $b_{it} = \mathbb{P}(o_t \mid q_t = S_i)$. With the existence of S_F, this matrix becomes:

$$B_n = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1T} & 0 \\ b_{21} & b_{22} & \cdots & b_{2T} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ b_{N1} & b_{N2} & \cdots & b_{NT} & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{4.12}$$

This modification is equivalent to assigning a virtual observation o_{T+1} to state S_F at the end of observation sequence of length T and disallowing any other feature at that state (i.e. the N first elements in the last line of the matrix are set to 0). In case the mixture Gaussian distribution is used as the observation model, an extra observation matrix B_2 of dimension $N \times K \times T$ where $B_2(i, k, t) = \mathbb{P}(o_t \mid q_t = S_i, K_t = k)$ is often used when applying the Baum-welch algorithm [129]. Here K_t denotes the mixture component at time t. A similar modification can also be done for this matrix by adding the corresponding elements for the state S_F.

With the above modifications, the standard Baum-Welch algorithm can be applied as usual for estimating the parameters of all the branches of the MB-HMM model. The remaining parameters that need to be estimated are the branch probabilities $\theta_1, \theta_2, \ldots, \theta_M$. Suppose that there are D_m observation sequences in the mth training data subset, the *a priori* probability of the branch m can be computed by:

$$\theta_m = \mathbb{P}(\lambda_m) = \frac{D_m}{\sum_{m=1}^{M} D_m} \tag{4.13}$$

where λ_m is the HMM model corresponding to the mth branch.

4.2.2 Online phase

In the online phase, the model learned from the off-line phase is used for the implementation of the diagnostics and the prognostics tasks for the monitored equipment. The deterioration data is assumed to be acquired "online" from the condition monitoring system and is used as test data. The monitored equipment is assumed to be identical and operate under the same conditions as the ones in the history. In the case of co-existence of several deterioration modes, an important diagnostics task is to detect the actual mode that the equipment is following. Given the detected mode, another important diagnostics task is to assess the current health state of the equipment. Based on the diagnostics results, the prognostics can follows for estimating the remaining useful life (RUL) of the equipment.

In the framework of the multi-branch modeling, each constitute branch is used to represents a deterioration mode. For this reason, in the followings, the two terminologies branch and mode are used interchangeable: the system is said to be in branch k is equivalent to the system is following the deterioration mode k.

4.2.2.1 Deterioration mode detection

The actual deterioration mode can be determined as the one that has the maximum posterior probability given the test data sequence [11]. That is:

$$\hat{m} = \arg\max_{m} \mathbb{P}\left(\lambda_m \mid \mathbf{O}\right) \tag{4.14}$$

where $\mathbf{O} = \{o_1, o_2, \ldots, o_t\}$ is the observation sequence until the current time t and the posterior probability of the mode m is calculated via the Bayes' theorem:

$$\mathbb{P}\left(\lambda_m \mid \mathbf{O}\right) = \frac{\mathbb{P}\left(\mathbf{O} \mid \lambda_m\right)\mathbb{P}\left(\lambda_m\right)}{\sum_{m=1}^{M}\mathbb{P}\left(\mathbf{O} \mid \lambda_m\right)\mathbb{P}\left(\lambda_m\right)} \tag{4.15}$$

where $\mathbb{P}\left(\mathbf{O} \mid \lambda_m\right)$ is the likelihood of the HMM model λ_m that corresponds to the mth branch with respect to the observation sequence $\mathbf{O}$ and can be computed via the forward-backward procedure (see Appendix A); $\mathbb{P}\left(\lambda_m\right)$ is the *a priori* probability of the branch m and is calculated by Equation (4.13).

4.2.2.2 Health state assessment

Once the deterioration mode has been detected, the next task is to assess the health state of the equipment. Given the sequence of test data $\mathbf{O}$, this task can be accomplished in two steps: Firstly, determine the most likely sequence of hidden states thanks to the Viterbi algorithm and then consider the last one in the estimated sequence as the current health state of the equipment. A detailed explication of the Viterbi algorithm can be

found in [128] or in Appendix A.

Specifically, suppose that the equipment is currently in the mode $\hat{m}$ and let $Q_{\hat{m}}$ denote a possible path of states under the mode $\hat{m}$. The best sequence of the states is given by:

$$Q^* = \arg\max_{Q_{\hat{m}}} \mathbb{P}\left(Q_{\hat{m}} \mid \mathbf{O}, \lambda_{\hat{m}}\right)$$

The last state in the sequence Q^* is the current health state of the equipment.

4.2.2.3 RUL estimation

Since the constituent branches of the MB-HMM model are exclusive, the equipment can only follow one branch for reaching the failure state S_F. In other words, if the equipment is currently staying in state S_i of the mth branch, the RUL can be defined as the time needed to pass the states $S_{i+1}, S_{i+2}, \ldots, S_N$ of that branch and hence can be computed in the same manner as in the standard HMM model case.

Consider an HMM model with a non-emitting state in Figure 4.5 and suppose that the equipment is currently in the state S_i. Under the discrete time context, the model is suppose to realize a state transition (including self-state one) after every one time step. In this case, the RUL can be defined as the number of transition steps to reach the final state S_F for the first time from the current one:

$$\text{RUL} = min\left\{n \geq 0 : q_{t+n} = S_F \mid q_t = S_i\right\} \tag{4.16}$$

where q_t denotes the system state at the time t.

By this definition, the RUL can be considered as a discrete variable. Its probability mass function (pmf) is given by:

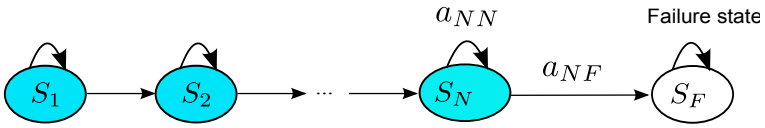
$$\mathbb{P}\left(RUL = n \mid q_t = S_i\right) = \mathbb{P}\left(q_{t+n} = S_F, q_{t+n-1} \neq S_F, \ldots, q_{t+1} \neq S_F \mid q_t = S_i\right) \tag{4.17}$$

That is, the probability that the system fails in n transition steps given that it is in state S_i at time t is the probability that it is in state S_F for the first time at time $t + n$.

Denote $h_i^{(n)} = \mathbb{P}\left(RUL = n \mid q_t = S_i\right)$. The pmf of the RUL can be computed by the following backward recursive equations:

- At state S_N:

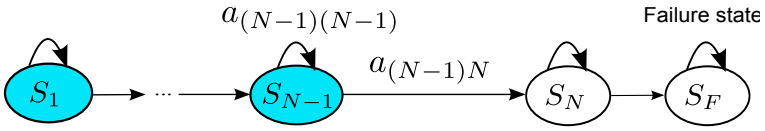
  Since at the next time step $t + 1$, the system can either staying in the state S_N or jump to the failure state S_F, the probability that the RUL equal 1 is exactly the probability of transition from state S_N to state S_F a_{NF}. The probability that RUL equal to n with $n > 1$ equal to probability that the system will stay in state S_N for more $n - 1$ time steps and then jump to state S_F.



We obtain:

$$\begin{aligned} h_N^{(1)} &= a_{NF} \\ h_N^{(2)} &= a_{NN} h_N^{(1)} \\ &\cdots \\ h_N^{(n)} &= a_{NN} h_N^{(n-1)} \end{aligned}$$

- At state S_{N-1}:

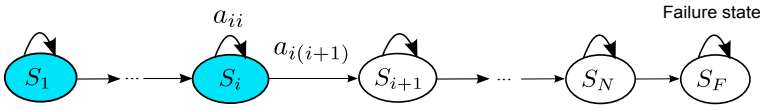


If the system is at state S_{N-1}, it can only jump to state S_N or stay at state S_{N-1} after one time step. Similar to the above calculation, we have:

$$\begin{aligned} h_{N-1}^{(1)} &= a_{(N-1)F} \\ h_{N-1}^{(2)} &= a_{(N-1)(N-1)} h_{N-1}^{(1)} + a_{(N-1)N} h_N^{(1)} \\ &\cdots \\ h_{N-1}^{(n)} &= a_{(N-1)(N-1)} h_{N-1}^{(n-1)} + a_{(N-1)N} h_N^{(n-1)} \end{aligned}$$

.....

- At state S_i By the same principle:



$$\begin{aligned} h_i^{(1)} &= a_{iF} \\ h_i^{(2)} &= a_{ii} h_i^{(1)} + a_{i(i+1)} h_{i+1}^{(1)} \\ &\cdots \\ h_i^{(n)} &= a_{ii} h_i^{(n-1)} + a_{i(i+1)} h_{i+1}^{(n-1)} \end{aligned}$$

where a_{ij} is the transition probability from state S_i to state S_j.

It should be noted that the above calculation is carried out for a given branch of the MB-HMM model. However, if the deterioration mode is wrong detected, especially

at the beginning of a defect propagation when the observations are insufficient to show an obvious trend, the RUL could be estimated with a very large bias. To take into consideration the uncertainties in mode detection, the Bayesian Model Averaging (BMA) technique [60] can be implemented. Specifically, the final RUL distribution is computed as an average of the RULs estimated for all branches, weighted by their posterior model probability:

$$\mathbb{P}\left(\text{RUL} \mid \mathbf{O}\right) = \sum_{m=1}^{M} \mathbb{P}\left(\text{RUL} \mid \lambda_m, \mathbf{O}\right) \mathbb{P}\left(\lambda_m \mid \mathbf{O}\right) \tag{4.18}$$

where $\mathbb{P}\left(\lambda_m \mid \mathbf{O}\right)$ is calculated from Equation (4.15).

4.3 Numerical results

In order to evaluate the performances of the proposed model, we investigate in this section two studies corresponding to two deterioration cases. In the first one, the health states of equipment are discrete and the observations are "artificially" generated from an MB-HMM model. The aim is to demonstrate the exactitude of the diagnostics and prognostics framework proposed in the last Section. In the second study, the MB-HMM model is used for approximating and representing the evolution of continuous health states, i.e. the depth of a crack appeared within a bearing. We show that by taking into account the co-existence of multiple modes of deterioration, the MB-HMM model can outperform the standard HMM, which is "mono-branch", in the estimation of the RUL.

4.3.1 Simulation study with synthetic data

4.3.1.1 Data generation

Motivated from engineering observations discussed in Section 4.1.2 about the difference in propagation rates of a crack related to its different apparition time, we consider in this study two deterioration rates of the crack: slow (mode 1) and rapid (mode 2). Under each mode, we suppose that the equipment passes through four discrete states before the complete failure. A two-branch HMM model with 4 states within each branch is hence implemented to model the deterioration process of the equipment. Without any *a priori* knowledge about the underlying mode once a crack initiates, the branch probabilities can be chosen to be equal $\theta = [\theta_1; \theta_2] = [0.5; 0.5]$.

We further suppose that two signals X_1, X_2 can be obtained through the condition monitoring system, i.e. in case of vibration monitoring, these two signals can correspond to the vibration signals acquired by two accelerometers located in horizontal and vertical direction on the equipment's house. The observations density are hence modeled by a bivariate Gaussian distribution. For the sake of simplicity, $K = 1$, i.e. one mixture

component is chosen for data generation purpose. Based on the works reported in [84] and [20], the following parameters are chosen:

- State transition probabilities

$$A^1 = \begin{bmatrix} 0.99 & 0.01 & 0 & 0 & 0 \\ 0 & 0.99 & 0.01 & 0 & 0 \\ 0 & 0 & 0.99 & 0.01 & 0 \\ 0 & 0 & 0 & 0.99 & 0.01 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad A^2 = \begin{bmatrix} 0.95 & 0.05 & 0 & 0 & 0 \\ 0 & 0.95 & 0.05 & 0 & 0 \\ 0 & 0 & 0.95 & 0.05 & 0 \\ 0 & 0 & 0 & 0.95 & 0.05 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- Observation model

$$\mu_1^1 = \mu_1^2 = \begin{bmatrix} 20 \\ 20 \end{bmatrix} \quad \mu_2^1 = \mu_2^2 = \begin{bmatrix} 20 \\ 35 \end{bmatrix} \quad \mu_3^1 = \mu_3^2 = \begin{bmatrix} 35 \\ 35 \end{bmatrix} \quad \mu_4^1 = \mu_4^2 = \begin{bmatrix} 35 \\ 20 \end{bmatrix}$$

$$\Sigma_1^1 = \Sigma_1^2 = \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix} \qquad \Sigma_2^1 = \Sigma_2^2 = \begin{bmatrix} 15 & 0 \\ 0 & 15 \end{bmatrix}$$

$$\Sigma_3^1 = \Sigma_3^2 = \begin{bmatrix} 15 & -2 \\ -2 & 15 \end{bmatrix} \qquad \Sigma_4^1 = \Sigma_4^2 = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$

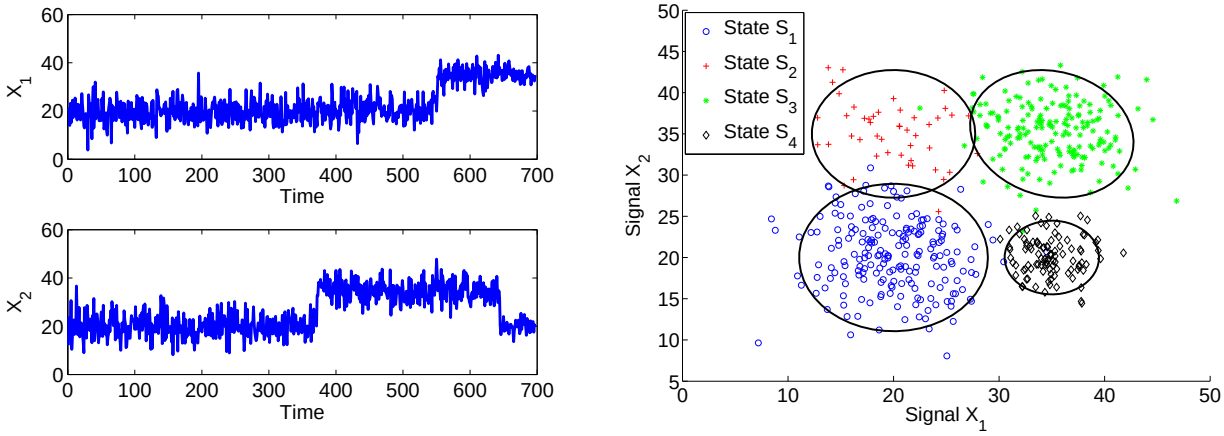
  where the superscript implies the corresponding mode.

Note that we have integrated the non-emitting failure state S_F in the model parameters. Furthermore, the observation models under the two modes are chosen to be the same reflecting the fact that the equipment could goes through the same health states under different deterioration modes. The only difference that distinguishes the two modes is the rate of state transitions. In effect, the self-transition probabilities in the matrix A^1 is greater than the ones in the matrix A^2 means that the deterioration rate under the mode 1 is lower than under the mode 2.

A training data set composed of 20 histories are generated. The choice of deterioration modes is done by randomly sampling from the mode probability vector θ. One example of the observed signals under mode 1 is illustrated in Figure 4.6.

4.3.1.2 Model training

As the model has two different branches, in order to train individually each branch by the Baum-Welch algorithm, the training data is divided into two subsets by applying the k-mean technique with $k = 2$ on the lifetime data. In effect, the equipment will last longer when it is in the mode 1 than when it is in the mode 2. Hence, by classifying the lifetime data into two "clusters", we obtain two data subsets for model training purpose.



(a) Temporal evolution (b) Observations by states

Figure 4.6: Example of training data

Before applying the Baum-welch algorithm for parameter estimation, we must determine the number of hidden states N and the number of mixture components K for each branch. These two quantities are estimated in the same time by using the BIC criterion introduced in Section 4.1.1.2. To this end, we do vary N from 2 to 8 states, K from 1 to 4 and compute the BIC value for each combination of N and K via Equation (4.7).

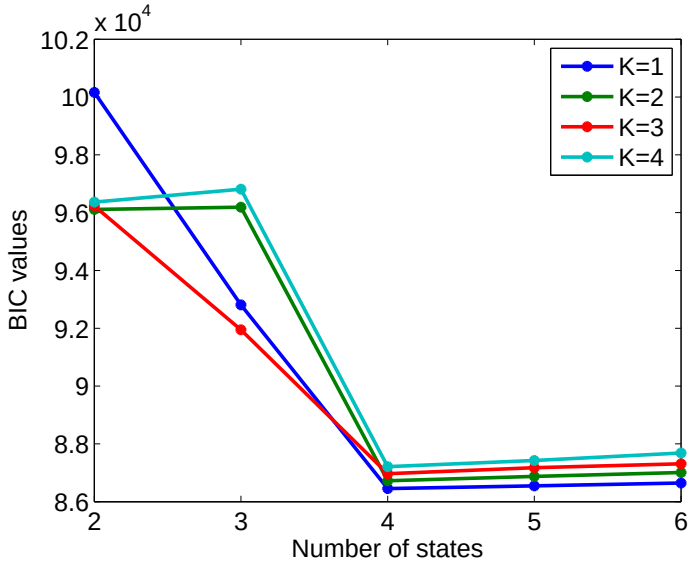


Figure 4.7: BIC values for the branch 1

Figure 4.7 shows the BIC values at different values of N and K for the branch 1. The minimum BIC value can be found at $N = 4$ and $K = 1$ which are the real values used for generating the data. A similar result is obtained for the branch 2.

The last step in training the MB-HMM is to calculate the *a priori* probabilities of the

two modes, i.e. by Equation (4.13).

4.3.1.3 Online phase

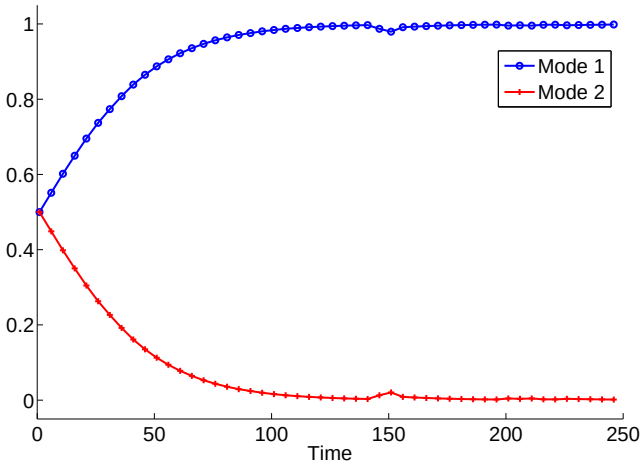


Figure 4.8: Posterior probability for the two branches

After having trained the MB-HMM model from the historical data, we now investigate its performances in accomplishing the diagnostics and prognostics tasks. For this end, we use the true model to generate one more data sequence, consider it as the test data. The deterioration mode is randomly chosen based on the mode probabilities vector θ. Note that we have only the observations and the real states are hidden. Our main goals are to identify the mode under which the equipment is deteriorating, to assess its current health state and to estimate the RUL. According to Equation (4.15), the posterior probabilities of the two modes at different times are calculated. Figure 4.8 shows the estimated result for the case where the real deterioration mode is the mode 1.

At time $t = 0$, the mode probabilities are the prior ones that are $\hat{\theta} = [0.5; 0.5]$. When the time passes, the posterior probability of the mode 1 approaches the value 1 whilst the mode 2 converges to 0. In other words, the mode 1 has been correctly detected.

Once the mode is detected, we apply the Viterbi algorithm to find the "optimal" single state path from which we can assess the current health state. Figure 4.9 shows that the health states of the equipment have been correctly estimated. For example, at time $t = 100$, the equipment is in state S_1 and at time $t = 160$, it is in state S_2.

Regarding the RUL estimation, Figure 4.10 shows the mean values of the estimated RUL and the corresponding 95% prediction interval at different instants. We can see that at some intervals, the RUL values do not change even when time passes. This is due to the inherent Markovian property of the MB-HMM model. In effect, because of the "memoryless" characteristic, the equipment's RUL are still the same when it has just entered in a state or when it has stayed for a while in that state. However, the estimated

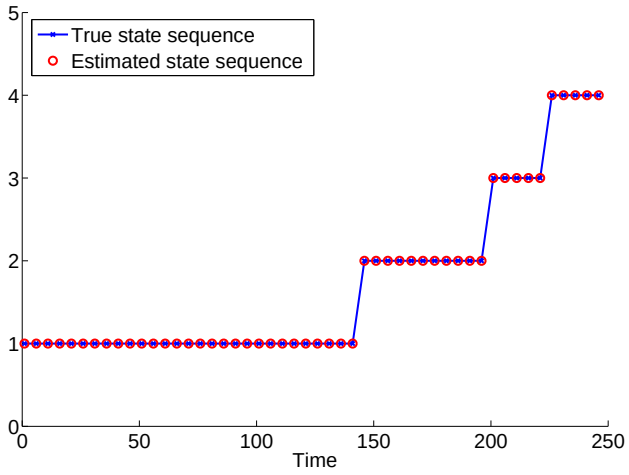


Figure 4.9: Optimal single state path estimated by the Viterbi algorithm

RUL converges to the real one and the prediction interval always covers these real values, which demonstrates the accuracy of the proposed RUL estimation method.

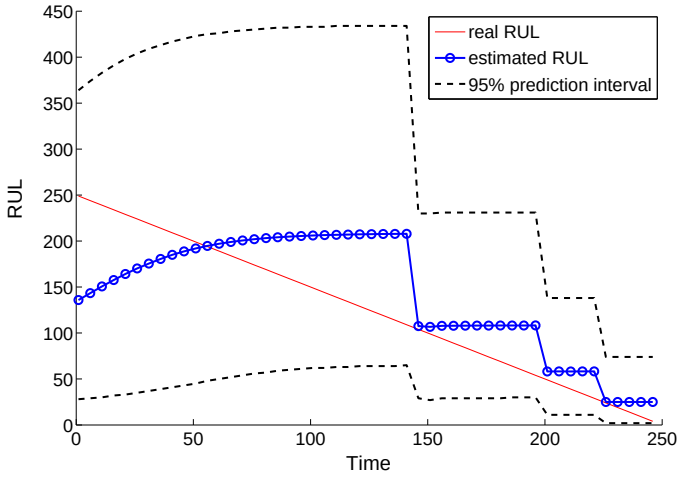


Figure 4.10: RUL estimation at different times

4.3.2 Numerical study with crack length data

In the previous section, we have investigated the performance of the proposed model with the synthetic data generated from an MB-HMM model. In this section, motivated from the effect of crack initiation time to its length propagation rate discussed in Section 4.1.2, the MB-HMM model is applied to represent the temporal evolution of a crack depth. As discussed in Section 2.2.3, a common used physical model in the literature is the one representing the evolution of a crack depth. To take uncertainties into consideration, a

stochastic version of the Fatigue Crack Growth (FCG) model presented in [108] is used to generate the deterioration data. It should be noted that the crack depth evolution is continuous in time, therefore the most natural way to model such process is use continuous stochastic processes such as the Brownian motion or Gamma process. However, by using a discrete-state model like the MB-HMM one to represent a continuous deterioration process, our main purposes are to show that:

- The proposed approach still performs correctly even in "non-favorable" conditions. This demonstrates also the robustness of the proposed model. In other words, we aim to show that the approach based on a discrete-state model is versatile enough to tackle both discrete and continuous deterioration phenomenon.
- The benefit of using a multi-branch model for taking into account the co-existence of several deterioration modes, even if continuous. For this end, a comparative study is given in Section 4.3.3.

4.3.2.1 Fatigue Crack Growth model for data generation

The Fatigue Crack Growth (FCG) model is constructed basing on the popular Paris-Erdogan equation and has been widely used in the literature to describe the evolution of a crack [65, 108]. In order to take into account the stochastic aspects of the crack propagation, we adapt in this study the discretized and randomized version of the FCG model introduced in [108]:

$$x_{t_i} = x_{t_{i-1}} + e^{w_{t_i}} C \left(\beta \sqrt{x_{t_{i-1}}}\right)^n \Delta t \tag{4.19}$$

where x_{t_i} denotes the crack depth at time t_i, C, and n are constants depending on the material property, β is a factor representing the relation between the stress intensity amplitude and the crack depth, w_{t_i} are independent and identically distributed according to a Normal distribution $\mathcal{N}(0, \sigma_w^2)$. By construction we have $0 < x_{t_{i-1}} < x_{t_i}\ \forall i$.

This discretized model allows to determine recursively the crack depth x_{t_i} at time t_i given the previous depth $x_{t_{i-1}}$ and the model parameters. By this way, the propagation data of the crack depth can be generated.

The different modes of deterioration are represented by the rates of propagation of the crack which are assumed to depend on the operating conditions. Specifically, we consider the coefficient β of Equation (4.19) as a function of the operating environment state, denoted by e, as below [65]:

$$\beta(e) = \beta_b \cdot e^{\gamma_e} \tag{4.20}$$

where β_b is the base stress level of system, the values $\gamma_e \geq 0$, $e = 1, 2, \ldots, M$ determine the level of extra stress linking with the state of the environment. Obviously, from Equations (4.19) and (4.20), the propagation rate is proportional to γ_e, i.e. the greater the value of

γ_e is, the more quickly the crack depth evaluates.

In practice, it is difficult to measure directly and accurately the depth of a crack without stopping the machine. For this reason, we suppose that the crack depth is unobservable but we can observe it directly through the measurements. In this study, the measurement is assumed to be the sum of the actual crack depth and a zero-mean Gaussian noise, that is:

$$y_{t_i} = x_{t_i} + \xi_{t_i} \tag{4.21}$$

where $\xi_{t_i} \sim \mathcal{N}\left(0, \sigma_\xi^2\right)$ is the measurement error.

4.3.2.2 Model training and RUL estimation

Similar to the first study, we consider in this section two modes of deterioration, i.e. two rates of propagation of the crack (slow and rapid). These two modes may correspond to the two states of operating conditions: normal and stressed. We choose therefore $\gamma_1 = 0$ for the slow mode and $\gamma_2 = 0.5$ for the rapid one (c.f. Equation (4.19)).

As in the last simulation study, the mode probability is chosen to be equal for the two modes: $\theta = [0.5; 0.5]$. Furthermore, the following parameters are chosen based on the work in [108]: $C = 0.005$, $n = 1.3$, $\beta_b = 1$, $\sigma_w = 1.7$, $\Delta t = 1$, $\sigma_{\xi 1}^2 = 2$, $\sigma_{\xi 2}^2 = 5$ where $\sigma_{\xi 1}$ and $\sigma_{\xi 2}$ are variances of the noise under the mode 1 and mode 2 respectively. The different values of σ_ξ signify that the measurements in the stressed working condition may have slightly more noises than in the normal one. In addition, the equipment is assumed to fail once the crack depth reaches for the first time a critical level $L = 100$.

By using the above FCG model, a set of 50 sequences of crack depths are generated in which there are 22 sequences for mode 1 and 28 for mode 2 (c.f. Figure 4.11).

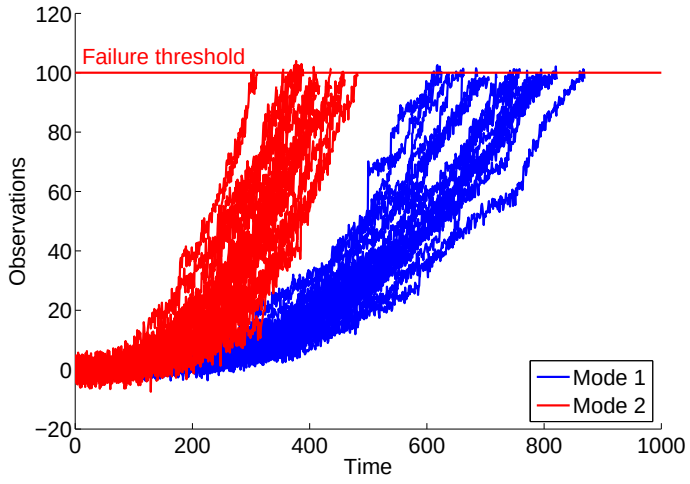


Figure 4.11: Measurements of crack depth in two modes

Suppose that the number of deterioration modes is already known, we adapt a two-branch HMM model to model the evolution of the crack depth. For training purpose, the "k-means" technique is implemented with $k = 2$ with the length data of observation sequences in order to classify them into 2 "clusters". By this way, the training data are grouped in two subsets which are then used to train the model. In this study, the number of states as well as the number of the mixture components for each branch are determined based on the BIC criterion presented in Section 4.1.1.2. In effect, by varying N between 2 and 15 and K between 1 and 4, we obtain the BIC values as shown in Figure 4.12.

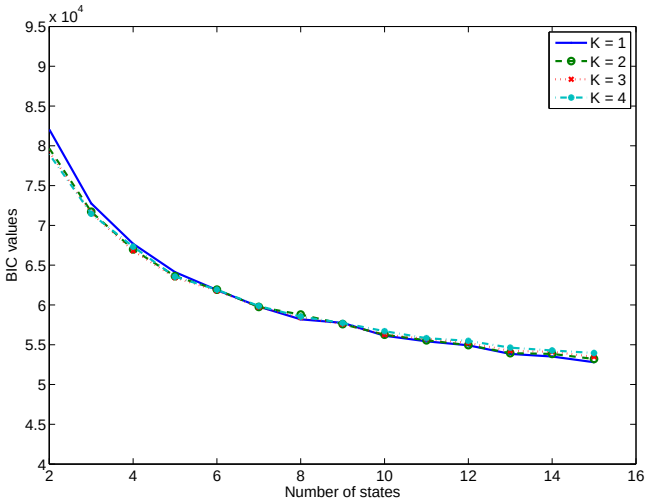


Figure 4.12: BIC for number of state selection

Since we are using a discrete-state model to represent the crack propagation which is a continuous process, the more discrete states we use, the better the approximation of the process. This leads to the decrease of BIC values according to the increase of N, for all the number of mixture component K. For this reason, to avoid overfitting it is sufficient to take the "knee" point on the BIC curve. From Figure 4.12, we choose $N = 10$ and $K = 1$ in this study.

After having determined the model structure, each branch including the non-emitting state S_F is then trained by the Baum-Welch algorithm with a slight modification as presented in 4.2.1. The last step is to calculate the *a priori* probability for each branch by Equation (4.13) which gives the result $\theta_1 = 0.44$ and $\theta_2 = 0.56$.

We are now move to the on-line phase. One more sequence of crack length from the beginning till the failure is generated from the FCG model (c.f. upper sub-figure in Figure 4.13). The actual time chosen for estimating the RUL is $t_{act} = 100h$: only the measured signals till this instant are available. The probability density function (pdf) of the RUL estimated is shown in the lower sub-figure of Figure 4.13.

To demonstrate the on-line RUL estimation problem, we shift gradually the actual time t_{act} by $30h$ towards the failure. After each replacement, we re-estimate the RUL as new observations have been available. The estimation results associated with its 95%

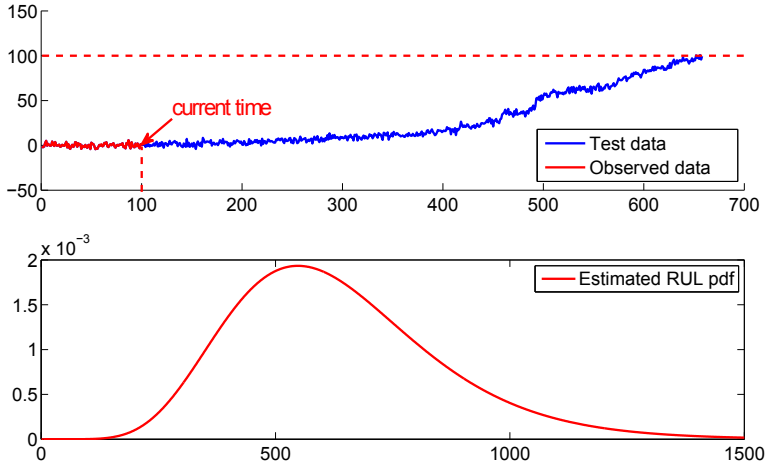


Figure 4.13: Observations and RUL estimation at $t_{act} = 100h$

prediction interval are shown in Figure 4.14. We can see that the estimation has a large bias at the beginning, i.e. before $t = 300h$. This can be explained by the lack of information for having a correct mode detection. The RUL in this period is hence the weighted average of the RULs calculated under the two modes. Nevertheless, just after this period, the measurements show a clearly trend in evolution, the estimated RUL converges hence to the real ones. Furthermore, once the deterioration mode has been correctly detected, the phenomenon of constant RUL due to the Markovian property discussed in Section 4.3.1.3 can be clearly observed. However, the real RUL values always lie within the prediction interval, which demonstrate the exactitude of the proposed RUL estimation method.

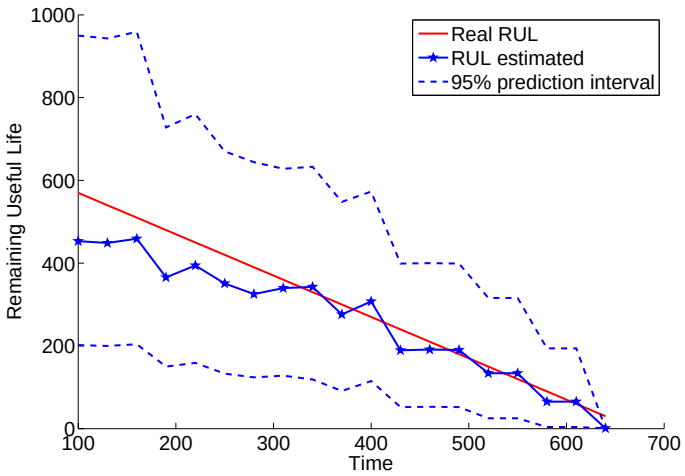


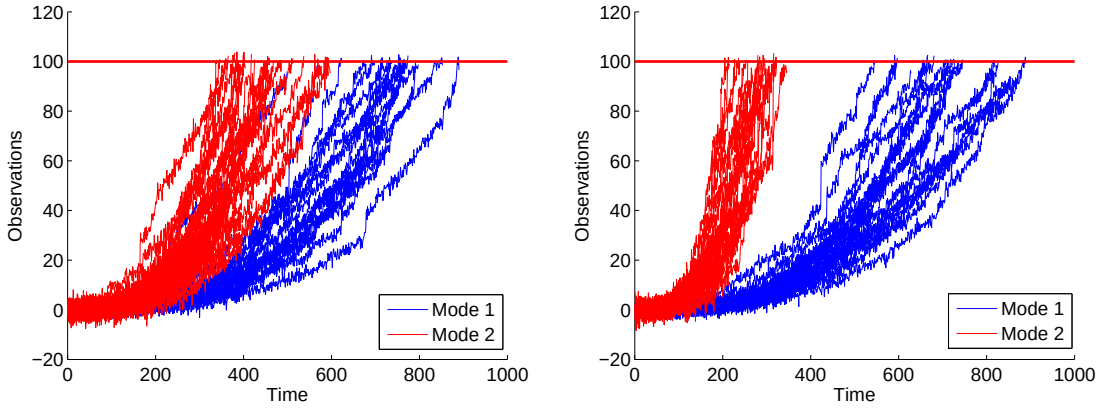
Figure 4.14: RUL estimation at different times

4.3.3 MB-HMM vs HMM

This section is dedicated to evaluate the performances of the MB-HMM model in RUL estimation in comparison with a standard HMM model. We denote the latter model by "average" HMM (AVG-HMM) model to emphasize its one-branch property. Two scenarios are considered: i) with the co-existence of two deterioration modes and ii) with the co-existence of three modes.

For each study, we follow the following steps:

- Firstly, we define the "distance" between the deterioration modes. This distance will help to distinguish one mode from the other ones. The purpose is to investigate the advantages of the multi-branch model in several scenarios where the modes could be very similar or very different. For example, consider the case of FCG data studied in the previous section, the difference in rates of propagation of the cracks could be considered as the distance between the normal and the severe modes. The "distance" in this case can be represented via the coefficients γ_e in the FCG model (c.f. Equation (4.19)). Indeed, by fixing $\gamma_1 = 0$ for the normal condition and varying γ_2 for the severe one, the "distance" between the two modes can be changed as shown in Figure 4.15.



(a) Small distance $\gamma_2 = 0.4$ (b) Large distance $\gamma_2 = 0.8$

Figure 4.15: Two different distances between the two modes

- Corresponding to each mode "distance", the two models MB-HMM and AVG-HMM are trained by the same data set. The learned models are then used for estimating the RUL in the online-phase as described in Section 4.2.2. In this study, 100 test data sequences were generated for the comparison purpose. The mode corresponding to each test data is randomly selected according to the probability vector θ. For each sequence, the RUL is estimated "online": we reestimate the RUL after each $30h$ with the arrival of new observations. The results are evaluated by the root mean

squared error (RMSE) metric, that is:

$$RMSE = \frac{1}{100}\sum_{k=1}^{100}\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\hat{RUL}_i - RUL_i\right)^2} \tag{4.22}$$

where n_k is the number of estimations realized for the kth sequence and $\hat{RUL}_i$ and RUL_i are the estimated and real RUL values at time t_i respectively. The model with smaller $RMSE$ value is the better in estimation of the RUL.

4.3.3.1 Co-existence of two deterioration modes

Figure 4.16 represents the $RMSE$ results for the two models MB-HMM and AVG-HMM at different mode "distances", i.e. at different values of γ_2 in case of two modes co-exist. In this study, $\gamma_2 = 0$ means that there is no distance between the two modes. In other words, there exists only one deterioration mode in the crack data. For training the MB-HMM model in this no distance case, the training data set is randomly divided into 2 groups with the equal number of observations sequences. Each group is then used for training one branch of the multi-branch model. We can realize that if there exists only one deterioration mode, the MB-HMM model give the same RUL estimation performance compared to the one obtained by the average HMM model. Although the MB-HMM has more parameters than the AVG-HMM one, this result can still explained by the fact that each branch of the MB-HMM model is trained separately. By consequence, there is no interest of using the multi-branch model in such mono-mode case.

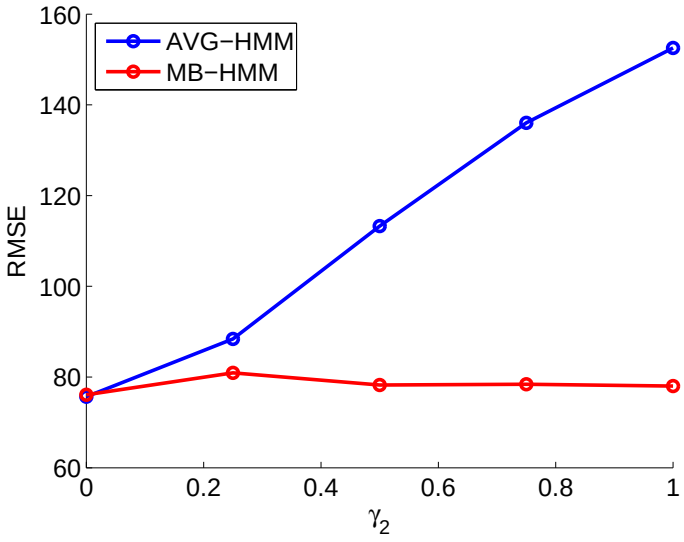


Figure 4.16: MB-HMM vs AVG-HMM at different mode distances

On the other hand, the advantage of the multi-branch model becomes clearer when several deterioration modes co-exist. In effect, the $RMSE$ values obtained by the MB-HMM model slightly decrease with the increase of the mode distance while they increase

for the AVG-HMM model. This result can be explained by the capacity of taking into account the co-existence of different deterioration modes. It can also be concluded that the larger the distance between the deterioration modes is, the better RUL estimation results can be obtained by the proposed MB-HMM model in comparison with the standard but mono-branch HMM model.

4.3.3.2 Co-existence of three deterioration modes

The above conclusion is still hold in case there are more than 2 modes that co-exist. Indeed, Figure 4.17 shows the errors in RUL estimation of the two models MB-HMM and AVG-HMM in case that there exists three deterioration modes. The two modes are the same as in the previous study while the third one represents the very fast deterioration rate, i.e. in severe mode of the component. The distance between the modes is also described through the parameter γ of the FCG model. For the sake of simplicity, we assume in this study that the distance between the severe mode and the fast mode equals to the one between the fast mode and the normal one.

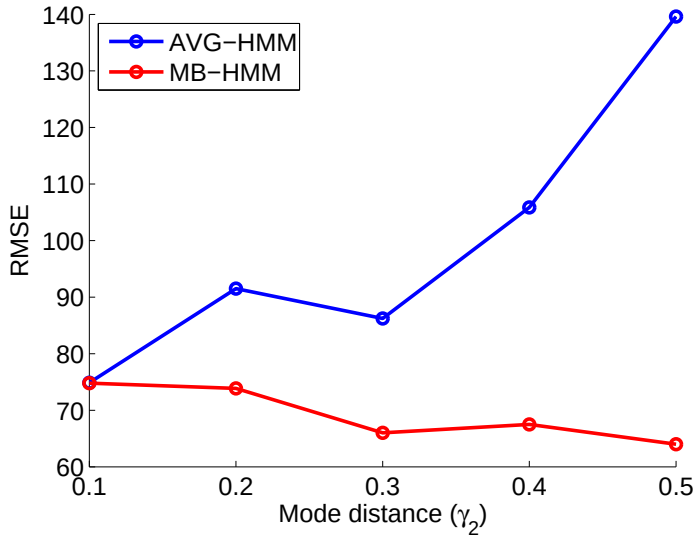


Figure 4.17: MB-HMM vs AVG-HMM in case of 3 modes

In Figure 4.17, the x-axis represents the different values of γ_2. We can see that when the distance between the modes is small, i.e. $\gamma_2 = 0.1$, the multi-branch and the average models give the same RUL estimation performances. However, the advantages of the multi-branch model becomes clearer with the increment of the distance.

4.4 Chapter summary

In this chapter, we have developed the multi-branch HMM model for dealing with the coexistence problem of different modes of deterioration. Based on the developed model, a two-phase diagnostics and prognostics framework has also been proposed. Through an application to the FCG data, we show that by using several branches for representing different modes, the MB-HMM model can give a better performance in RUL estimation compared to an "average" HMM one, especially when there is a large "distance" between the deterioration modes.

However, due to its inherent Markovian property, the sojourn time within a state of the Markov-based model follows either an exponential or a geometric distribution, which may not be hold in several practical applications. This leads to the inaccurate RUL estimations as discussed in Section 4.3.1.3 and 4.3.2.2. In the next chapter, we will extend the MB-HMM model by relaxing the "Markovian property" to overcome this limitation.

Chapter 5

Multi-branch Hidden semi-Markov Model

5.1 Introduction

In the previous chapter, we have developed the multi-branch model based on the Hidden Markov Model for dealing with the co-existence problem of multiple deterioration modes. However, due to its inherent Markovian property, the sojourn time staying in a state of the model is either geometrical or exponential distributed [128, 165]. As argued in [133], this can be a source of inaccurate duration modeling with the Markov-based models since most real-life systems do not exhibit this property. To overcome this problem, the underlying Markov process can be replaced by a semi-Markov one which allows the state sojourn time following any arbitrary probability distributions.

The goal of this chapter is to develop a multi-branch Hidden semi-Markov Model to overcome the limitation of the MB-HMM introduced in the last chapter. Firstly, we review some mathematical background of the HsMM model as well as its three associated problems. A two-phase framework for diagnostics and prognostics similarly to the case of the MB-HMM model is provided. Due to the extension to the semi-Markov, the difficulties and challenges in model training and in RUL estimation will be addressed. The performance of the proposed model is evaluated through two numerical studies: In the first simulated study with the FCG data, we show that the new MB-HsMM model can outperform the standard HsMM model as well as the MB-HMM model in estimating the RUL of equipment. After that, through a case study, we show that by effectively taking into account the coexistence of several deterioration modes, the MB-HsMM model can give promising results in comparison with the other continuous-state based models.

5.2 Hidden semi-Markov models background

5.2.1 Elements of HsMM model

According to Shun-Zheng Yu [165], the HsMM model is an extension of HMM by modeling the underlying state process as a semi-Markov chain. The key difference that

distinguishes the HsMM model from the HMM one is that the Markov property does not hold for every time steps in the HsMM model. More specifically, once entering into a state S_i, the HsMM model will stay in this state for a duration d determined by an arbitrary distribution. However, the Markov property holds at the end of this duration: the model transits from the state S_i to another state S_j with $j \neq i$ according to the transition probability matrix. This is where the word "semi-Markov" comes from. Figure 5.1 illustrates an example explaining more clearly this semi-Markov property.

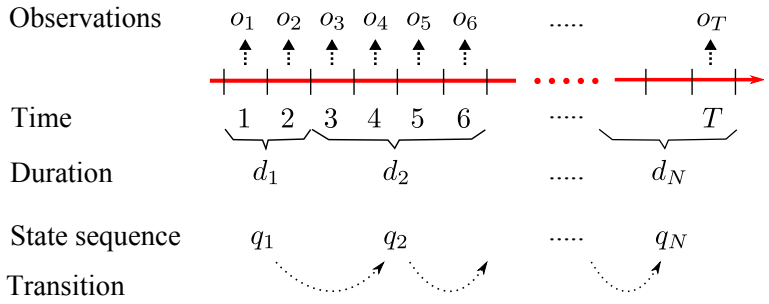


Figure 5.1: General HsMM [165]

In this figure, the state at time $t = 1$ denoted by q_1 and its duration d_1 are selected randomly according to the initial state probabilities π and the duration distribution corresponding to this state. The state q_1 then lasts for $d_1 = 2$ time units in this example. It emits two observations (o_1, o_2) according to the emission probability corresponding to state q_1. After that, it transits, according to the state transition probability matrix A to state q_2. It stays in this state for $d_2 = 4$ time units and produces four observations (o_3, o_4, o_5, o_6) according to the emission probability distribution of state q_2. This procedure is repeated until the final observation is produced.

To describe a HsMM model, apart from the parameters π, A, B as used for the HMM model, an additional parameter taking into account the distribution of the state duration is needed. In the above example, the state duration is assumed to be a discrete random variable taking integer values from the set $\mathcal{D} = \{1, 2, \ldots\}$. From a model learning point of view, this leads to a great number of parameters to be estimated. To overcome this problem, many studies in the literature have modeled the state duration by a continuous parametric density function such as Gamma or Gaussian [33, 32, 94, 87, 103].

In the framework of the deterioration modeling, similarly to the case of the HMM model, we are interested in the continuous left-right HsMM model in this chapter. The initial state distribution π and the observation models B are the same as in the HMM case (c.f. Equations (4.1) and (4.5)). Regarding the state transition probability matrix A, since the sojourn time in a state is determined by a state duration distribution and not by the state transition matrix, we can let the self-transition probabilities of the matrix equal 0: $a_{ii} = 0 \quad \forall i = 1..N$. It means that right after the end of period staying in state

S_i, the model will certainly move to an another state. We have:

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{5.1}$$

where for the state N: $a_{Ni} = 0 \quad \forall i$. It should be noted that the matrix A in this case is known and not need to be estimated.

An HsMM can be denoted by the compact notation: $\lambda = (\pi, A, B, D)$ where the letter D is used to denote the parameters relating the state sojourn time distributions.

Similar to the HMM model, three basic problems (i.e. the evaluation, the decoding and the training problems) must be solved in order to be able to use the HsMM model in real-life applications. Since the Markovian is relaxed in the semi-Markov based models, in the next section we will investigate a modified forward-backward algorithm to solve these three problems for the HsMM case.

5.2.2 Forward-Backward algorithm for HsMM models

The forward-backward (FB) procedure aims to answer the first evaluation problem of HsMM: given an HsMM model λ and a sequence of observations $\mathbf{O} = \{o_1, o_2, \dots, o_T\}$, how to compute the probability that the observed sequence was produced by the model. In other words, we want to compute a score of how well the given HsMM model λ matches the given observation sequence $\mathbf{O}$. In the case of HMM model, the forward-backward variables are defined as the joint probabilities of the state ending at a certain time and a series of observations up to that time (c.f. Appendix A). This approach has been firstly applied for the HsMM cases by Ferguson [37] in 1980. Based on the Ferguson's algorithm, Levinson [87] and Mitchell *et al.* [102] then proposed a recursive method for calculating more efficiently the joint probability distribution of a sequence of observations required by Ferguson's algorithm. However, the refined algorithms are still computationally too intensive in some applications [167]. A more efficient FB algorithm for HsMM models has been recently proposed by Yu and Kobayashi in [166] and [167]. In effect, by defining the FB variables based on the notion of a state together with its remaining sojourn time, Yu and Kobayashi showed that it is the most efficient one among the existing approaches, leading it to be practical in many applications [167]. In the next sections, we adapt this FB algorithm for our case of continuous observations and continuous state duration. For a more detailed explication of the algorithm, the reader can refer to [166] and [167].

5.2.2.1 Forward recursion formula

In [167], Yu modeled the state duration by a discrete variable taking value d with probability $p_i(d)$ from a finite set of integer values: $d \in \{1, 2, \ldots, D_{max}\}$ where D_{max} is the maximum time steps that the model could stay in a state. Regarding industrial applications, the condition monitoring data are also often acquired periodically in discrete time steps. This assumption is therefore adequate for the deterioration modeling purpose. However, with the aim of adapting continuous density functions such as the Gaussian distribution for state duration in order to reduce the number of parameters needed to be estimated, the discrete counterpart of the chosen parametric pdf can be used. The idea is to model the state duration with the best fitting parametric pdf, and then consider the discrete counterpart of this density function as the best probability mass function (pmf) [3, 4, 20].

Similar to the HMM case, we denote q_t the state of the HsMM model at time t that can take the value in a finite set: $S = \{S_1, S_2, \ldots, S_N\}$; o_t the observation at time t, $t = 1, 2, \ldots, T$ and $b_i(o_t)$ the probability of observing o_t under the state S_i, $i = 1, 2, \ldots, N$. In addition, let τ_t denote the remaining sojourn time of the current state q_t, the forward variable can be defined as follows:

$$\alpha_{t|x}(i, d) = P(q_t = S_i, \tau_t = d \mid \mathbf{o}_1^x) \tag{5.2}$$

where $x = t - 1$, t or T corresponding to the "predicted", "filtered" or "smoothed" probabilities of (q_t, τ_t) and $\mathbf{o}_a^b$ denote the observation sequence from time a to b.

This variable can be interpreted as the joint probability that the model is currently in state S_i and will stay in this state for the more d time steps. For a recursive computation of this variable, we need to define some intermediate variables:

- The ratio of the filtered probability $\alpha_{t|t}(i, d)$ over the predicted one is defined by:

$$b_i^*(o_t) \equiv \frac{\alpha_{t|t}(i, d)}{\alpha_{t|t-1}(i, d)} = \frac{b_i(o_t)}{\mathbb{P}\left(o_t \mid \mathbf{o}_1^{t-1}\right)} \tag{5.3}$$

  where $\mathbb{P}\left(o_t \mid \mathbf{o}_1^{t-1}\right)$ is the one-step prediction of the observation and can be determined by

$$\mathbb{P}\left(o_t \mid \mathbf{o}_1^{t-1}\right) = \sum_{i,d} \alpha_{t|t-1}(i, d)\, b_i(o_t) \tag{5.4}$$

- The conditional probability of a state ending at t given the observation up to time t is defined as:

$$\mathcal{E}_t(i) = P\left(q_t = S_i, \tau_t = 1 \mid \mathbf{o}_1^t\right) = \alpha_{t|t-1}(i, 1)\, b_i^*(o_t) \tag{5.5}$$

- The conditional probability of a state starting at $t+1$ given o_1^t:

$$\mathcal{S}_t(i) = P\left(\tau_t = 1, q_{t+1} = S_i \mid \mathbf{o}_1^t\right) = \sum_j \mathcal{E}_t(j)\, a_{ji} \tag{5.6}$$

  where a_{ji} is the transition probability from state S_j to state S_i.

The recursion formula for computing the forward variable can be obtained by examining all the possible state transition at time $t-1$ to reach the state $(q_t, \tau_t) = (S_i, d)$ at time t. We realize that this state may transit either from any other states $(S_j, 1)$ with $\forall j \neq i$ or continue the previous state, i.e. $(q_{t-1}, \tau_{t-1}) = (S_i, d+1)$. Therefore, the forward variable can be recursively calculated by:

$$\alpha_{t|t-1}(i,d) = S_{t-1}(i)\, p_i(d) + b_i^*(o_{t-1})\, \alpha_{t-1|t-2}(i, d+1) \tag{5.7}$$

with the initial value:

$$\alpha_{1|0}(i,d) = \pi_i p_i(d) \tag{5.8}$$

Given the "predicted" forward variables, the likelihood function of the observation sequence $\mathbf{o}_1^T$ can be calculated by:

$$\mathbb{P}\left(\mathbf{o}_1^T\right) = \prod_{t=1}^{T} \mathbb{P}\left(o_t \mid \mathbf{o}_1^{t-1}\right) \tag{5.9}$$

As in the case of the HMM models, the forward variables only are efficient to compute the likelihood function of the observations, i.e. answer the evaluation problem. However, the backward variables are still needed for the parameters learning purpose.

5.2.2.2 Backward recursion formula

We define the backward variable by the ratio of the smoothed probability $\alpha_{t|T}(i,d)$ over the predicted one $\alpha_{t|t-1}(i,d)$:

$$\beta_t(i,d) = \frac{P\left(q_t = S_i, \tau_t = d \mid \mathbf{o}_1^T\right)}{P\left(q_t = S_i, \tau_t = d \mid \mathbf{o}_1^{t-1}\right)} \tag{5.10}$$

with the initial value:

$$\beta_T(i,d) = b_i^*(o_T) \tag{5.11}$$

To derive the recursion formula for the backward variable, we examine all possible states that follow the state $(q_t, \tau_t) = (S_i, d)$. When $d = 1$, the next state can be $(q_{t+1}, \tau_{t+1}) = (S_j, d')$ for any $j \neq j$ and $d' \geq 1$. In contrast, when $d > 1$, the next state must be $(q_{t+1}, \tau_{t+1}) = (S_i, d-1)$. Therefore, the backward variables can be calculated by

the following recursion formula:

$$\beta_t(i,d) = \begin{cases} \mathcal{S}^*_{t+1}(i)\, b^*_i(o_t), & d = 1 \\ \beta_{t+1}(i,d-1)\, b^*_i(o_t), & d > 1 \end{cases} \tag{5.12}$$

where

$$\begin{aligned} \mathcal{S}^*_t(i) &= \frac{P\left(\mathbf{o}_1^T \mid q_{t-1} = S_i, \tau_{t-1} = 1\right)}{P\left(\mathbf{o}_1^T \mid \mathbf{o}_1^{t-1}\right)} \\ &= \sum_j a_{ij} \mathcal{E}^*_t(j) \end{aligned} \tag{5.13}$$

and

$$\begin{aligned} \mathcal{E}^*_t(i) &= \frac{P\left(\mathbf{o}_1^T \mid q_t = S_i, \tau_{t-1} = 1\right)}{P\left(\mathbf{o}_1^T \mid \mathbf{o}_1^{t-1}\right)} \\ &= \sum_d p_i(d)\, \beta_t(i,d) \end{aligned} \tag{5.14}$$

The FB algorithm for the HsMM models is summarized in Algorithm 1.

Algorithm 1 Forward-Backward algorithm for HsMM models

1. **The forward recursion**

   For $t = 1, 2, ..., T$

   - Calculate $\alpha_{t|t-1}(i,d)$ by Equations (5.8) and (5.7);
   - Calculate $b^*_i(o_t)$ by Equation (5.3);
   - Calculate $\mathcal{E}_t(i)$ by Equation (5.5);
   - Calculate $\mathcal{S}_t(i)$ by Equation (5.6).

2. **The backward recursion**

   For $t = T, T-1, ..., 1$

   - Calculate $b^*_i(o_t)$ by Equations (5.11) and (5.12);
   - Calculate $\mathcal{E}^*_t(i)$ by Equation (5.14);
   - Calculate $\mathcal{S}^*_t(i)$ by Equation (5.13).

5.2.3 Parameters re-estimation

Based on the Forward-Backward variables, the model parameters can be estimated by the principle of the Expectation-Maximization algorithm [29]. To this end, we construct firstly the re-estimation formulas for the model parameters. The following "smoothed"

probabilities that are the conditional probabilities given all the observations are defined:

- Smoothed probability that a transition occurs from state S_i to state S_j at time t:

$$\begin{aligned}\mathcal{T}_{t|T}(i,j) &= \mathbb{P}\left(q_{t-1}=S_i, \tau_{t-1}=1, q_t=S_j \mid \mathbf{o}_1^T\right)\\ &= \mathcal{E}_{t-1}(i)\, a_{ij} \mathcal{E}_t^*(j)\end{aligned} \tag{5.15}$$

- Smoothed probability that state S_i is entered at time t and lasts for d time units:

$$\begin{aligned}\mathcal{D}_{t|T}(i,d) &= \mathbb{P}\left(\tau_{t-1}=1, q_t=S_i, \tau_t=d \mid \mathbf{o}_1^T\right)\\ &= \mathcal{S}_{t-1}(i)\, p_i(d)\, \beta_t(i,d)\end{aligned} \tag{5.16}$$

- Smoothed conditional probability of being in state S_i at time t given the observations:

$$\begin{aligned}\gamma_{t|T}(i) &= \mathbb{P}\left(q_t=S_i \mid \mathbf{o}_1^T\right)\\ &= \sum_d \alpha_{t|T}(i,d)\end{aligned} \tag{5.17}$$

  where

$$\alpha_{t|T}(i,d) = \alpha_{t|t-1}(i,d)\, \beta_t(i,d) \tag{5.18}$$

  Note that this smoothed probability can also be calculated in a recursive way leading to a considerable reduction in computation [167]:

$$\gamma_{t-1|T}(i) = \gamma_{t|T}(i) + \mathcal{E}_{t-1}(i)\, \mathcal{S}_t^*(i) - \mathcal{S}_{t-1}(i)\, \mathcal{E}_t^*(i) \tag{5.19}$$

The above smoothed probabilities are calculated given that the model parameters λ are known. In case that λ is not known, based on the principle of the EM algorithm, we train the model for given observations $\mathbf{o}_1^T$, where λ is initially estimated and then re-estimated multiple times until the likelihood of the observations increases and converges to a certain value. Such training and updating process is referred to as parameter re-estimation [167].

Specifically, given the above smoothed probabilities and by applying the maximum-likelihood principle as in the HMM case, the model parameters can be re-estimated by [167]

$$\hat{\pi}_i = \frac{\gamma_{t|T}(i)}{N_\pi} \tag{5.20}$$

$$\hat{a}_{ij} = \sum_{t=2}^{T} \frac{\mathcal{T}_{t|T}(i,j)}{N_a} \tag{5.21}$$

$$\hat{p}_i(d) = \sum_{t=2}^{T} \frac{\mathcal{D}_{t|T}(i,d)}{N_p} \tag{5.22}$$

where N_π, N_a, N_p are normalized factor so that $\sum_i \hat{\pi}_i = 1$; $\sum_n \hat{a}_{ij} = 1$ and $\sum_d \hat{p}_i(d) = 1$. Note that the re-estimation formula of the duration d is used to estimate its probability mass function. In case of continuous parametric distribution, its parameters can be estimated by approximating the estimated mass function by a probability density function, i.e. by choosing the mean and variance values of the continuous distribution coincide with the ones of the mass function. For instant, in case of using the univariate Gaussian distribution for the sojourn time, the mean and the variance parameters are re-estimated by:

$$\hat{\mu}_d(i) = \frac{\sum_d d \cdot \hat{p}_i(d)}{\sum_d \hat{p}_i(d)} \tag{5.23}$$

$$\hat{\sigma}_d^2(i) = \frac{\sum_d \hat{p}_i(d)\left(d - \hat{\mu}_d(i)\right)^2}{\sum_d \hat{p}_i(d)} \tag{5.24}$$

Regarding the observation model, the parameters of the mixture Gaussian distribution can be re-estimated by the same principle used for the HMM model case [128]. To this end, we need firstly calculate the smoothed probability $\gamma_t(i,k)$ which is the probability of being in state S_i at time t with the kth mixture component, i.e. $\gamma_t(i,k) = \mathbb{P}\left(q_t = S_i, K_t = k \mid \mathbf{o}_1^T\right)$ where K_t denotes the mixture component at time t.

This smoothed probability can be readily computed given the smoothed probability $\gamma_{t|T}(i)$ as follows [105]. Let $z_t = \{o_1, \ldots, o_{t-1}, o_{t+1}, \ldots, o_T\}$ be all the observations except o_t. From the Bayesian formula, we have:

$$\begin{aligned}
\gamma_t(i,k) &= \mathbb{P}(q_t = S_i, K_t = k \mid o_t, z_t) \\
&= \frac{\mathbb{P}(o_t \mid q_t = S_i, K_t = k, z_t)\,\mathbb{P}(q_t = S_i, K_t = k \mid z_t)}{\mathbb{P}(o_t \mid z_t)} \\
&= \frac{\mathbb{P}(o_t \mid q_t = S_i, K_t = k)\,\mathbb{P}(K_t = k \mid q_t = S_i)\,\mathbb{P}(q_t = S_i \mid z_t)}{\mathbb{P}(o_t \mid z_t)}
\end{aligned} \tag{5.25}$$

On the other hands, from its definition:

$$\begin{aligned}
\gamma_{t|T}(i) &= \mathbb{P}(q_t = S_i \mid o_t, z_t) \\
&= \frac{\mathbb{P}(o_t \mid q_t = S_i)\,\mathbb{P}(q_t = S_i \mid z_t)}{\mathbb{P}(o_t \mid z_t)}
\end{aligned} \tag{5.26}$$

This leads to:

$$\frac{\mathbb{P}(q_t = S_i \mid z_t)}{\mathbb{P}(o_t \mid z_t)} = \frac{\mathbb{P}(q_t = S_i \mid o_t, z_t)}{\mathbb{P}(o_t \mid q_t = S_i)} \tag{5.27}$$

Replace this into Equation (5.25), we obtain:

$$\gamma_t(i,k) = \frac{\mathbb{P}(o_t \mid q_t = S_i, K_t = k)\,\mathbb{P}(K_t = k \mid q_t = S_i)\,\gamma_{t|T}(i)}{\mathbb{P}(o_t \mid q_t = S_i)} \tag{5.28}$$

where $\mathbb{P}(o_t \mid q_t = S_i, K_t = k)$ is the probability density function at o_t given the state S_i and kth mixture component; $\mathbb{P}(K_t = k \mid q_t = S_i)$ is the probability of the kth mixture component given the state S_i and $\mathbb{P}(o_t \mid q_t = S_i) = \sum_k \mathbb{P}(o_t \mid q_t = S_i, K_t = k)$.

Given the smoothed probability $\gamma_t(i,k)$, the re-estimation formulas for the coefficients of the mixture density, i.e. c_{jk}, μ_{jk}, Σ_{jk} (c.f. Equation (4.1)) are given by [128]:

$$\hat{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t(j,k)} \tag{5.29}$$

$$\hat{\mu}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot o_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{5.30}$$

$$\hat{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (o_t - \hat{\mu}_{jk})(o_t - \hat{\mu}_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{5.31}$$

where prime denotes vector transpose.

The re-estimation formula for c_{jk} can be interpreted as the ratio between the expected number of times the model is in state S_j using the kth mixture component, and the expected number of times the model is in state S_j. Similarly, Equation (5.30) gives the expected value of the portion of the observation vector accounted for by the kth mixture component. A similar interpretation can be easily obtained for Equation (5.31) [128].

5.2.4 Viterbi algorithm

Given the smooth probabilities $\gamma_{t|T}(i)$, one possible solution to find out the "optimal" state sequence for the HsMM model compared to the HMM one is to modify the optimality criterion. In effect, instead of finding the single best state sequence (path) as in the HMM case, we choose the states q_t which are individually most likely, that is:

$$\hat{q}_t = \arg\max_i \gamma_{t|T}(i) \tag{5.32}$$

In this manner, a sequence of states $(\hat{q}_1, \ldots, \hat{q}_T)$ can be estimated.

5.3 Extension to Multi-branch HsMM

As already mentioned, the standard HsMM model can take into account only one deterioration mode. To deal with the problem of co-existence of multiple deterioration modes, we extend the HsMM to a multi-branch HsMM model (MB-HsMM) in this section. The principle is similar to the one used in Chapter 4: The MB-HsMM model consists of several branches in which each one represents a deterioration mode and the two non-emitting states in common. The extended model is also based on the continuous left-right HsMM model (c.f. Figure 5.2).

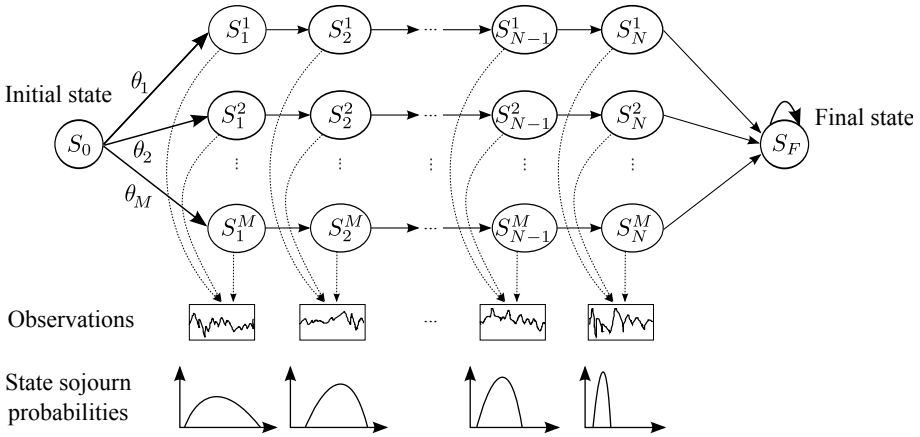


Figure 5.2: A left-right multi-branch HsMM model

Different from the MB-HMM case, there is no self-transition of the states in this model. Instead, the model will stay in a given state S_i for a duration d_i which is determined by the state sojourn probabilities. In addition, when staying in this state, the model emits d_i observations according to the observation model corresponding to state S_i. Furthermore, one can realize that similar to the MB-HMM case, mode switching is not allowed in this model. In other words, the model can follow only one of the branches from the initial state S_0 till it reaches the final one S_F. The probability that the model follows the mth branch is denoted by θ_m for $m = 1, 2, \ldots, M$.

In the context of deterioration modeling, as discussed in Section 4.1.2, the initial and final states correspond to the normal health and failure ones respectively and they are assumed to be the non-emitting states. The equipment is said to be failed once it reaches for the first time the final state S_F. Furthermore, S_F is an absorbing state. Each branch together with this state forms a continuous left-right HsMM model with a final non-emitting state. Therefore we must make some modifications for the initial state probabilities π, the state transition matrix A, and the observation matrix as done in Section 4.2.1 in order to be able to apply the Forward-Backward algorithm presented in the last section. For example, the state transition matrix with the non-emitting state

S_F becomes:

$$A_n = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & \mathbf{1} \end{bmatrix} \tag{5.33}$$

It should be noted that in case that the continuous left-right topology is used, this matrix does not need to be re-estimated.

5.3.1 MB-HsMM based framework for diagnostics and prognostics

Based on the MB-HsMM model, a two-phase framework is also proposed for the diagnostics and the prognostics in Figure 5.3.

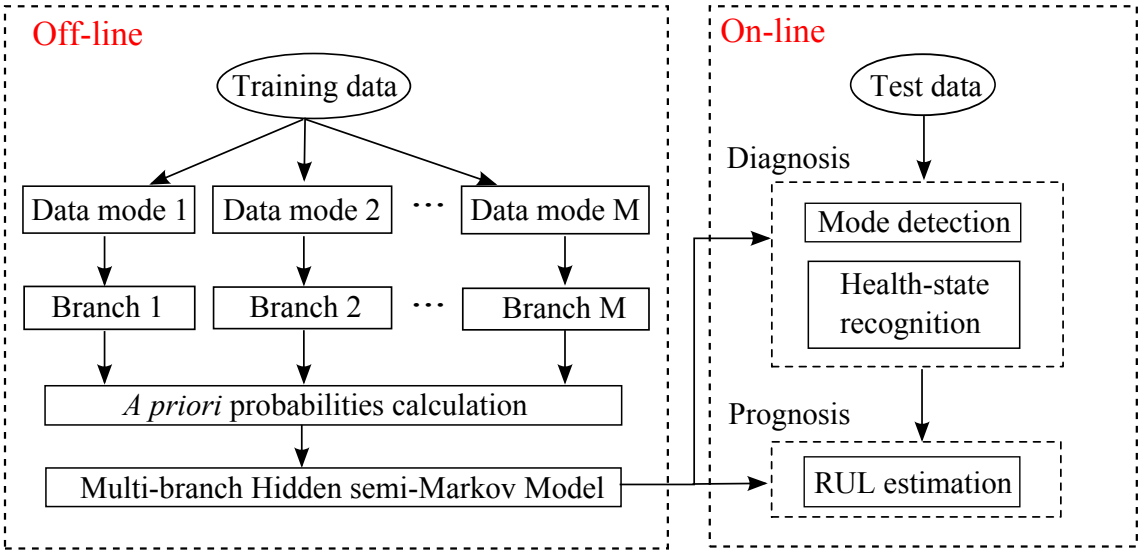


Figure 5.3: MB-HsMM framework for diagnostics and prognostics

This framework is similar to the one presented in Section 4.2 for the MB-HMM model. It consists of the two phases: off-line and on-line phases. In the off-line phase, the idea is also to divide the training data set into M subsets corresponding to the M deterioration modes of equipment and then use each subset for training one branch of the multi-branch model. While in the MB-HMM case this task can be fulfilled by the standard Baum-Welch algorithm, the parameters re-estimation procedure based on the modified forward-backward algorithm presented in Section 5.2.3 will be implemented for the MB-HsMM model. The *a priori* mode probabilities can be computed similarly as for the MB-HMM model, i.e. by Equation (4.13).

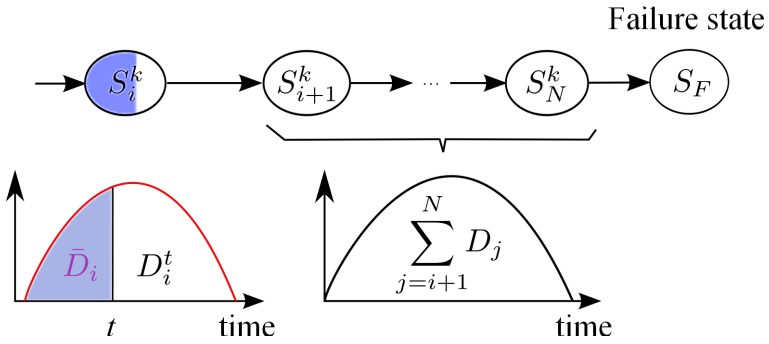


Figure 5.4: Illustration of RUL estimation for HsMM

In the on-line phase, the forward-backward algorithm is firstly applied for calculating the likelihood of each individual branch of the model given a sequence of test data. The mode detection task is then accomplished by choosing the branch that has the maximum posterior probability given the test data (c.f. Equation (4.14)). Once the underlying deterioration mode has been detected, the FB-based Viterbi algorithm presented in the previous section can be carried out for estimating the actual health state of the monitored equipment.

It should be noted that as the inherent Markov property has been relaxed, the RUL estimation procedure for the MB-HsMM model is different from the MB-HMM case. This task will be tackled in the next section.

5.3.2 RUL calculation

Suppose that in the last mode detection step, the model is detected to be currently in the mode k and suppose that thanks to the Viterbi algorithm, we know that it is actually in the state S_i at time t and have been in this state for $\overline{D}_i$ time units. The RUL - the remaining time units for reaching state S_F in this case is the sum of the two variables: one concerning the residual sojourn time staying in state S_i and one for the total times spending for the next health states before entering state S_F (c.f. Figure 5.4).

Denote D_i and D_i^r the random variables representing the total and the residual sojourn time respectively staying in state S_i for $i = 1, 2, \ldots N$. The RUL at time t is given by:

$$RUL_t = D_i^r + \sum_{j=i+1}^{N} D_j \tag{5.34}$$

The residual sojourn time D_i^r is considered as a conditional random variable $D_i^t = D_i - \overline{D}_i | D_i > \overline{D}_i$. If D_i is Gaussian distributed, e.g. $D_i \sim N(\mu_d(i), \sigma_d(i))$, it can be shown that the residual time $D_i^t | D_i^t > 0$ follows a truncated Gaussian distribution with

mean $\mu_d(i) - \overline{D}_i$ and standard deviation $\sigma_d(i)$ with a truncation to the left of zero.

Furthermore, we know that the sum of Gaussian distributed random variables is also a Gaussian distributed one. The sum of sojourn times staying in states $S_{j+1}, \ldots, S_N$ is hence Gaussian distributed. In effect, denote $Z = \sum_{j=i+1}^{N} D_j$, we have $Z \sim N(\mu_z, \sigma_z)$ where $\mu_z = \sum_{j=i+1}^{N} \mu_d(j)$ and $\sigma_z = \sqrt{\sum_{j=i+1}^{N} \sigma_d^2(j)}$. The estimation of the RUL at time t hence becomes the computation of the sum of two variables, one follows a Gaussian distribution and the other one follows a truncated Gaussian distribution.

According to Nelson in [112], if we denote $L(h, k, r) = P(X > h, Y > k)$ where (X, Y) are two random variables following standard bivariate Normal distribution with the correlation coefficient r, the cumulative distribution function of the RUL can be given by:

$$F_{RUL}(x) = \frac{-L(h, k, r) + 1 - \Phi(k)}{\Phi(h)} \tag{5.35}$$

where $h = \dfrac{\mu_d(i) - \overline{D}_i}{\sigma_d(i)}$, $k = \dfrac{\mu_d(i) + \mu_z - x}{\sqrt{\sigma_d^2(i) + \sigma_z^2}}$, $r = \dfrac{\sigma_d(i)}{\sqrt{\sigma_d^2(i) + \sigma_z^2}}$ and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Normal distribution.

It should be noted that this calculation is carried out for the current detected branch k of the multi-branch model. To calculate the RUL for the whole model and similar to the case of the MB-HMM model, we apply the Bayesian Model Averaging (BMA) technique ([60]) in order to take into account the uncertainty in mode detection step. Indeed, the final RUL distribution is calculated as an average of the posterior RUL distributions for all constituent branches weighted by their posterior branch probabilities:

$$P(RUL \mid \mathbf{O}) = \sum_{m=1}^{M} \mathbb{P}(RUL \mid \lambda_m, \mathbf{O}) \, \mathbb{P}(\lambda_m \mid \mathbf{O}) \tag{5.36}$$

where $\mathbb{P}(\lambda_m \mid \mathbf{O})$ is calculated as in Equation (4.15).

5.4 Numerical results

In order to investigate the effectiveness of the proposed MB-HsMM model, several numerical studies are performed both on simulated and real data. The simulated data are generated using the same stochastic FCG model as described in Chapter 4. The main purpose is to conduct a comparative study in RUL estimation between the multi-branch HsMM model versus the standard HsMM one as well as versus the multi-branch HMM model developed in the last chapter. After that, a case study taking from the Prognostic and Health Management 2008 competition will be considered in Subsection 5.4.2.

5.4.1 Simulation study

Consider the case of deterioration due to the apparition and propagation of a crack, we use the discretized and randomized version of the FCG model for data generating purpose (c.f. Section 4.3.2.1):

$$x_{t_i} = x_{t_{i-1}} + e^{w_{t_i}} C \left(\beta \sqrt{x_{t_{i-1}}}\right)^n \Delta t \tag{5.37}$$

where x_{t_i} denotes the crack depth at time t_i, C, and n are constants depending on the material property, β is a factor representing the relation between the stress intensity amplitude and the crack depth, w_{t_i} are independent and identically distributed according to a Normal distribution $\mathcal{N}(0, \sigma_w^2)$.

The parameter β is suppose to be a function of operating environment state for the purpose of representing different propagation rates:

$$\beta(e) = \beta_b \cdot e^{\gamma_e} \tag{5.38}$$

where β_b is the base stress level of system, the values $\gamma_e \geq 0$, $e = 1, 2, \ldots, M$ determine the propagation rate of the crack depth.

The observation model is chosen by:

$$y_{t_i} = x_{t_i} + \xi_{t_i} \tag{5.39}$$

where $\xi_{t_i} \sim \mathcal{N}\left(0, \sigma_\xi^2\right)$ is the measurement error.

Similar to the study conducted in the MB-HMM case, two deterioration modes will be considered in this section. These two modes correspond to the two values of γ: $\gamma_1 = 0$ for slow-rate mode and $\gamma_2 = 0.5$ for fast-rate mode. The following parameters of the FCG model are used for the data generation: $C = 0.005$, $n = 1.3$, $\beta_b = 1$, $\sigma_w = 1.7$, $\Delta t = 1$, $\sigma_{\xi 1}^2 = 2$, $\sigma_{\xi 2}^2 = 5$. Besides that, the critical threshold $L = 100$ is also chosen. The two-mode crack depth data are shown in Figure 5.5.

In this study, 22 and 28 data sequences are generated for the slow-rate and fast-rate modes respectively. This data is then used as training data to estimate the parameters of the MB-HsMM model. Similar to the MB-HMM case, the first step in model training is to determine the structure of the MB-HsMM model to be trained. Specifically, we need to determine the number of hidden state for each branch N, the number of mixture components K in observation model. By using the BIC criterion, we select from Figure 5.6 the "knee" point $N = 10$ for the number of hidden states. Concerning the number of mixture components, we can see that there is no big difference in BIC values for $K = 1$, $K = 2$, $K = 3$ or $K = 4$. We choose therefore the least complex model corresponding to $K = 1$.

The parameters of the model are then estimated by the re-estimation procedure basing on the forward-backward algorithm introduced in Section 5.2.2. Figure 5.7 represents the

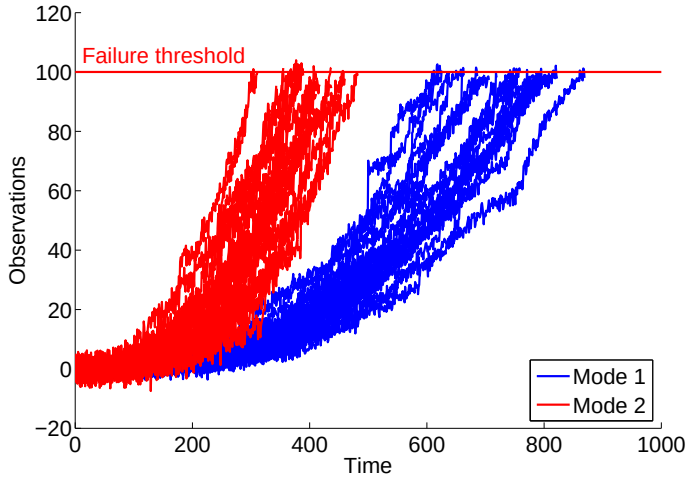


Figure 5.5: Measurements of crack depth in two modes

online RUL estimation result associated with the 95% prediction interval. By shifting the time by $50h$ from the beginning toward the failure, the RUL is re-estimated after each replacement with the update of new observations. The result demonstrates the accuracy of the RUL estimation method for the MB-HsMM model.

5.4.1.1 MB-HsMM vs HsMM and MB-HsMM vs MB-HMM

This section is dedicated to demonstrate the advantages of the MB-HsMM model in comparison with a standard HsMM model as well as with the MB-HMM developed in Chapter 4. We denote the standard HsMM model by "average" HsMM (AVG-HsMM) to emphasize that it has only one branch compared to the MB-HsMM and MB-HMM models.

Similar to the study in Section 4.3.3, we consider the crack depth data generated from the FCG model as the deterioration data and use the coefficient γ_e to represent the "distance" between the two modes, i.e. two crack propagation rates. Specifically, we fix $\gamma_1 = 0$ for the mode 1 (slow-rate mode) and do varying γ_2 within the interval $[0.25; 1]$ for the mode 2 (fast-rate mode). Corresponding to each value of γ_2, we generate 50 sequences of observations for the two modes which are used as the training data and we train the three models by this data set. After that, for the evaluation purpose, we generate in the online phase 100 other observation sequences from the same FCG model in which each sequence is considered as a test data. Corresponding to each sequence, we use the three trained models to estimate the RUL "online", i.e. by shifting the time-to-estimate by $50h$ from the beginning towards the failure. The estimation results are evaluated by using the root mean squared error (RMSE) criterion. The model that has the minimum RMSE value will be the best one.

Figure 5.8 represents the $RMSE$ values for the four models MB-HsMM, AVG-HsMM,

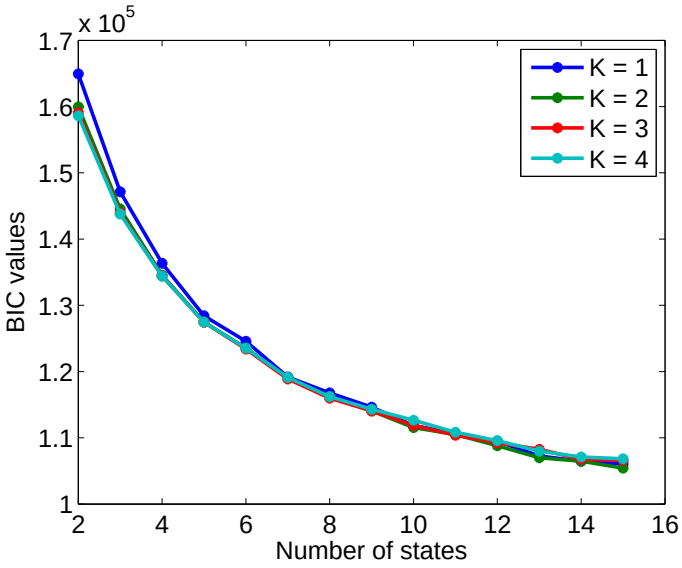


Figure 5.6: BIC values for determining N and K

MB-HMM and AVG-HMM respectively at different values of γ_2 or equivalently at different mode "distances". Obviously, we can see that the RMSE values of the MB-HsMM and MB-HMM models decrease while they are increase for the AVG-HsMM and AVG-HMM ones with the increment of the mode "distance". It means that when the distance between the deterioration modes is large, the MB-HsMM and MB-HMM models outperform the AVG-HsMM and AVG-HMM models in RUL estimation. This demonstrates the effectiveness of the multi-branch models in comparison with the standard mono-mode ones. Furthermore, the MB-HsMM model gives better estimation results than the MB-HMM one. This can be explained by the fact of using the semi-Markov chain in the former model to relax the inherent "Markovian" property in the MB-HMM one.

5.4.2 A case study

In this section, the MB-HsMM model is implemented to estimate the RUL for a real data case: the 2008 Prognostic and Health Management (PHM) competition.

5.4.2.1 PHM08 competition data

Prognostic and Health Management (PHM) 2008 is the first annual international conference sponsored by the IEEE reliability society. During this conference, a RUL estimation competition was proposed to the participants. There are two data sets that are available: one for training purpose and one for the test. Each data set consists of multivariate time series collected from 218 identical and independent units of an unspecified component. For the training data set, information about operational settings as well as sensor measurements are complete from the starting cycle until the failure while for the

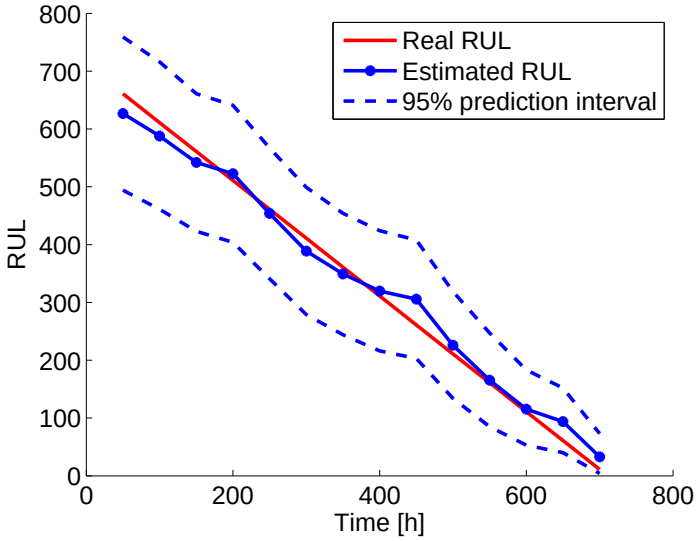


Figure 5.7: RUL estimation with MB-HsMM model

testing one, the time series are truncated at some random time prior to the failure. The objective of the competition is that the participants have to propose and construct a prognostics method by basing on the training data set and then apply it for estimating the number of remaining operational cycles of each unit in the testing data set. The accuracy of the proposed prognostics method will be evaluated by the following score:

$$S = \sum_{i=1}^{218} S_i \tag{5.40}$$

where S_i is the penalty score for unit i, computed as follows:

$$S_i = \begin{cases} e^{-d_i/13} - 1, & d_i \leq 0 \\ e^{d_i/10} - 1, & d_i > 0 \end{cases} \tag{5.41}$$

where $d_i = \hat{RUL}(i) - RUL(i)$ is the estimation error; $\hat{RUL}(i)$ and $RUL(i)$ are the estimated and the real RUL values respectively of the unit i.

This score originally introduced by the competition is asymmetrical and penalize more the late failure estimations where the error d_i is positive. In the literature, some papers use another criterion to evaluate the accuracy of methods, such as the root squared error (RSE):

$$RSE = \sqrt{\sum_{i=1}^{218} d_i^2} \tag{5.42}$$

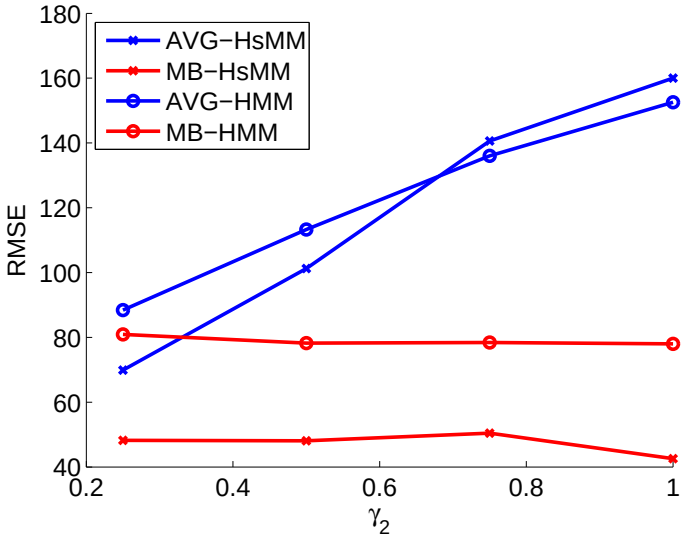


Figure 5.8: RMSE values at different mode distances

or the mean squared error (MSE):

$$MSE = \sum_{i=1}^{218} \frac{d_i^2}{218} \tag{5.43}$$

For all above scores, the lower the score is, the more accurately the RUL is estimated. In the framework of PHM 2008 challenge, three winners obtained the following results: [154] obtained a total score of $S = 5636$, [58] achieved a prediction error of $RSE = 519.8$ and [121] obtained a mean square error of $MSE = 984$.

5.4.2.2 Data description and health indicators

The PHM 2008 data sets are constructed by the C-MAPSS model (Commercial Modular Aero-Propulsion System Simulation), which is a tool developed by NASA research center for simulating a realistic large commercial turbofan engine (c.f. Figure 5.9). The objective is to generate the training and test data sets for the competition.

In the framework of the PHM08 competition, C-MAPSS is used to simulate an engine model of the $90,000lb$ thrust class operating at (i) altitudes ranging from sea level to $42,000ft$, (ii) Mach numbers from 0 to 0.84, and (iii) Throttle Resolver Angle (TRA) from 20 to 100. The engine is simulated to operate under six operational modes corresponding to different values of these parameters as summarized in Table 5.1.

The C-MAPSS model has 14 inputs including fuel flow and a set of 13 health-parameter inputs that allow the user to simulate the effects of faults and deterioration in any of the

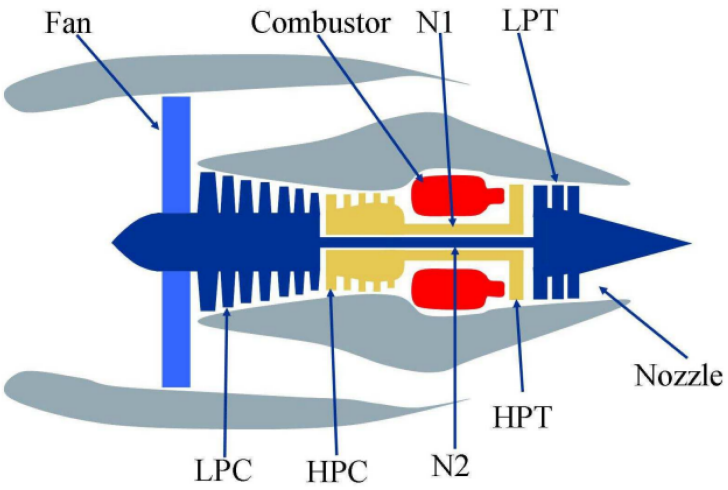


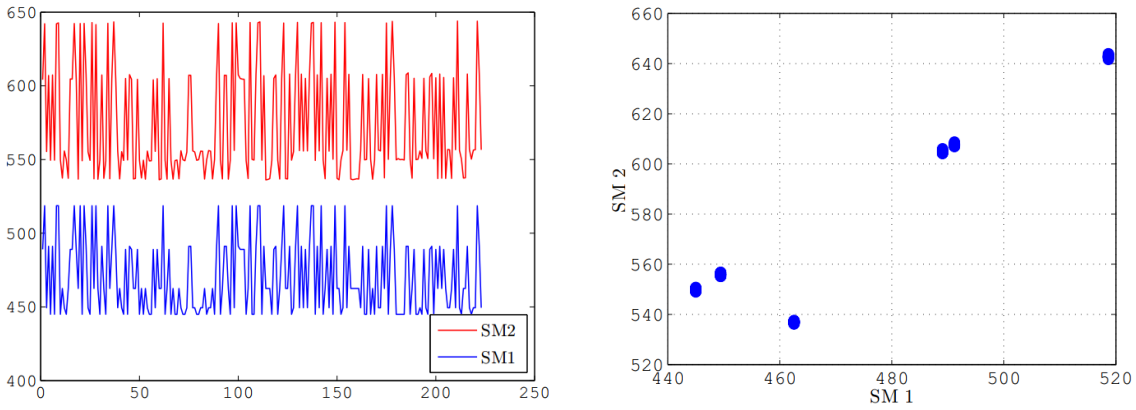
Figure 5.9: Simplified diagram of engine simulated in C-MAPSS [134]

Table 5.1: Operational modes of the simulated engine

Mode	Altitude (ft)	Mach numbers	TRA
1	0	0	100
2	$25K$	0.62	80
3	$35K$	0.84	60
4	$42K$	0.84	40
5	$20K$	0.25	20
6	$20K$	0.7	0

engine's five rotating components such as Fan, LPC, HPC, HPT, and LPT. The outputs include various sensor response surfaces and operability margins. A total of 21 variables out of 58 different outputs available from the model were provided to the participants of the competition. These variables are the sensor measurements of the temperature, the pressure, the speed, etc. of the 218 independent and identical units [134]. For example, Figure 5.10a illustrate the temporal evolution of the measurements obtained for the unit 1 from the two sensors SM1 and SM2 which represent the total temperatures (in Rankine degree) at fan inlet and at LPC outlet respectively.

It can be seen that the temporal evolution of such measurements does not exhibit a clear trend that can help to identify the level of deterioration of units. In addition, if we plot one measurement in function of the other, i.e. SM2 vs SM1, Figure 5.10b shows that the 6 points obtained represent exactly the 6 modes of operations. As the transitions between these points are stochastic and we do not know exactly the point that corresponds to the failure for every unit (we know the exact point for each training unit but it differs from one unit to the others), this plot does not hence provide the useful information concerning the underlying deterioration process. From this point,



(a) Temporal evolution of SM1 and SM2 (b) Relation between SM1 and SM2

Figure 5.10: Measurements from the sensors SM1 and SM2

it is necessary to construct a "health indicator" which represents the evolution of the deterioration level in time of every unit in order to apply the proposed MB-HsMM model. In addition, as the MB-HsMM model is based on the left-right topology, the indicators to be constructed should exhibit the same trend in evolution from the initial point until the end of life of the unit. In this study, we decide to use the indicators proposed by Le Son *et al.* [80] for two reasons. Firstly, their constructed indicators show clearly and homogeneous trends time for all the units (c.f. figure 5.11) and secondly, by using these indicators, the authors achieved better accuracy in RUL estimation compared to the winners of the competition. In effect, by using continuous stochastic processes such as a non homogeneous Wiener process and a non homogeneous Gamma process, the estimation results of ($S = 5520$; $RSE = 423$; $MSE = 819$) and ($S = 4170$; $RSE = 434$; $MSE = 864$) were obtained in [80] and [79] respectively. More details about the indicator construction, readers are referred to [80].

Although these indicators are continuous in time, we can consider them as the continuous observations of the MB-HsMM model and assume that the health (or the deterioration level) of the units can be classified into discrete states. In the next section, we will introduce the application of the MB-HsMM model for estimating the RUL of the competition based on the above indicators.

5.4.2.3 Application of the MB-HsMM model

To train the MB-HsMM model from the provided training data set, the first step is to determine the model structure, i.e. the number of branches M, the number of hidden states N and the number of mixture component K from the data. Firstly, consider the number of branches M: it equals to the number of modes of the underlying deterioration processes. It can be identified by re-examining the manner in which the deterioration is

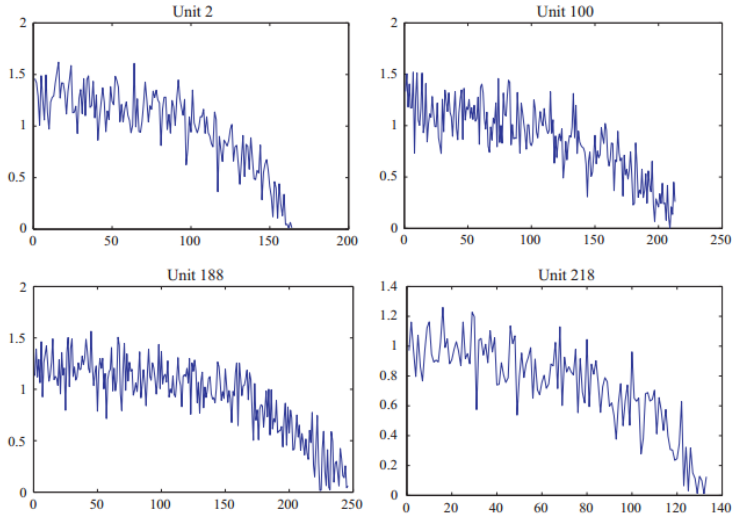


Figure 5.11: Some examples of degradation indicators ([80])

generated with the C-MAPSS model.

According to Saxena *et al.*, the deterioration data are generated by reducing exponentially the efficiency (e) and flow (f) parameters at the input of the C-MAPSS model, until the engine exceeds a safe operating region - which is determined via operability margins [134]. Depending on the direction of the failure evolution trajectory (c.f. Figure 5.12), a failure threshold may or may not be crossed. The overall stopping condition is hence determined by the margin that approaches the corresponding limit first.

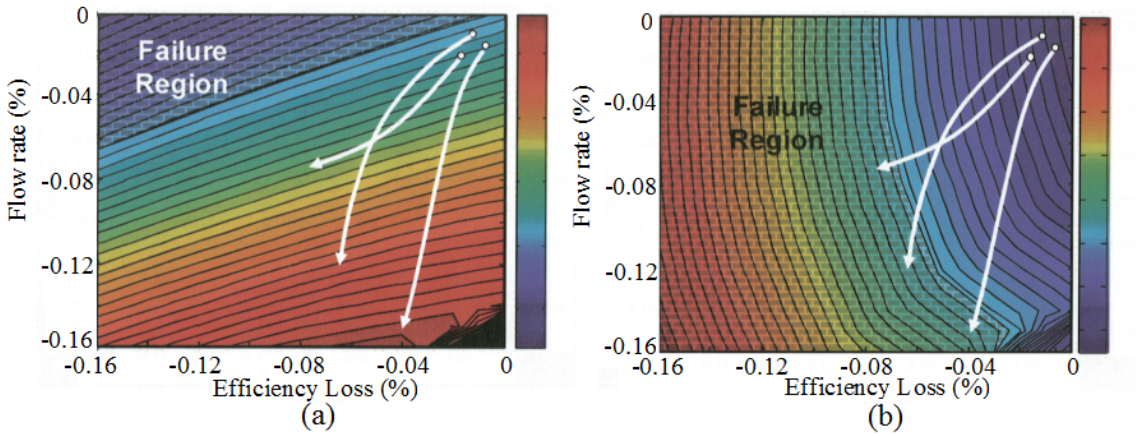


Figure 5.12: Fault propagation trajectories on HPC stall margin (a) and EGT (b) contour map with failure threshold ([134])

From Figure 5.12, we can see that the fault propagation trajectories vary depending on the correlation between the rates of reduction of the two parameters e and f. For example, an unit seems to last longer when the parameter f decreases faster than the parameter e and vice versa. It motivates us to divide the training data sets into different

groups corresponding to different "fault trajectories" for the purpose of training the multi-branch model. In this study, we assume that the training units can be classified into two, three or four groups depending on the differences in decrease rates of e and f. These groups correspond to the two, three of four modes of deterioration and are summarized in Table 5.2.

Table 5.2: Training data grouping

	2 groups	3 groups	4 groups
Group 1	$f < e$	$f < e$	$f \ll e$
Group 2	$f > e$	$f \approx e$	$f < e$
Group 3		$f > e$	$f > e$
Group 4			$f \gg e$

where the notation $f < e$ implies that the parameter f decreases more slowly than the parameter e and $f \approx e$ signifies that there is no big difference in the decreases of these parameters, etc.

Corresponding to each case, we have $M = 2$, $M = 3$ or $M = 4$ and the MB-HsMM model will have 2, 3, or 4 branches respectively. The next step is to divide the training data set into individual group so that each branch can be trained as presented in Section 5.2.2. In this study, we propose to base on the two sources of information for the data classification purpose. The first one is the lifetimes of every units. Since for training data set, the observations are complete from the beginning to the failure, the lifetime of an unit is equivalent to the length of the data sequence corresponding to that unit. Furthermore, to avoid the case that the two units could have the same lifetimes but with different fault propagation trajectories, another information is used which is the "area under curve" (AUC). This can be done by approximating, i.e. fitting the propagation trajectory of each unit by a two-degree polynomial and than calculate the area between this polynomial and the two axis (c.f. Figure 5.13). Based on the lifetimes and the AUC information, we can implement the "k-means" technique in order to classify the training data set into 2, 3 or 4 groups respectively described in Table 5.2.

Given the training data groups, the number of hidden states for each branch N as well as the number of mixture components K are determined via the Bayesian Information Criterion (BIC) as presented in Section 4.1.1.2. For example, Figure 5.14 illustrates the BIC values corresponding to different values of N and K, computed for the case of co-existence of 4 deterioration modes. From this figure, we obtain $N = 7$ and $K = 2$.

Now, corresponding to each case of M, we train the MB-HsMM model via the parameter re-estimation procedure presented in Section 5.2.3. The learned model is then used for estimating the RUL of the 218 units from the test data set. The results are evaluated through the three scores presented in 5.4.2.1. Table 5.3 shows the score results for different cases of number of deterioration modes.

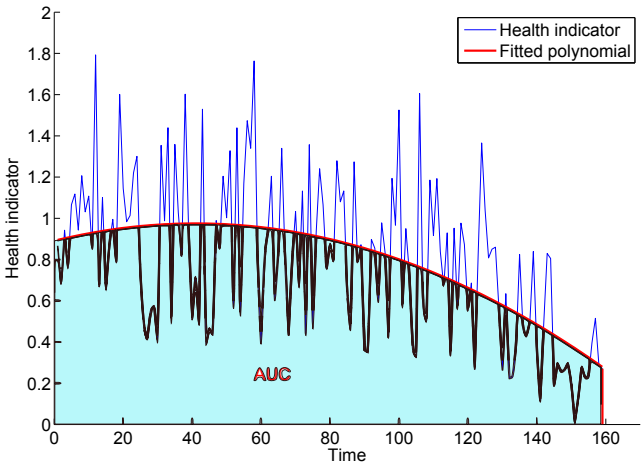


Figure 5.13: Illustration of the area under curve for an unit

In this table, the scores obtained for the "1-branch" HsMM model, which is the standard HsMM one, is also introduced. Moreover, for comparison purpose, the scores obtained with the Wiener process-based methods in [80] and the Gamma process-based method in [79] are given. There are two points that could be concluded:

- Firstly, concerning the scores obtained with the MB-HsMM models: the score decreases with the increase of number of branches M. This implies that the more deterioration modes that are taken into consideration, the better RUL estimation performance can be obtained. However, it should be also noted that the model complexity will increase with the increase of the number of branches. For this reason, a good choice of M is hence crucial and should be carried out by taking into account the compromise between the RUL estimation result and the model complexity. The scores obtained in Table 5.3 also demonstrate the advantages of the multi-branch model in comparison with the standard mono-branch one.
- Secondly, in this case study, the best RUL estimation performance is given by the four-branch HsMM model ($S = 3791$). Noting that by basing on the Hidden semi-

Table 5.3: RUL estimation scores result

Method	Score	RSE	MSE
1-branch HsMM	12246	502	1157
2-branch HsMM	6456	451	936
3-branch HsMM	**5458**	**410**	**773**
4-branch HsMM	**3791**	**389**	**694**
Wiener-based method in [80]	5575	423	823
Gamma-based method in [79]	4107	434	864

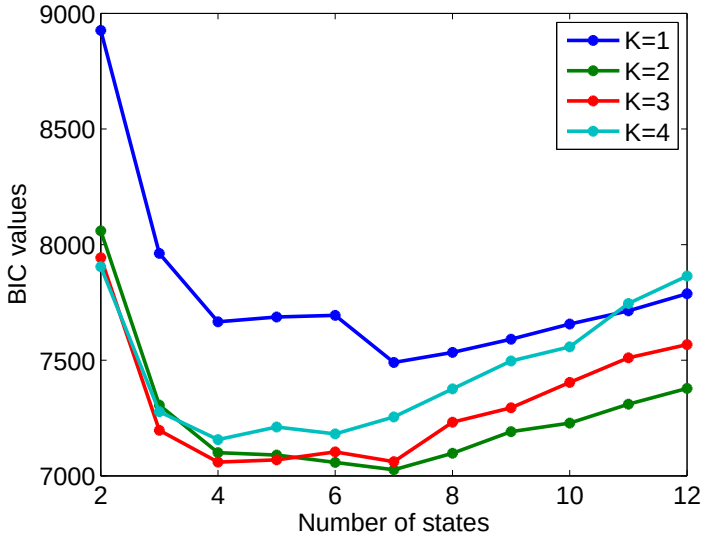


Figure 5.14: BIC values for the 4 modes case

Markov model, we are approximating and representing the continuous temporal evolution of health indicators by discrete states. Compared to the results obtained by the methods based on continuous stochastic processes, such as the Wiener process [80] or the non homogeneous Gamma process [79], the better scores given by the MB-HsMM model with 3 or 4 branches demonstrate once again the benefits of the proposed multi-branch modeling approach.

5.5 Chapter summary

This chapter extends the concept of the MB-HMM model to a more general one: the MB-HsMM model. The key idea is to allow the sojourn time staying in a state of the model to follow any arbitrary distributions (including non-parametric or parametric ones). This helps to relax the "Markovian" property of the HMM-based models but it also makes the model being more complex, leading to the difficulties in training the model parameters. To tackle this problem, a modified version of the Forward-backward algorithm introduced by Yu and Kobayashi in [166, 167] was represented and adapted for the case where the observations and the state sojourn times are continuous. From this procedure, the training and the decoding problems can be solved for the HsMM model.

By modeling several deterioration modes and through the investigation of a case study, the multi-branch model showed that it gives very promising results in RUL estimation compared to a mono-mode model, both in discrete or continuous states.

Chapter 6

Jump Markov Linear Systems for deterioration modeling

6.1 Introduction

In the two last chapters, we introduced the multi-branch model concept in order to deal with the co-existing problem of multiple modes of deterioration. Based on the Markov and semi-Markov models, the MB-HMM and MB-HsMM models as well as the associated RUL estimation methods were developed. Through several numerical studies, it is shown that by taking into account and distinguishing of several deterioration modes, the multi-branch models could give better results in RUL estimation compared to the standard "mono-branch" HMM and HsMM models respectively.

However, there are still two important limitations existing for these "multi-branch" models. Firstly, by basing on the Markov and semi-Markov chains, the health of equipment are represented by discrete states. Although this assumption leads to easy-to-interpret diagnostics results for the maintenance personnel (i.e. the equipment is currently in deteriorated state level 1, level 2, etc.), almost engineering assets deteriorate continuously in time. The continuous-state models can be therefore more appropriate for modeling the deterioration process in such cases. The second limitation of the Markovian-based multi-branch models developed in Chapter 4 and Chapter 5 comes from their assumption that the deterioration modes are exclusive once initiated. Although it can be true in some cases, i.e. in the case of the apparition and development of a crack, this assumption often cannot be hold in several real-life systems. This limits hence the applications of the MB-HMM and MB-HsMM in practice. In effect, nowadays, practical systems usually operate in dynamic environment. The deterioration process is hence affected by several external factors such as the temperature, the humidity, the applied load, etc. Such factors are called the co-variables and they could vary over time, leading to the changes in deterioration dynamic. In such cases, allowing the mode transitions could help to model more precisely and approach closely the real deterioration phenomena.

In this chapter, we extend the multi-branch models for the continuous-state cases with the allowance of "switching" between the constituent branches during the equipment's life. Firstly, we base on the state-space model for representing the temporal evolution of the continuous health states under one deterioration mode. The state-space model is formulated by two equations: one describing the system state evolution and the other one

representing the relationship between the observations and these states. Noting that the Hidden Markov Model is itself a special case of the state-space model where the states are discrete. The state-space model is hence a suitable tool for modeling the deterioration processes in the continuous-state cases [8, 39, 140, 171]. Secondly, similar to the extensions done for the MB-HMM and MB-HsMM models, several individual state-space models can be used and combined into one unified model in order to represent the coexistence of multiple deterioration modes. Each constituent model corresponds to a deterioration mode. To allow the transition between these models, an extra discrete random variable is added. This variable is assumed to follow a Markov chain and it governs the manner of switching between the individual models.

More specifically, suppose that there exists M different deterioration modes and let $t \in \{1, 2, \ldots\}$ denote the discrete time. Under each mode, the dynamic of the deterioration process is represented by a state-space model:

$$\begin{aligned} &\text{Mode 1:} && x_t = f^{(1)}\left(x_{t-1}, \omega_{t-1}^{(1)}\right) && y_t = g^{(1)}\left(x_t, \nu_t^{(1)}\right) \\ & && \ldots \\ &\text{Mode M:} && x_t = f^{(M)}\left(x_{t-1}, \omega_{t-1}^{(M)}\right) && y_t = g^{(M)}\left(x_t, \nu_t^{(M)}\right) \end{aligned} \tag{6.1}$$

where x_t and y_t are the state and observation vectors at time t and ω_t and ν_t represent the process and measurement noises. The superscript implies the underlying mode.

To incorporate these equations into one unified model, we use a discrete variable S that takes the values in $\{1, 2, \ldots, M\}$ and denote q_t its realization at time t. Equation (6.1) becomes:

$$\begin{aligned} x_t &= f^{(q_t)}\left(x_{t-1}, \omega_{t-1}^{(q_t)}\right) \\ y_t &= g^{(q_t)}\left(x_t, \nu_t^{(q_t)}\right) \end{aligned} \tag{6.2}$$

In the literature, this type of model is called the "switching state-space model" (Switching SSM), a state-space model with regimes switching that has been widely used in the control engineering and economic communities [71, 46]. Graphically, this model can be represented as a case of the dynamic bayesian network (DBN) [107] as shown in Figure 6.1.

Recently, the switching SSM model has been studied as a tool for modeling the deterioration processes as well as for RUL estimation in the literature. For example, Luo *et al.* applied the multiple models for modeling the deterioration of a half-car suspension system [97]. The authors supposed that the system could deteriorate according to three different modes under three road conditions respectively and used the Interacting Multiple Model (IMM) filter to estimate the probability of each mode at a given time. Based on this tracking result, the remaining life time was estimated [97]. More recently, Compare *et al.* implemented the switching SSM model for taking into account different

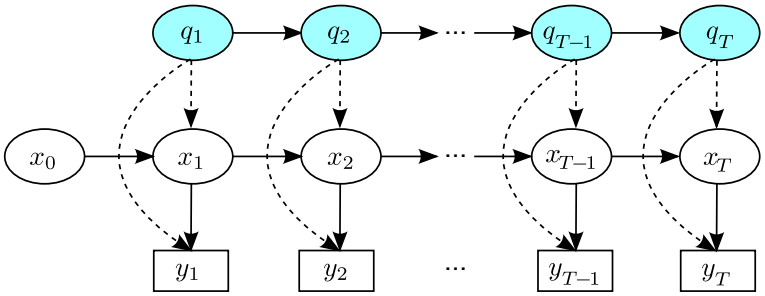


Figure 6.1: Graphical representation of the switching state-space model

deterioration mechanisms of a component and proposed a particle filter based method for fault diagnostics [23]. However, the authors did not deal with the prognostics problem in their paper. It is worth noting that almost all of works on the switching SSM model for prognostics in the literature have assumed that the model parameters are already known. This assumption can not be hold in several practical situations. In this chapter, we concentrate not only on the RUL estimation problem, but also on the learning of the model parameters.

Generally, the state and observation equations in model (6.2) are non-linear, which are powerful in describing a broad range of dynamic systems. However, their analysis and identification might be relatively complex and unfeasible [39]. In some practical cases, it may be sufficient to approximate the model (6.2) by a more simple one, i.e. by its linearized version. This leads to a model that is common used in the control systems community called the "Jump Markov Linear Systems" (JMLS) [25]. For the seek of simplicity, this linear model will be investigated in this chapter for modeling the deterioration processes.

In the next section, we first review the background of the JMLS model as well as address the problem of parameters learning. We then introduce an approximated Expectation-Maximization algorithm to overcome this problem in Section 6.3. After that, based on the learned model, the diagnostics and prognostics tasks in the context of the predictive maintenance are presented in Section 6.4. To evaluate the performance of the proposed model, a numerical study is implemented in Section 6.5.

6.2 Jump Markov Linear Systems (JMLS)

6.2.1 Model formulation

Jump Markov Linear Systems (JMLS) share the same form as linear state-space models but with discrete-values time-variant parameters. Let $t \in \{1, 2, \ldots\}$ denote the discrete

time, a general JMLS can be represented by:

$$\begin{aligned} x_t &= A^{(q_t)} x_{t-1} + \omega_t^{(q_t)} \\ y_t &= C^{(q_t)} x_t + \nu_t^{(q_t)} \end{aligned} \tag{6.3}$$

where x_t and y_t are the continuous hidden state and the observation at time t respectively; ω_t and ν_t are the state and measurement noises that, together with the initial continuous state x_0, are assumed to be independently Gaussian distributed:

$$\begin{aligned} \omega_t^{(q_t)} &\sim \mathcal{N}\left(0, Q^{(q_t)}\right) \\ \nu_t^{(q_t)} &\sim \mathcal{N}\left(0, R^{(q_t)}\right) \\ x_0 &\sim \mathcal{N}\left(\mu_0, \Sigma_0\right) \end{aligned} \tag{6.4}$$

The main difference of the JMLS model from the traditional linear state-space one is that its parameters are not fixed but change over time depending on the realization of the variable S. This variable is added for representing the transition between different "regimes". In effect, denote q_t a realization at time t of S, we can see in the above equations that the state as well as the observation dynamics are dependent on the value of q_t. Assuming that there are M "regimes" in total, denoted by M discrete symbols $\{m_1, m_2, \ldots, m_M\}$, the temporal evolution of the variable S can be modeled by a discrete-time, homogeneous, first-order Markov chain with initial distribution π_1 and transition probability matrix Π where:

$$\Pi\left(i,j\right) \triangleq \mathbb{P}\left(q_{t+1} = m_j \mid q_t = m_i\right) \quad \forall i,j \in \{1,2,\ldots,M\} \tag{6.5}$$

For notation convenience, we denote A_i, C_i, Q_i, R_i for implying $A^{(m_i)}, C^{(m_i)}, Q^{(m_i)}, R^{(m_i)}$ respectively in the subsequent sections of this chapter.

The parameters vector of the JMLS model is hence denoted by:

$$\Theta = \left\{\left(A_i, C_i, Q_i, R_i\right)_{i=1,\ldots,M}, \mu_0, \Sigma_0, \Pi, \pi_1\right\} \tag{6.6}$$

6.2.2 Identifiability issue

An essential task while implementing a mathematical model like the JMLS one to represent the dynamics of a system is to learn its parameters from the observed data. A crucial issue in the design of any learning algorithms is to study the model's identifiability [68, 148, 150]. In this section, a discussion about the identifiability of the JMLS model is given.

The identifiability is the capacity of learning the true value of model's underlying parameter after obtaining an infinite number of observations. The identifiability study seeks to answer the uniqueness question of the model to be learned: Given observations

$\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$ generated by the JMLS model Θ, what is the set of models that may produce the same output sequence $\mathbf{O}$? This question is crucial since if the set contains several different models which yield equal likelihood, any inference algorithms could converge to any point within the set.

Concerning the case of the JMLS model, it is well known that all minimal state-space representations equivalent from an input-output point of view can be deduced one from another by a state-space similarity transformation [68, 150]. As pointed out by Vidal *et al.* in [148], given observations alone, there exists infinite systems that generate them, which differ in the trajectories of both discrete and continuous states and model parameters. This means that the model that can generate the given observations is not unique.

For this reason, constraints must be imposed on the model structure in order to guarantee its identifiability [148]. Kalawoun *et al.* [68] recently show that the parameters of the JMLS models are globally structurally identifiable if the two followings constraints are satisfied:

1. A prior information at $t_0 = 0$ is available: $y_0 = C^{(q_0)} x_0$, where x_0 and q_0 are known
2. $\forall\,(i,j),\ \Pi\,(i,j) \neq 0$

The first condition is equivalent to imposing a constraint on the parameter C while the second one implies that the Markov chain used to describe the evolution of the discrete variable S must be irreducible and aperiodic. To guarantee the identifiability of the JMLS model, we suppose $C_i = 1$ for $\forall i = 1, 2, \ldots, M$ in the subsequent sections of this chapter.

6.2.3 Parameters learning problem

Let $\mathcal{Y}_t = \{y_1, \ldots, y_t\}$, $\mathcal{X}_t = \{x_0, \ldots, x_t\}$ and $\mathcal{S}_t = \{q_1, \ldots, q_t\}$ denote the sequences of observations, continuous state and discrete state values until the time step t respectively. The model learning problem is to determine the parameters Θ that maximize the likelihood function $\mathbb{P}\,(\mathcal{Y}_T \mid \Theta)$ given a finite sequence of observations $\mathcal{Y}_T$.

As $\mathcal{X}_T$ and $\mathcal{S}_T$ are unobservable, the expectation-maximization (EM), an iterative algorithm for finding a local maximum of a likelihood function, stands out to be a suitable tool for model learning in such incomplete data problem [29]. The EM algorithm calculates, in the first "E-step", the expected value of the complete-data log-likelihood function with respect to unknown data and the current parameters estimate, that is:

$$\mathcal{Q}\left(\Theta \mid \Theta^{(k)}\right) = \mathbf{E}\left[\log \mathbb{P}\,(\mathcal{X}_T, \mathcal{S}_T, \mathcal{Y}_T \mid \Theta) \mid \mathcal{Y}_T, \Theta^{(k)}\right] \tag{6.7}$$

where $\Theta^{(k)}$ is the parameters estimate at the iteration k. Then in the second step, namely the "M-step", the expectation computed in the first step is maximized to find a new

estimation of the parameters. In other words, we find:

$$\Theta^{(k+1)} = \arg\max_{\Theta} \mathcal{Q}\left(\Theta \mid \Theta^{(k)}\right) \tag{6.8}$$

These two steps are repeated until a convergence criterion is satisfied. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function [10].

It is worth noting that the expectation in (6.7) is calculated over the conditional distribution $p\left(\mathcal{X}_T, \mathcal{S}_T \mid \mathcal{Y}_T, \Theta^{(k)}\right)$ of hidden variables given the observation and under the current parameters estimate. If there were no switching dynamics, i.e. in case of linear dynamical systems (LDS), this conditional distribution could be evaluated thanks to an Rauch-Tunng-Striebel (RTS) smoother [135]. However, due to the presence of switching dynamics embedded in the transition matrix Π, the E-step becomes intractable in most cases for the JMLS models [46, 107]. An approximation is therefore necessary for computing the conditional distribution of the hidden states $p\left(\mathcal{X}_T, \mathcal{S}_T \mid \mathcal{Y}_T, \Theta^{(k)}\right)$. Regarding this problem, several approaches have been proposed in the literature [106], such as Generalized Pseudo Bayesian algorithm (GPB) [70], Interacting Multiple Model algorithm (IMM) [13], variational method [46] or MCMC sampling [34], etc. Inspired from the "pruning" technique introduced in [120, 159, 12], we adapt in the next section a Viterbi approximation-based EM approach for estimating the parameters of the JMLS models.

6.3 Approximated EM algorithm for JMLS learning

To explain the main idea of the algorithm, we first rewrite the $\mathcal{Q}$ function in (6.7) by:

$$\begin{aligned}\mathcal{Q}\left(\Theta \mid \Theta^{(k)}\right) &= \mathbf{E}_{\mathcal{S}_T|\mathcal{Y}_T,\Theta^{(k)}}\left[\mathbf{E}_{\mathcal{X}_T|\mathcal{S}_T,\mathcal{Y}_T,\Theta^{(k)}}\left[LL\right]\right] \\ &= \sum_{\mathcal{S}_T}\left(\mathbb{P}\left(\mathcal{S}_T \mid \mathcal{Y}_T, \Theta^{(k)}\right)\int p\left(\mathcal{X}_T \mid \mathcal{S}_T, \mathcal{Y}_T, \Theta^{(k)}\right)\log \mathbb{P}\left(\mathcal{X}_T, \mathcal{S}_T, \mathcal{Y}_T \mid \Theta\right) d\mathcal{X}_T\right)\end{aligned} \tag{6.9}$$

where $LL = \log \mathbb{P}\left(\mathcal{X}_T, \mathcal{S}_T, \mathcal{Y}_T \mid \Theta\right)$ is the complete-data log-likelihood.

As the expectation is computed by considering all possible sequences of discrete states $\mathcal{S}_T$, doing this is intractable in practice since the number of possible paths $\mathcal{S}_T$ increases exponentially along with the increasement of time steps. Based on the "pruning" principle, the main idea here is to restrict the space of paths over which the expectation (6.9) is carried out, i.e. by removing all the "low likelihood" sequence of $\mathcal{S}_T$ and retaining only the most likely one [159]. As pointed out in [12], although this approximation does not guarantee the convergence of the EM algorithm, it is still sufficient in most cases and, most importantly, the algorithm is linear in the number of time steps. Inspired from the decoding problem of the Hidden Markov Model, we adapt the Viterbi algorithm, a

technique based on the dynamic programming methods [129], to find out a single most likely state sequence for the JMLS model given the observations and then approximate the expectation in (6.9) by calculating the sum over this single estimated sequence.

6.3.1 Viterbi-based E-step

Applying the principle of the Viterbi algorithm for the case of the JMLS model, we first define, at each time step t, the best "partial cost" as follows [120]:

$$J_t(i) = \max_{\mathcal{S}_{t-1}, \mathcal{X}_t} \log \mathbb{P}\left(\mathcal{X}_t, \mathcal{Y}_t, \mathcal{S}_{t-1}, q_t = m_i\right) \tag{6.10}$$

i.e. among all possible sequences of continuous states $\mathcal{X}_t$ and discrete states $\mathcal{S}_t$ that ends in m_i, $J_t(i)$ is the best score for the complete-data log-likelihood $\log \mathbb{P}\left(\mathcal{X}_t, \mathcal{Y}_t, \mathcal{S}_t \mid \Theta\right)$. Compared to the traditional Viterbi algorithm applied for the discrete state HMM model, the partial cost $J_t(i)$ plays the same role as the quantity $\delta_t(i)$ which is the best score along a single path at time t accounting for the first t observations and ends in state S_i [128]. To calculate this cost recursively, we firstly investigate the calculation of the complete-data log-likelihood $\log \mathbb{P}\left(\mathcal{X}_t, \mathcal{Y}_t, \mathcal{S}_t \mid \Theta\right)$.

Based on the Markovian property of the model (6.3) and from the Bayesian theorem, the complete-data likelihood at time T of the JMLS model can be evaluated as follows:

$$\begin{aligned}\mathbb{P}\left(\mathcal{X}_T, \mathcal{Y}_T, \mathcal{S}_T \mid \Theta\right) &= \mathbb{P}\left(\mathcal{Y}_T \mid \mathcal{X}_T, \mathcal{S}_T, \Theta\right) \cdot \mathbb{P}\left(\mathcal{X}_T \mid \mathcal{S}_T, \Theta\right) \cdot \mathbb{P}\left(\mathcal{S}_T \mid \Theta\right) \\ &= \mathbb{P}_\Theta\left(x_0\right) \prod_{t=1}^{T} \mathbb{P}_\Theta\left(x_t \mid x_{t-1}, q_t\right) \prod_{t=1}^{T} \mathbb{P}_\Theta\left(y_t \mid x_t, q_t\right) \prod_{t=2}^{T} \Pi\left(q_{t-1}, q_t\right)\end{aligned} \tag{6.11}$$

Note that from the Gaussian assumption of the state and measurement noises, it follows that the conditional probability distributions of x_t and y_t given the value of x_{t-1} and q_t are also Gaussian ones. In other words:

$$\begin{aligned} x_t \mid x_{t-1}, q_t = m_i &\sim \mathcal{N}\left(A_i x_{t-1}, Q_i\right) \\ y_t \mid x_t, q_t = m_i &\sim \mathcal{N}\left(C_i x_t, R_i\right) \end{aligned}$$

Therefore, the complete-data log-likelihood in (6.11) can be expressed by:

$$\begin{aligned}\log \mathbb{P}\left(\mathcal{X}_T, \mathcal{Y}_T, \mathcal{S}_T\right) \sim & -\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{M}\left[\left(x_t - A_i x_{t-1}\right)' Q_i^{-1}\left(x_t - A_i x_{t-1}\right) + \log\left|Q_i\right|\right]\mathbf{1}_{\{q_t=m_i\}} \\ & -\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{M}\left[\left(y_t - C_i x_t\right)' R_i^{-1}\left(y_t - C_i x_t\right) + \log\left|R_i\right|\right]\mathbf{1}_{\{q_t=m_i\}} \\ & -\frac{1}{2}\left[\left(x_0 - \mu_0\right)'\Sigma_0^{-1}\left(x_0 - \mu_0\right) + \log\left|\Sigma_0\right|\right] + \sum_{t=2}^{T}\log\Pi\left(q_{t-1}, q_t\right) + \log\pi_1\left(q_1\right)\end{aligned}$$

where x' denote the transpose of the vector x and $\mathbf{1}_{\{\}}$ is the indicator function.

In the case of the discrete state HMM model, the score $\delta_{t+1}(i)$ at time $t+1$ is calculated by examining all possible transitions of the hidden state from S_j to S_i and retaining the transition that has the best score. For the JMLS model, we also associate an "innovation cost" for each of $m_j \to m_i$ transition by taking into account at the same time: i) the cost that reflects the LDS state transition (i.e. of x_t and y_t) and ii) the cost due to discrete state switching from m_j to m_i. The "innovation cost" is given in Equation (6.12).

$$\begin{aligned}J_{t,t+1}^{j,i} = & -\frac{1}{2}\left(y_t - C_i x_{t+1|t,j,i}\right)'\left(C_i\Sigma_{t+1|t,j,i}C_i' + R_i\right)^{-1}\left(y_t - C_i x_{t+1|t,j,i}\right) \\ & -\frac{1}{2}\log\left|C_i\Sigma_{t+1|t,j,i}C_i' + R_i\right| + \log\Pi\left(j,i\right)\end{aligned} \tag{6.12}$$

where $\hat{x}_{t+1|t,j,i}$ and $\Sigma_{t+1|t,j,i}$ are the one-step predicted mean and variance of continuous state X at time t respectively, given that the switch transits from state m_j to state m_i at time t.

$$\begin{aligned}\hat{x}_{t+1|t,j,i} &= \mathbf{E}\left[x_{t+1} \mid \mathcal{Y}_t, q_t = m_j, q_{t+1} = m_i\right] \\ \Sigma_{t+1|t,j,i} &= \mathbf{E}\left[\left(x_{t+1} - \hat{x}_{t+1|t,j,i}\right)\left(x_{t+1} - \hat{x}_{t+1|t,j,i}\right)' \mid \mathcal{Y}_t, q_t = m_j, q_{t+1} = m_i\right]\end{aligned}$$

These quantities can be calculated through a Kalman filter applying for the JMLS model. For this end, we further define the following mode-dependent states and variances terms:

$$\begin{aligned}\hat{x}_{t|t,i} &= \mathbf{E}\left[x_t \mid \mathcal{Y}_t, q_t = m_i\right] \\ \Sigma_{t|t,i} &= \mathbf{E}\left[\left(x_t - \hat{x}_{t|t,i}\right)\left(x_t - \hat{x}_{t|t,i}\right)' \mid \mathcal{Y}_t, q_t = m_i\right] \\ \hat{x}_{t|t,j,i} &= \mathbf{E}\left[x_t \mid \mathcal{Y}_t, q_{t-1} = m_j, q_t = m_i\right] \\ \Sigma_{t|t,j,i} &= \mathbf{E}\left[\left(x_t - \hat{x}_{t|t-1,i,j}\right)\left(x_t - \hat{x}_{t|t-1,i,j}\right)' \mid \mathcal{Y}_t, q_{t-1} = m_j, q_t = m_i\right]\end{aligned}$$

where $\hat{x}_{t|t,i}$ and $\Sigma_{t|t,i}$ is the best filtered estimates for mean and variance of the continuous state respectively at time t when the switch variable is in state m_i at t given the sequence of measurement $\mathcal{Y}_t$. similarly, $\hat{x}_{t|t,i,j}$ and $\Sigma_{t|t,i,j}$ are the best filtered estimates at time t given that the switch transits from state m_j to state m_i at time t.

Consider the transition $m_j \to m_i$ of the discrete variable S at time t, the following prediction equations are hold based on the theory of the Kalman filter:

$$\begin{aligned}\hat{x}_{t+1|t,j,i} &= A_i \hat{x}_{t|t,j} \\ \Sigma_{t+1|t,j,i} &= A_i \Sigma_{t|t,j} A_i' + Q_i\end{aligned} \tag{6.13}$$

The filtered quantities are updated by:

$$\begin{aligned}\hat{x}_{t+1|t+1,j,i} &= \hat{x}_{t+1|t,j,i} + K_{j,i}\left(y_t - C_i \hat{x}_{t+1|t,j,i}\right) \\ \Sigma_{t+1|t+1,j,i} &= \Sigma_{t+1|t,j,i} - K_{j,i} C_i \Sigma_{t+1|t,j,i}\end{aligned}$$

where $K_{j,i}$ is the Kalman gain corresponding to the transition $m_j \to m_i$ of the switch variable.

Given $J_t(j)$ at time t for $\forall j = 1, \ldots, M$, the partial cost $J_{t+1}(i)$ at time $t+1$ is hence calculated by examining all possible transitions that can occur to reach the regime m_i. Each of these $m_j \to m_i$ transitions has a certain innovation cost $J^{j,i}_{t,t+1}$ associated with it. The "best" *partial cost* corresponding to state m_i at time $t+1$ is therefore selected by:

$$J_{t+1}(i) = \max_j \left(J^{j,i}_{t,t+1} + J_t(j)\right) \tag{6.14}$$

Figure 6.2 illustrates the computation of this partial cost at time $t+i$ for the regime m_i.

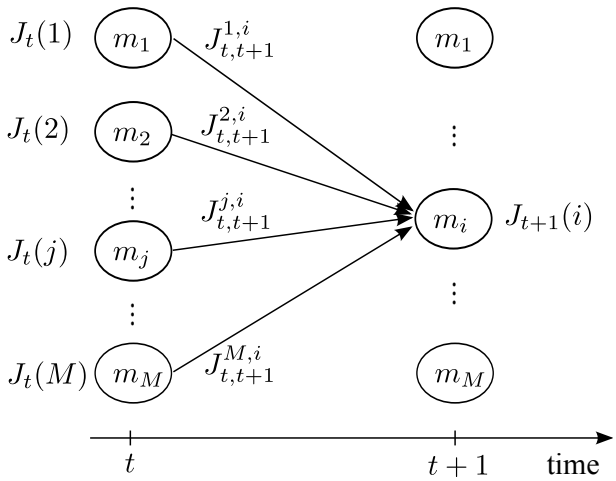


Figure 6.2: Illustration for the computation of the partial cost $J_{t+1}(i)$

Similar to the traditional Viterbi algorithm, we make use of the array $\psi_t(i)$ to keep track of the argument that maximizes (6.14) for each t and i. We have:

$$\psi_t(i) = \arg\max_j \left(J^{j,i}_{t,t+1} + J_t(j)\right) \tag{6.15}$$

Consequently, the best filtered mean $\hat{x}_{t+1|t+1,i}$ and variance $\Sigma_{t+1|t+1,i}$ at time $t+1$ are the ones that correspond to the argument kept in $\psi_t(i)$:

$$\hat{x}_{t+1|t+1,i} = \hat{x}_{t+1|t+1,\psi_t(i),i}$$
$$\Sigma_{t+1|t+1,i} = \Sigma_{t+1|t+1,\psi_t(i),i}$$

The above procedure is repeated for all time steps $t = 1, \ldots, T$. At time T, the best overall cost can be chosen by:

$$J_T^* = \max_i J_{T,i} \tag{6.16}$$

By a backtracking step as used in the traditional Viterbi algorithm, the "best" switching state sequence can be decoded.

Denote $\mathcal{S}_T^*$ the most likely state sequence at time T, the approximated $\mathcal{Q}$ can be computed by:

$$\mathcal{Q}\left(\Theta \mid \Theta^{(k)}\right) \approx \int p\left(\mathcal{X}_T \mid \mathcal{S}_T^*, \mathcal{Y}_T, \Theta^{(k)}\right) \log \mathbb{P}\left(\mathcal{X}_T, \mathcal{S}_T^*, \mathcal{Y}_T \mid \Theta\right) d\mathcal{X}_T \tag{6.17}$$

To evaluate this function, it is sufficient to carry out an Rauch-Tung-Streiber (RTS) smoother [130] to estimate the smoothed quantities $p\left(\mathcal{X}_T \mid \mathcal{S}_T^*, \mathcal{Y}_T, \Theta^{(k)}\right)$. The complete-data log-likelihood function $\log \mathbb{P}\left(\mathcal{X}_T, \mathcal{S}_T^*, \mathcal{Y}_T \mid \Theta\right)$ can be evaluated by (6.11).

6.3.2 M-step

To estimate the parameters Θ of the JMLS, one can take the derivatives of the $\mathcal{Q}$ function in (6.17) with respect to every parameter and set them to zero. Based on a single sequence of switching states estimated in the E-step, the equations for parameters estimation are a generalization of the ones for the classical (non-switching) linear dynamical systems [120]. These parameter update equations can also be generalized for the case of multiple observation sequences as follows.

Suppose that there are K observation sequences, the update equations for parameters

concerning the continuous states of the JMLS models are given by:

$$\begin{aligned}
\hat{A}_i &= \mathcal{S}_{10,i}\,\mathcal{S}_{00,i}^{-1} \\
\hat{Q}_i &= \frac{1}{\sum_{k=1}^{K} T_i^{(k)}}\,\mathcal{S}_{11,i} - \hat{A}_i \mathcal{S}_{10,i}' \\
\hat{R}_i &= \frac{1}{\sum_{k=1}^{K} T_i^{(k)}}\,\mathcal{S}_{22,i} - \hat{C}_i \mathcal{S}_{20,i}' \\
\hat{\mu}_0 &= \frac{1}{K}\sum_{k=1}^{K} \hat{\mu}_0^{(k)} \\
\hat{\Sigma}_0 &= \frac{1}{K}\sum_{k=1}^{K}\left[\hat{P}_0^{(k)} + \hat{\mu}_0^{(k)}\left(\hat{\mu}_0^{(k)}\right)'\right] - \hat{\mu}_0\left(\hat{\mu}_0\right)'
\end{aligned}$$

where the subscript i and the superscript (k) represent the corresponding mode m_i and kth observation sequence respectively and

$$\begin{aligned}
\mathcal{S}_{11,i} &= \sum_{k=1}^{K}\sum_{t\in\mathcal{F}_i(S_T^*)}\left(\hat{x}_{t|T}^{(k)}\left(\hat{x}_{t|T}^{(k)}\right)' + \hat{\Sigma}_{t|T}^{(k)}\right) \\
\mathcal{S}_{10,i} &= \sum_{k=1}^{K}\sum_{t\in\mathcal{F}_i(S_T^*)}\left(\hat{x}_{t|T}^{(k)}\left(\hat{x}_{t-1|T}^{(k)}\right)' + \hat{\Sigma}_{t|T}\right) \\
\mathcal{S}_{00,i} &= \sum_{k=1}^{K}\sum_{t\in\mathcal{F}_i(S_T^*)}\left(\hat{x}_{t-1|T}^{(k)}\left(\hat{x}_{t-1|T}^{(k)}\right)' + \hat{\Sigma}_{t-1|T}\right) \\
\mathcal{S}_{22,i} &= \sum_{k=1}^{K}\sum_{t\in\mathcal{F}_i(S_T^*)} y_t^{(k)}\left(y_t^{(k)}\right)' \\
\mathcal{S}_{20,i} &= \sum_{k=1}^{K}\sum_{t\in\mathcal{F}_i(S_T^*)} y_t^{(k)}\left(\hat{x}_{t|T}^{(k)}\right)'
\end{aligned}$$

where $\mathcal{F}_i(S_T)$ is the set of time steps in the sequence S_T for which the switching state is m_i (members of $\mathcal{F}_i(S_T)$ are integers in the range $[1, T]$). Given this "best" sequence, $\hat{x}_{t|T}$ and $\hat{\Sigma}_{t|T}$ are the mean value and covariance matrix of the continuous state that are estimated via the Rauch-Tung-Streiber smoother [130].

The discrete part of the parameters can be estimated by applying the same principle of the Hidden Markov Model [129], that is:

$$\begin{aligned}
\hat{\pi}_1(i) &= \text{number of times in state } i \text{ at time } (t=1) \\
\hat{\Pi}(i,j) &= \frac{\text{number of transitions from state } i \text{ to state } j}{\text{number of transitions from state } i}
\end{aligned}$$

Denote e_i an unit vector of dimension M with a non-zero element in the ith position, we

obtain:

$$\hat{\pi}_1 = \frac{1}{K}\sum_{k=1}^{K} e_{S_T^{*(k)}(t=1)}$$

$$\hat{\Pi} = \frac{1}{K}\sum_{k=1}^{K}\left(\sum_{t=1}^{T^{(k)}}\xi_t^{(k)}\left(\xi_t^{(k)}\right)'\right)\text{diag}\left(\sum_{t=1}^{T^{(k)}}\xi_t^{(k)}\right)^{-1}$$

where $\xi_t^{(k)} = e_{S_T^{*(k)}(t)}$.

6.4 JMLS-based diagnostics and prognostics

In this section, we applied the trained JMLS model for the diagnostics and prognostics of the monitored equipment. Given a sequence of observations $\mathcal{Y}_t$ up to time t, the diagnostics task deals with: (i) detecting the current deterioration mode; and (ii) assessing the actual deterioration state x_t. After that, the RUL of the equipment will be evaluated in the prognostics step.

6.4.1 Diagnostics

Suppose that the M "regimes" m_i, $i = 1, \ldots, M$ of the JMLS model corresponds to M different modes of deterioration of the equipment, i.e if $q_t = m_i$, the equipment is said to be deteriorated in the mode m_i at time t.

To estimate the actual mode q_t, we base on the Viterbi-like procedure described in Section 6.3.1. Specifically, instead of choosing only one "best" state sequence $\mathcal{S}_t^*$, we retain, at the time t, the M sequences in which each one is the "best" among those ending in state m_i for $\forall i = 1, 2, \ldots, M$. The corresponding partial costs of these sequences are used to compute the probability of each being in each mode at time t. Denote the best state sequence that ends in m_i by $\mathcal{S}_{t,i}^*$, the probability that the mode m_i is active at time t is given by:

$$\mu_t(i) \triangleq \frac{\mathbb{P}\left(\mathcal{S}_{t,i}^*\right)}{\sum_{i=1}^{M}\mathbb{P}\left(\mathcal{S}_{t,i}^*\right)} = \frac{1}{1+\exp\left(\sum_{j\neq i} J_{t,j} - J_{t,i}\right)} \tag{6.18}$$

where $\mu_t(i) = \mathbb{P}\left(q_t = m_i\right)$.

Corresponding to each state sequence, the current health state of the equipment at time t is estimated via an RTS smoother. The overall health state and its variance are

considered as the mixture of these estimates and are given by:

$$\begin{aligned}\hat{x}_t &= \sum_{i=1}^{M} \mu_t(i)\hat{x}_{t,i} \\ \hat{\Sigma}_t &= \sum_{i=1}^{M} \mu_i(t)\Sigma_{t,i} + \sum_{i=1}^{M} \mu_t(i)\left(\hat{x}_{t,i} - \hat{x}_t\right)\left(\hat{x}_{t,i} - \hat{x}_t\right)^T \end{aligned} \tag{6.19}$$

where $\hat{x}_{t,i}$ and $\hat{\Sigma}_{t,i}$ are the mean value and the variance of x_t corresponding to the state sequence ending in m_i.

6.4.2 Prognostics

The main objective of the prognostics task in this section is to predict the RUL of the monitored equipment. We assume that the equipment fails once its health state reaches for the first time a predefined critical threshold denoted by L. In the JMLS framework, the evolution of x_t depends both on the current state as well as the future evolution of the switching variable. Since this future information is stochastic and unknown beforehand, we suppose that the stationary law of the switching variable remains unchanged on the prediction horizon.

Under the discrete time assumption, the RUL is defined as the time steps for the equipment's health state to reach the threshold L: $RUL = (\min k : x_{t+k} \geq L \mid x_t < L)$. Given the current health state and its variance estimated from the diagnostics step, the RUL can be estimated by projecting this state in the future until it exceeds L. If the future evolution of the discrete switch S is deterministic, this task can be easily accomplished by applying the Kalman predictor for the system (6.3) for a multi-step ahead prediction of x_t. However, since the discrete switch variable evaluates randomly following a chain of Markov, applying the Kalman predictor for the JMLS model will produce an M-fold increase in the number of Gaussian distributions to consider, leading the computation be intractable [70].

One solution to overcome this limitation is to approximate, i.e. by merging, at each time $t+1$, the M Gaussian distributions by only one Gaussian and then consider it as the starting point for the next prediction at time $t+2$. Specifically, denote $\mathcal{N}\left(\hat{x}_{t+1|t,i}, \hat{\Sigma}_{t+1|t,i}\right)$ the Gaussian distribution predicted at time $t+1$ under the regime m_i where $\hat{x}_{t+1|t,i}$ and $\hat{\Sigma}_{t+1|t,i}$ are its mean and variance respectively, the merged distribution is given by [13]:

$$p\left(\hat{x}_{t+1|t}\right) \approx \sum_{i=1}^{M} \mu_{t+1}(i)\mathcal{N}\left(\mathbf{x}_{t+1|t,i}, \hat{\Sigma}_{t+1|t,i}\right) \tag{6.20}$$

where $\mu_{t+1}(i)$ is the probability of mode m_i at time $t+1$.

To accomplish the procedure, the mode probabilities $\{\mu_{t+1}(i)\}_{i=1}^{M}$ must be recursively

calculated from its previous values $\{\mu_t(i)\}_{i=1}^{M}$. Based on the principle of the Interacting Multiple Model (IMM) filter [13], the recursions for the mode probabilities are given by:

$$\mu_{t+1}(i) = \frac{\mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t,i}, q_{t+1} = m_i\right) \sum_{j=1}^{M} \Pi\left(j,i\right) \mu_t(j)}{\sum_{l=1}^{M} \mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t,l}, q_{t+1} = m_l\right) \sum_{j=1}^{M} \Pi\left(j,l\right) \mu_t(j)} \tag{6.21}$$

where $\hat{y}_{t+1|t,i}, q_{t+1} = m_i = C_i \hat{x}_{t+1|t,i}$ is the innovation quantity computed from the Kalman filter.

In the case of prediction, since we do not have any new observations after the time step t, we assign $\mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t,i}, q_{t+1} = m_i\right) = 1$ and Equation (6.21) becomes:

$$\mu_{t+1}(i) = \frac{\sum_{j=1}^{M} \Pi\left(j,i\right) \mu_t(j)}{\sum_{l=1}^{M} \sum_{j=1}^{M} \Pi\left(j,l\right) \mu_t(j)} = \sum_{j=1}^{M} \Pi\left(j,i\right) \mu_t(j) \tag{6.22}$$

Figure 6.3 illustrates the prediction procedure for the case of $M = 2$ for the JMLS model.

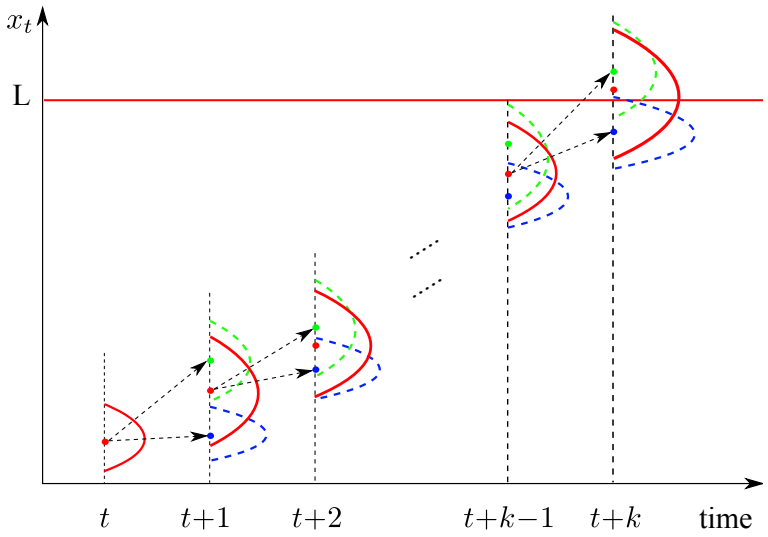


Figure 6.3: Illustration of RUL estimation for the JMLS model with $M = 2$

From the definition of the RUL and under the discrete time assumption, we have:

$$\mathbb{P}\left(RUL = i\right) = \mathbb{P}\left(\hat{x}_{t+i|t} \geq L \cap \hat{x}_{t+i-1|t} < L \cap \ldots \cap \hat{x}_t < L\right)$$

The distribution of the RUL can be calculated recursively as follows:

$$\begin{aligned}
\mathbb{P}(RUL=1) &= \mathbb{P}\left(\hat{x}_{t+1|t} \geq L \cap \hat{x}_t < L\right) \\
&= \mathbb{P}\left(\hat{x}_{t+1|t} \geq L \mid \hat{x}_t < L\right) \mathbb{P}(\hat{x}_t < L)
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(RUL=2) &= \mathbb{P}\left(\hat{x}_{t+2|t} \geq L \cap \hat{x}_{t+1|t} < L \cap \hat{x}_t < L\right) \\
&= \mathbb{P}\left(\hat{x}_{t+2|t} \geq L \mid \hat{x}_{t+1|t} < L\right) \left[1 - \mathbb{P}\left(\hat{x}_{t+1|t} \geq L \mid \hat{x}_t < L\right)\right] \mathbb{P}(\hat{x}_t < L) \\
&= \mathbb{P}\left(\hat{x}_{t+2|t} \geq L \mid \hat{x}_{t+1|t} < L\right) \left[\mathbb{P}(\hat{x}_t < L) - \mathbb{P}(RUL=1)\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(RUL=3) &= \mathbb{P}\left(\hat{x}_{t+3|t} \geq L \cap \hat{x}_{t+2|t} < L \cap \hat{x}_{t+1|t} < L \cap \hat{x}_t < L\right) \\
&= \mathbb{P}\left(\hat{x}_{t+3|t} \geq L \mid \hat{x}_{t+2|t} < L\right) \left[\mathbb{P}(\hat{x}_t < L) - \mathbb{P}(RUL=2) - \mathbb{P}(RUL=1)\right] \\
&\vdots \\
\mathbb{P}(RUL=k) &= \mathbb{P}\left(\hat{x}_{t+k|t} \geq L \cap \hat{x}_{t+k-1|t} < L \cap \ldots \cap \hat{x}_{t+1|t} < L \cap \hat{x}_t < L\right) \\
&= \mathbb{P}\left(\hat{x}_{t+k|t} \geq L \mid \hat{x}_{t+k-1|t} < L\right) \left[\mathbb{P}(\hat{x}_t < L) - \sum_{l=1}^{k-1} \mathbb{P}(RUL=l)\right]
\end{aligned}$$

In these equations, the following approximation was made:

$$\mathbb{P}\left(\hat{x}_{t+k|t} \geq L \mid \hat{x}_{t+k-1|t} < L \cap \ldots \cap \hat{x}_{t+1|t} < L \cap \hat{x}_t < L\right) \approx \mathbb{P}\left(\hat{x}_{t+k|t} \geq L \mid \hat{x}_{t+k-1|t} < L\right)$$

The fact of conditioning on restricted area of the past states, i.e. $x_{t+k-1} < L$ does not assure the Markovian property in this equation. However, this approximation becomes exact if x_t can only follow a increasing evolution, i.e. when $A_i \geq 1$, $\forall i$. Indeed, in that case, $x_{t+k-1} < L$ assures that all the remaining conditions could be hold.

To calculate the RUL, the two quantities: $\mathbb{P}(\hat{x}_t < L)$ and $\mathbb{P}\left(\hat{x}_{t+k|t} \geq L \mid \hat{x}_{t+k-1|t} < L\right)$ must be computed. Since $\hat{x}_t$ is Gaussian distributed due to the merge approximation, compute the first one is a trivial task. By assuming that $A_i > 0$ $\forall i$, the second quantity can be expressed by:

$$\begin{aligned}
\mathbb{P}\left(\hat{x}_{t+k|t} \geq L \mid \hat{x}_{t+k-1|t} < L\right) &= \mathbb{P}\left(\sum_{i=1}^{M} \mu_{t+k}(i) \cdot A_i \cdot \hat{x}_{t+k-1|t} \geq L \mid \hat{x}_{t+k-1|t} < L\right) \\
&= \mathbb{P}\left(L < \hat{x}_{t+k-1|t} \leq \frac{L}{\sum_{i=1}^{M} \mu_{t+k}(i) \cdot A_i}\right)
\end{aligned}$$

As $\hat{x}_{t+k-1|t}$ is Gaussian distributed after the merge operation, the above quantity can be easily deduced from its probability density function.

6.5 Numerical results

In this section, the performance of the JMLS model is evaluated through a simulation study. Similar to the study presented in Chapter 4, two deterioration modes corresponding to the two different propagation rates are considered: quick-rate (mode 1) and normal-rate (mode 2). These two modes may correspond to the severe or normal operating conditions of equipment. The switching variable is therefore modeled by a homogeneous Markov chain with two discrete states. The real parameters of the JMLS model are chosen as follows:

$$\begin{array}{llll} A_1 = 1.01, & C_1 = 1, & Q_1 = 0.015, & R_1 = 2.5, \\ A_2 = 1.002, & C_2 = 1, & Q_2 = 0.005, & R_2 = 0.25, \\ \mu_0 = 2, & \Sigma_0 = 0.2 & & \end{array}$$

and

$$\pi_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad \Pi = \begin{bmatrix} 0.99 & 0.01 \\ 0.02 & 0.98 \end{bmatrix}$$

Supposing that in the "severe" mode, the state process and observation noises may be greater than the ones in the "normal" mode, the parameters Q_1 and R_1 are chosen to be greater than Q_2 and R_2 respectively. These "real" parameters are used in the next sections to generate the data for both the training and testing purposes. In this study, the equipment is assumed to fail once its health state reaches for the first time the critical level $L_f = 20$.

6.5.1 Parameters estimation

To demonstrate the parameters estimation results, a total number of 50 data sequences are generated from the real JMLS model and used as the training data.

Given this training data set, the Viterbi-based EM algorithm described in Section 6.3 is implemented to re-estimate the parameters of the JMLS model. The starting values for the algorithm are randomly initiated. After 100 iterations, the following results are obtained:

$$\begin{array}{lll} \hat{A}_1 = 1.0097, & \hat{Q}_1 = 0.02, & \hat{R}_1 = 2.43, \\ \hat{A}_2 = 1.002, & \hat{Q}_2 = 0.01, & \hat{R}_2 = 0.21, \\ \hat{\mu}_0 = 2.01, & \hat{\Sigma}_0 = 0.1 & \end{array}$$

and

$$\hat{\pi}_1 = \begin{bmatrix} 0.62 \\ 0.38 \end{bmatrix}, \quad \hat{\Pi} = \begin{bmatrix} 0.994 & 0.006 \\ 0.011 & 0.989 \end{bmatrix}$$

where the hat symbol implies the estimation.

The parameters estimated are very closed to the real ones suggesting the correctness of the proposed Viterbi-based learning algorithm. Given this learned model, in the next sections, we move to an "online" phase in which both the diagnostics and the prognostics tasks are carried out.

6.5.2 Diagnostics results

The diagnostics in this study deals with the two problems: i) assessing the current health state of the monitored equipment and ii) detecting the actual deterioration mode that the equipment if following. For this purpose, from the real JMLS model, a sequence of observations are generated and used as the test data. They are shown in Figure 6.4. In the lower sub-figure, the red line represents the real evolution of the switching variable S. In this study, the equipment firstly follows mode 1 and changes to mode 2 at about 55h. It then jumps to mode 1 for a while and then comes back to mode 2 before transits to mode 1 and stays there until it fails at time $T_f = 257$h.

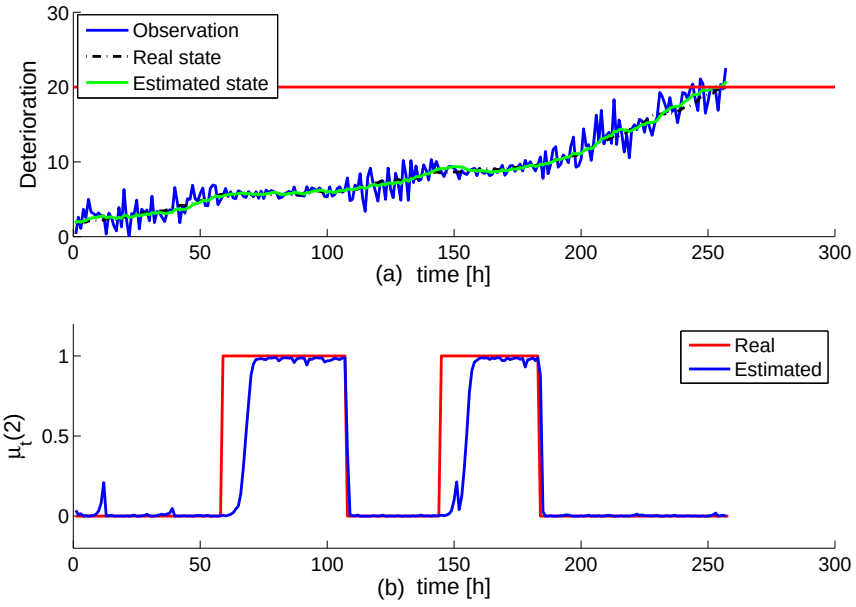


Figure 6.4: Test data

Suppose that we have no information about the continuous and discrete states. Given the noisy observations until time t, the procedure described in Section 6.4.1 is applied to estimate the mode probabilities $\mu_t(i)$ as well as continuous states x_t. As the time passes,

the procedure is repeated once more observations are available. The green line in Figure 6.4(a) and the blue line in Figure 6.4(b) represent the estimation results for x_t and $\mu_t(i)$ respectively. We can see that the continuous states were well re-estimated. Concerning the mode detection, when the transits from mode 2 to mode 1 were well detected while there are some delays that can be noticed at the transits from mode 1 to mode 2. This can be explained by the fact that under the mode 1, the continuous states propagate more quickly and the noises are higher than when the equipment is under the mode 2.

6.5.3 RUL estimation

Once the diagnostics task is accomplished, the equipment's RUL can be estimated. For this end, we assume that the stationary law of the switching variable remains unchanged on the prediction horizon. We also suppose that only the observations from the time $t = 0$ to the actual time t_{act} are available for the estimation purpose. Given the results obtained from the diagnostics stage, the procedure described in Section 6.4.2 is applied on the noisy observations in order to estimate the RUL of the equipment.

For comparison purpose, a Monte Carlo estimation is also carried out in this study. Specifically, based on the diagnostics results, the continuous state at time t_{act} is propagated until it reaches for the first time the failure threshold L_f and an estimate of RUL is obtained. By repeating this procedure a large number of times, i.e. 5000 times and by recording all the values obtained, a histogram of the RUL can be obtained.

Figure 6.5 shows the results of RUL estimation at time $t_{act} = 150$h. In the above subfigure, only the observations in red color are available for RUL estimation while the blue ones refer to the full data from the beginning to the failure of the equipment. The red line curve in the below figure represents the RUL probability mass function (pmf) obtained by the procedure described in Section 6.4.2 (hereafter considered as analytic calculation) while the RUL histogram obtained by the Monte Carlo simulation is illustrated in blue.

At time $t_{act} = 150$h, the real RUL of the equipment is $RUL_{real} = 107$h. One can notice that the Monte Carlo simulation gives better performance than the analytic calculation in this particular case. The difference between the RUL pmf and the histogram can be explained by the approximations applied in the RUL analytic calculation. In effect, in that procedure, we have merged the mixture of M Gaussian distributions by only one Gaussian after each prediction step. However, the error at this particular time does not impose a serious problem in RUL estimation as we will see in the next study, the real RUL still lies close to the 95% prediction interval.

To demonstrate the "online" estimation of the RUL, we do increase periodically t_{act} by 30h. After each increment, new observations are obtained and available and the RUL is re-estimated with the same procedure as above. Figure 6.6a shows the mean values of the RUL estimated at different time steps compared to the real ones.

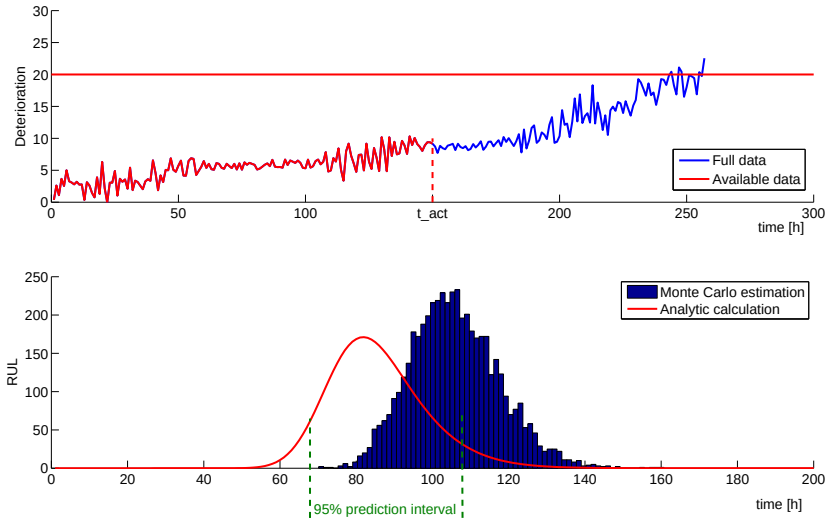


Figure 6.5: RUL estimation at $t = 150[h]$

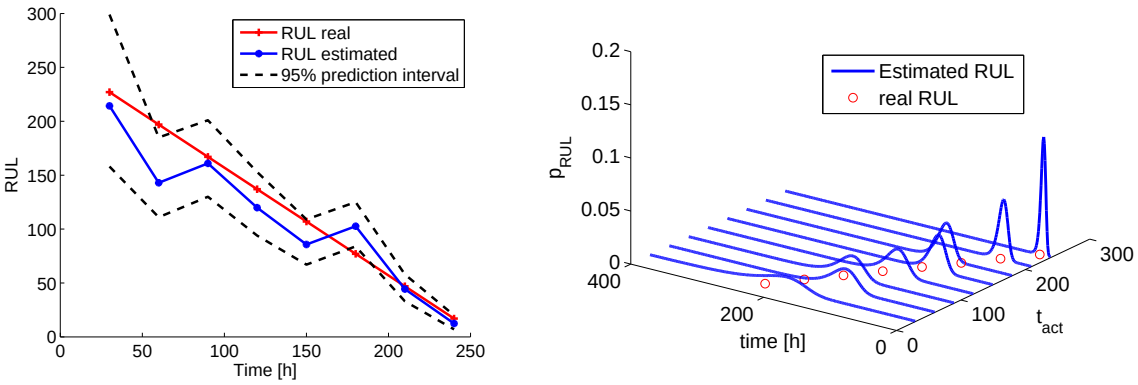
One can notice that the estimated RUL (represented by the blue line) converges to the real ones (represented by the red curve) as time passes. In addition, the real RUL values almost lie within the 95% prediction interval which demonstrates the accuracy of the proposed RUL estimation procedure. Moreover, at time $t = 150$h, we see again the difference between the estimated RUL and the real one. As already mentioned, this difference is not so important here from a point of view for the overall RUL estimation.

Figure 6.6b represents the evolution of the estimated RUL pmf in the function of time together with the real RUL values (represented by the red dots). We can see that the variance or the uncertainties in RUL estimation decreases gradually when more observations and information about the deterioration process are available.

6.6 Chapter summary

In this chapter, we have extended the multi-branch concept applied in the two last chapters for the continuous-state case. The jump Markov linear systems have been implemented for taking into account the co-existence of different deterioration modes. Compared to the conventional dynamic linear systems, this model employs an extra discrete switching variable, playing the role of a "switch". The equipment is therefore allowed to transit from one mode of deterioration to another, providing a deterioration modeling framework that is more closed to real phenomena in practice.

To estimate the parameters of the model, the expectation-maximization algorithm was carried out to obtain the maximum likelihood estimates. As the E-step is attractable for the JMLS model, a Viterbi-based approximation method is adapted. The numerical results showed the good performance of the learning algorithm.



(a) RUL at different time steps (b) Evolution of the RUL density functions

Figure 6.6: RUL estimation results

Furthermore, the diagnostics and the prognostics tasks are also considered in this chapter. Because of the co-existence of several deterioration modes, we adapted the Viterbi approximation technique to evaluate the probabilities of each mode at a given time t. The health state assessment was then done by conducting the Kalman filter and smoother under each mode. Based on this diagnostics information, a RUL estimation procedure based on the principle of the IMM filter was proposed. The numerical studies showed that the obtained results are very promising.

Conclusion and perspectives

General conclusion

With a ratio between maintenance costs and added value higher than 25%, maintenance is a major issue for all manufacturing or production plants. For continuous production industries such as food, cement or paper where ensuring continuity of production is one of the key requirements, predictive maintenance is a critical issue thanks to its ability to propose an optimal maintenance actions planning by anticipating the failure before it occurs. The damage prognostics and residual life prediction play hence the center role in implementing a predictive maintenance program.

The works presented in this manuscript deal with the deterioration modeling and remaining useful life estimation problems within the framework of the European project SUPREME. Specifically, the thesis' works were divided into two main parts. The first one gives a comprehensive review of the deterioration models and RUL estimation methods existing in the literature. By analyzing their advantages and disadvantages, the so-called incubation/propagation model was adapted to model the two-phase deterioration process found within a test bench of the project. In addition, to take into account the impacts of environmental factors (i.e. the covariates) on the deterioration processes, a load-dependent deterioration model was studied and investigated. Furthermore, some practical implementation aspects, i.e. the issue of information exchange between the project partners are also detailed in this first part.

The second part was dedicated for the development beyond the state-of-the-art in order to have some "ready-to-use" deterioration models and RUL estimation methods for potential application cases that can be found in the SUPREME framework. Specifically, we were interested in the phenomenon in which several deterioration modes co-exist in competition, even within one component. To tackle this problem, the concept of the "multi-branch" models was introduced and developed. Indeed, they are the models that consist of several branches in which each one is used to represent a deterioration mode. In the framework of this thesis, two multi-branch model types were presented corresponding to the discrete and continuous cases of the health state of the systems. In the discrete case, the so-called multi-branch Hidden Markov Model (Mb-HMM) were firstly constructed based on the Hidden Markov Model. Based on this model, a diagnostics and prognostics framework was implemented. The numerical results showed that the multi-branch HMM models, by taking into account the co-existence of multiple deterioration modes, can give better performances in RUL estimation compared to the ones obtained by standard "single branch" HMM models. These advantages become more clearly when the dynamics of the deterioration modes are very different.

One limitation of the Markov-based model is that the sojourn time in a state always

follows exponential or geometrical distributions. From a deterioration modeling point of view, this property could not hold in several practical situations. To overcome this problem, the Mb-HsMM model, an extension of the Mb-HMM model is then introduced. This model was applied to a case study and the obtained results show that even being used for representing a continuous deterioration process, the Mb-HsMM models can outperform other continuous stochastic process-based approaches in RUL estimation.

Concerning the continuous health state case, the Jump Markov Linear System (JMLS) was presented and investigated. Since the model learning by the Expectation-Maximization algorithm becomes intractable due to the existence of a switching discrete state, an approximated solution based on the Viterbi algorithm was proposed and implemented. In addition, a RUL estimation procedure based on the interacting multiple model (IMM) filter principle was also developed for the JMLS model and gave promising results.

Perspectives

In a short term

In the near future, some extensions could be envisaged in order to overcome the limitations of the presented works. Firstly, in Chapter 3, the incubation/propagation was adapted to the available information from the three completed tests of the test-bench. As another tests are still in progress, the data obtained from these tests should be investigated once they are finished. Another point that could be extended concerns the discrete-state multi-branch models developed in Chapter 4 and 5. In effect, while constructing the MB-HMM model (Chapter 4) and the MB-HsMM model (Chapter 5), we assumed that they obey the strictly left-right topology. Such assumption could help to facilitate the RUL estimation tasks, but it also limits the application range of the models. For example, in continuous production industries, maintenance interventions are implemented regularly to prevent machines from high damaged states. Under the effects of these actions, the deterioration processes could not be monotone anymore and the left-right topology becomes inappropriate. Extension to the ergodic topology [128] could help to take into account such real situations in deterioration modeling.

The other possible extension of the MB-HMM and MB-HsMM models comes from the structure selection issue. Indeed, in the numerical studies presented in this manuscript, the number of branches or equivalently of the deterioration modes is assumed to be known beforehand. In practice, however, such information is not always available. Therefore, more studies should be conducted on how we can determine the number of the deterioration modes that exist in a given training data set. For this end, the cross-validation technique proposed by Celeux and Durand in [21] could be considered and investigated.

Also in developing the MB-HMM and Mb-HsMM models, the underlying deterioration modes are assumed to be exclusive once initiated. This implies that branch switching

is not allowed in our models. Although such assumption was inspired from a real phenomenon observed within the SUPREME project as discussed in Section 4.1.2, it could still be relaxed to approach more closely to real dynamic deterioration phenomena in practice. It should be noted that, allowing transitions between the branches also imposes difficulties on the model parameters learning problem which may lead to other works to be done in the future.

The continuous-state multi-branch models developed in Chapter 6, although allow the branches transitions, assumes that the underlying deterioration dynamics can be approximated by linear processes. Such approximations could give good results in our study, but it may not hold in several practical cases. Extension to the multi-branch model concept for non-linear case is hence an obvious direction. An example is to model the temporal evolution of the underlying continuous health states by non-linear stochastic processes such as the Gamma or Wiener processes studied in Chapter 2. Another point that can be extended for this model is the switching state modeling. Indeed, in the future works, the Markovian assumption imposed on the transitions of the discrete variable could be relaxed, i.e. by using a semi-Markov chain. In such cases, the model parameters estimation may become much more complicated and more advanced numerical techniques, such as the particle filter or Monte Carlo simulation could be required.

Another research work to be done in a near future is to investigate how the RUL estimates could be used for a dynamical adaptation of maintenance strategies. This problem has been currently studied in the literature but only the mean value of the RUL was taken into consideration. In our study, the RUL is characterized by a probability distribution, which contains more information than the mean residual life value. It is hence hoped to be able to achieve more efficient maintenance strategies in case that all RUL information is taken into the decision making process.

In a long term

For a long term perspective, a validation of the developed models with real-world data should be envisaged. Indeed, although the data from a test-bench and a case study (i.e. the PHM08 competition) were studied in this thesis, it still does not refer to real-life system data. For this end, the prognostics data repository provided by the prognostics center of excellence of NASA [111] could be a promising direction. In effect, this data repository gives access to several prognostics data sets that can be used for both development and validation of prognostics algorithms.

In addition, the models developed in this thesis are the generic ones. They could be hence applied for modeling the deterioration processes of another kinds of components and/or systems. Therefore, without limiting ourselves in the framework of the SUPREME project, in the future, the developed models could be applied for different systems in different industries.

Appendix A

Formulations for Chapter IV

As described in Chapters 2 and 3, given the specification of a Hidden Markov model, there are three basic problems to solve, in order to use the model in practical applications. Denote $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$ the generic observation sequence (continuous or discrete), these problems are:

1) Evaluation problem

   Given the HMM model λ and a sequence of observations $\mathbf{O}$, compute the probability that the observed sequence was produced by the model. This problem can be solved by the so-called Forward-Backward algorithm [128].

2) Decoding problem

   Given the HMM model λ and a sequence of observations $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$, the decoding problem deals with finding the optimal hidden state sequence $Q = \{q_1, q_2, \ldots, q_T\}$ that have most likely produced the observation sequence. There are several criteria existing for defining the optimality, but the most commonly used criterion is to find the single best state sequence (path) Q that maximizes $P(Q \mid \mathbf{O}, \lambda)$ or equivalently $P(Q, \mathbf{O} \mid \lambda)$. The problem can be solved through the Viterbi algorithm, a technique based on dynamic programming methods [38].

3) Training problem

   Given the observations $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$, find the HMM model $\lambda = (A, C, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \pi)$ that best describe how the observations come about. The best solution to this problem for HMM is represented by the Baum-Welch algorithm [128].

In this appendix, the three algorithms that solve the three problems discussed above for HMMs are reported, as formulated in the work of Rabiner [128, 129].

A.1 The forward-backward algorithm

The goal of the forward-backward algorithm is, given an observation sequence $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$ and a model λ, to calculate the model likelihood, i.e., $\mathbb{P}(\mathbf{O} \mid \lambda)$.

At this purpose, we consider the forward variable $\alpha_t(i)$, for each time t, defined as

$$\alpha_t(i) = \mathbb{P}\left(o_1, o_2, \ldots, o_t, q_t = S_i\right) \qquad 1 \leq i \leq N$$

where N is the number of the HMM states.

The calculation of the forward variable is performed inductively, through the follows steps:

1) Initialization: for $1 \leq i \leq N$

$$\alpha_1(i) = \pi_i b_i\left(o_1\right)$$

2) Induction: for $1 \leq j \leq N$ and $1 \leq t \leq T-1$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j\left(o_{t+1}\right)$$

3) Termination

$$\mathbb{P}\left(\mathbf{O} \mid \lambda\right) = \sum_{i=1}^{N} \mathbb{P}\left(\alpha_T(i)\right)$$

where the last quation gives the solution to the evaluation problem described above.

Although only the forward part of the forward-backward algorithm is needed to calculate the model likelihood, the backward algorithm is also reported, since it will be used in the following to solve the learning problem.

In a similar way as previously done, we define the backward variable $\beta_t(i)$ as

$$\beta_t(i) = \mathbb{P}\left(o_{t+1}, o_{t+2}, \ldots, o_T \mid q_t = S_i, \lambda\right) \qquad 1 \leq i \leq N$$

To solve for $\beta_t(i)$ the following induction formula is used:

1) Initialization: for $1 \leq i \leq N$

$$\beta_T(i) = 1$$

2) Induction: for $1 \leq i \leq N$ and $t = T-1, T-2, \ldots, 1$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j\left(o_{t+1}\right) \beta_{t+1}(j)$$

A.2 The Viterbi algorithm

The goal of the Viterbi algorithm is to find the single best state sequence (path) $Q = \{q_1, q_2, \ldots, q_T\}$ that have most likely produced the observation sequence $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$. At this purpose, we define the variable $\delta_t(i)$ that, for each time t, represents the highest probability along a single path which accounts for the first t observations and ends in state S_i, as

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} \mathbb{P}\left(q_1, q_2, \ldots, q_{t-1}, q_t = S_i, o_1, o_2, \ldots, o_t \lambda\right)$$

By induction we have

$$\delta_{t+1}(j) = \max_{1 \leq i \leq N} \left[\delta_t(i) a_{ij}\right] b_j\left(o_{t+1}\right) \tag{A.1}$$

To actually retrieve the state sequence, we keep track of the argument that maximizes (A.1) for each t and j via the array $\psi_t(j)$. The complete procedure for the Viterbi algorithm is given as follows:

1) Initialization: for $1 \leq i \leq N$

$$\begin{aligned} \delta_1(i) &= \pi_i b_i\left(o_1\right) \\ \psi_1(i) &= 0 \end{aligned}$$

2) Recursion: for $1 \leq j \leq N$ and $2 \leq t \leq T$

$$\begin{aligned} \delta_t(j) &= \max_{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right] b_j\left(o_t\right) \\ \psi_t(j) &= \arg\max_{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right] \end{aligned}$$

3) Termination

$$\begin{aligned} P^* &= \max_{1 \leq i \leq N} \left[\delta_T(i)\right] \\ q_T^* &= \arg\max_{1 \leq i \leq N} \left[\delta_T(i)\right] \end{aligned}$$

4) Path backtracking: for $t = T-1, T-2, \ldots, 1$

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right)$$

A.3 The Baum-Welch algorithm

The solution to the learning problem aims to find the model parameters λ^* which, given the observations sequence $\mathbf{O}$, maximizes the model likelihood, i.e.,

$$\lambda^* = \arg\max_{\lambda} \mathbb{P}(\mathbf{O} \mid \lambda)$$

Since this problem cannot be solved analytically, the Baum-Welch algorithm uses the well known Expectation-Maximization [29] algorithm to find the model parameters λ that locally maximize the likelihood $\mathbb{P}(\mathbf{O} \mid \lambda)$. The Baum-Welch algorithm makes use of the forward-backward algorithm described in Section A.1 [128].

To derive the re-estimation formulas, the variable $\xi_t(i,j)$ is firstly defined, as the probability of being in state S_i at time t, and in state S_j at time $t+1$, given the model parameters and the observation sequence, i.e.,

$$\begin{aligned}\xi_t(i,j) &= \mathbb{P}(q_t = S_i, q_{t+1} = S_j \mid \mathbf{O}, \lambda) \\ &= \frac{\mathbb{P}(q_t = S_i, q_{t+1} = S_j, \mathbf{O} \mid \lambda)}{\mathbb{P}(\mathbf{O} \mid \lambda)}\end{aligned}$$

which can be expressed in terms of forward and backward variables as:

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}$$

From $\xi_t(i,j)$ we can derive the probability of being in state S_i at time t given the observation sequence and the model parameters, represented by the variable $\gamma_t(i)$ as

$$\gamma_t(i) = \mathbb{P}(q_t = S_i \mid \mathbf{O}, \lambda) = \sum_{j=1}^{N} \xi_t(i,j)$$

Note that, if we sum $\xi_t(i,j)$ and $\gamma_t(i)$ over the time index t, we get a quantity that represents the expected (over time) number of transition from S_i to S_j and number of transition from S_i respectively. Using the concept of counting event occurrences, a reasonable re-estimation formulas for the HMM parameters can be given by:

$$\begin{aligned}\overline{\pi}_i &= \text{expected number of times in state } S_i \text{at time } t = 1 = \gamma_1(i) \\ \overline{a}_{ij} &= \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i} \\ &= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}\end{aligned}$$

$$\begin{aligned}\bar{b}_j(k) &= \frac{\text{expected number of times in state } S_j \text{ and observing the symbol } v_k}{\text{expected number of times in state } S_j} \\ &= \frac{\sum_{\substack{t=1 \\ s.t.\ o_t=v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}\end{aligned}$$

Note that the last one is for re-estimation of the observation parameters in case the observations are discrete.

If the observations are continuous and are supposed to follow mixture of Gaussian distributions (c.f. Equation 4.1), re-estimation formulas for the parameters C, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by [128]:

$$\hat{c}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(j,k)} \tag{A.2}$$

$$\hat{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot o_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{A.3}$$

$$\hat{\boldsymbol{\Sigma}}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot \left(o_t - \hat{\boldsymbol{\mu}}_{jk}\right)\left(o_t - \hat{\boldsymbol{\mu}}_{jk}\right)'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{A.4}$$

where $\gamma_t(i,k)$ is the probability of being in state S_i at time t with the kth mixture component, i.e. $\gamma_t(i,k) = \mathbb{P}\left(q_t = S_i, K_t = k \mid o_1^T\right)$ where K_t denotes the mixture component at time t.

The re-estimation formula for c_{jk} can be interpreted as the ratio between the expected number of times the model is in state S_j using the kth mixture component, and the expected number of times the model is in state S_j. Similarly, Equation A.3 gives the expected value of the portion of the observation vector accounted for by the kth mixture component. A similar interpretation can be easily obtained for Equation A.4 [128].

Bibliography

[1] Abdel-Hameed, M. (1975). A gamma wear process. *Reliability, IEEE Transactions on*, R-24(2):152–153.

[2] Applebaum, D. (2009). *Lévy processes and stochastic calculus*. Cambridge university press.

[3] Azimi, M. (2004). *Data transmission schemes for a new generation of interactive digital television*. PhD thesis, The University of British Columbia.

[4] Azimi, M., Nasiopoulos, P., and Ward, R. K. (2005). Offline and online identification of hidden semi-markov models. *IEEE Transactions on Signal Processing*, 53(8-1):2658–2663.

[5] Bagdonavicius, V. and Nikulin, M. S. (2001). Estimation in degradation models with explanatory variables. *Lifetime Data Analysis*, 7(1):85–103.

[6] Baruah, P. and Chinnam, R. B. (2005). Hmms for diagnostics and prognostics in machining processes. *International Journal of Production Research*, 43(6):1275–1293.

[7] Bechhoefer, E., Bernhard, A., He, D., and Banerjee, P. (2006). Use of hidden semi-markov models in the prognostics of shaft failure. *HIP*, 1(1):3.

[8] Bechhoefer, E., Clark, S., and He, D. (2010). A state space model for vibration based prognostics. *Proc. PHM*, pages 10–14.

[9] Biem, A., Ha, J.-Y., and Subrahmonia, J. (2002). A bayesian model selection criterion for hmm topology optimization. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–989. IEEE.

[10] Bilmes, J. A. et al. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126.

[11] Bishop, C. M. et al. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

[12] Blackmore, L., Gil, S., Chung, S., and Williams, B. (2007). Model learning for switching linear systems with autonomous mode transitions. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*.

[13] Blom, H. A. and Bar-Shalom, Y. (1988). The interacting multiple model algorithm for systems with Markovian switching coefficients. *Automatic Control, IEEE Transactions on*, 33(8):780–783.

[14] Bonissone, P. and Goebel, K. (2002). When will it break? a hybrid soft computing model to predict time-to-break margins in paper machines. In *Proceedings of SPIE 47th Annual Meeting, International Symposium on Optical Science and Technology*, volume 4787, pages 53–64.

[15] Bunks, C., McCarthy, D., and Al-Ani, T. (2000). Condition-based maintenance of machines using hidden markov models. *Mechanical Systems and Signal Processing*, 14(4):597–612.

[16] Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

[17] Byington, C. S. and Roemer, M. J. (2002). Prognostic enhancements to diagnostic systems for improved condition-based maintenance [military aircraft]. In *Aerospace Conference Proceedings, 2002. IEEE*, volume 6, pages 6–2815. IEEE.

[18] Caesarendra, W., Widodo, A., Thom, P. H., Yang, B.-S., and Setiawan, J. D. (2011). Combined probability approach and indirect data-driven method for bearing degradation prognostics. *Reliability, IEEE Transactions on*, 60(1):14–20.

[19] Camci, F. and Chinnam, R. B. (2010). Health-state estimation and prognostics in machining processes. *Automation Science and Engineering, IEEE Transactions on*, 7(3):581–597.

[20] Cartella, F., Lemeire, J., Dimiccoli, L., and Sahli, H. (2015). Hidden semi-markov models for predictive maintenance. *Mathematical Problems in Engineering*.

[21] Celeux, G. and Durand, J.-B. (2008). Selecting hidden markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564.

[22] Christer, A., Wang, W., and Sharp, J. (1997). A state space condition monitoring model for furnace erosion prediction and replacement. *European Journal of Operational Research*, 101(1):1–14.

[23] Compare, M., Baraldi, P., Turati, P., and Zio, E. (2015). Interacting multiple-models, state augmented particle filtering for fault diagnostics. *Probabilistic Engineering Mechanics*, 40:12–24.

[24] Cooke, R., Mendel, M., and Vrijling, J. (2013). *Engineering probabilistic design and maintenance for flood protection*. Springer Science & Business Media.

[25] Costa, O. L. V., Fragoso, M. D., and Marques, R. P. (2006). *Discrete-time Markov jump linear systems*. Springer Science & Business Media.

[26] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220.

[27] Deloux, E., Castanier, B., and Bérenguer, C. (2008). Maintenance policy for a deteriorating system evolving in a stressful environment. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 222(4):613–622.

[28] Deloux, E., Castanier, B., and Bérenguer, C. (2009). Predictive maintenance policy for a gradually deteriorating system subject to stress. *Reliability Engineering & System Safety*, 94(2):418–431.

[29] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[30] Doksum, K. A. and Hbyland, A. (1992). Models for variable-stress accelerated life testing experiments based on wener processes and the inverse gaussian distribution. *Technometrics*, 34(1):74–82.

[31] Dong, M. (2008). A novel approach to equipment health management based on auto-regressive hidden semi-markov model (ar-hsmm). *Science in China Series F: Information Sciences*, 51(9):1291–1304.

[32] Dong, M. and He, D. (2007). A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology. *Mechanical Systems and Signal Processing*, 21(5):2248–2266.

[33] Dong, M., He, D., Banerjee, P., and Keller, J. (2006). Equipment health diagnosis and prognosis using hidden semi-markov models. *The International Journal of Advanced Manufacturing Technology*, 30(7-8):738–749.

[34] Doucet, A., Gordon, N. J., and Krishnamurthy, V. (2001). Particle filters for state estimation of jump Markov linear systems. *Signal Processing, IEEE Transactions on*, 49(3):613–624.

[35] Engel, S. J., Gilmartin, B. J., Bongort, K., and Hess, A. (2000). Prognostics, the real issues involved with predicting life remaining. In *Aerospace Conference Proceedings, 2000 IEEE*, volume 6, pages 457–469. IEEE.

[36] Feng, Q., Peng, H., and Coit, D. W. (2010). A degradation-based model for joint optimization of burn-in, quality inspection, and maintenance: a light display device application. *The International Journal of Advanced Manufacturing Technology*, 50(5-8):801–808.

[37] Ferguson, J. D. (1980). Variable duration models for speech. In *Proceedings of the Symposium on the Application of HMMs to Text and Speech*, pages 143–179.

[38] Forney Jr, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

[39] Gašperin, M., Juričić, Đ., Boškoski, P., and Vižintin, J. (2011). Model-based prognostics of gear health using stochastic dynamical models. *Mechanical Systems and Signal Processing*, 25(2):537–548.

[40] Gebraeel, N. (2006). Sensory-updated residual life distributions for components with exponential degradation patterns. *Automation Science and Engineering, IEEE Transactions on*, 3(4):382–393.

[41] Gebraeel, N., Elwany, A., and Pan, J. (2009). Residual life predictions in the absence of prior degradation knowledge. *Reliability, IEEE Transactions on*, 58(1):106–117.

[42] Gebraeel, N., Lawley, M., Liu, R., and Parmeshwaran, V. (2004). Residual life predictions from vibration-based degradation signals: a neural network approach. *Industrial Electronics, IEEE Transactions on*, 51(3):694–700.

[43] Gebraeel, N. and Pan, J. (2008). Prognostic degradation models for computing and updating residual life distributions in a time-varying environment. *Reliability, IEEE Transactions on*, 57(4):539–550.

[44] Gebraeel, N. Z., Lawley, M. A., Li, R., and Ryan, J. K. (2005). Residual-life distributions from component degradation signals: A Bayesian approach. *IIE Transactions*, 37(6):543–557.

[45] Geramifard, O., Xu, J.-X., Zhou, J.-H., and Li, X. (2012). A physically segmented hidden markov model approach for continuous tool condition monitoring: Diagnostics and prognostics. *Industrial Informatics, IEEE Transactions on*, 8(4):964–973.

[46] Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural computation*, 12(4):831–864.

[47] Ghasemi, A. and Hodkiewicz, M. R. (2012). Estimating mean residual life for a case study of rail wagon bearings.

[48] Gorjian, N., Ma, L., Mittinty, M., Yarlagadda, P., and Sun, Y. (2010a). A review on degradation models in reliability analysis. In *Engineering Asset Lifecycle Management*, pages 369–384. Springer.

[49] Gorjian, N., Ma, L., Mittinty, M., Yarlagadda, P., and Sun, Y. (2010b). A review on reliability models with covariates. In *Engineering Asset Lifecycle Management*, pages 385–397. Springer.

[50] Haghighi, F., Nooraee, N., and Rad, N. N. (2010). On the general degradation path model: Review and simulation. In *Advances in Degradation Modeling*, pages 147–155. Springer.

[51] Hämäläinen, A., Ten Bosch, L., and Boves, L. (2009). Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider. *Speech Communication*, 51(2):130–150.

[52] Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 155–176.

[53] Haug, R., Lucke, D., Adolf, T., and Johannes, V. (2013a). Deliverable 2.2 - supreme reference model. Confidential deliverable, The SUPREME Project (FP7/2007-2013 grant agreement No 314311).

[54] Haug, R., Lucke, D., Adolf, T., Le, T. T., Bérenguer, C., Chatelain, F., and Christien, J. (2014a). Deliverable 4.1 - enhancement of existing competencies, methodologies and tools. Confidential deliverable, The SUPREME Project (FP7/2007-2013 grant agreement No 314311).

[55] Haug, R., Lucke, D., Adolf, T., Le, T. T., Bérenguer, C., Chatelain, F., and Christien, J. (2014b). Deliverable 4.2 - development of the new required competencies, methodologies and tools. Confidential deliverable, The SUPREME Project (FP7/2007-2013 grant agreement No 314311).

[56] Haug, R., Lucke, D., Adolf, T., Le, T. T., Bérenguer, C., and Christien, J. (2015). Deliverable 4.3 - realisation of supreme reliability and maintainability modules in living labs. Confidential deliverable, The SUPREME Project (FP7/2007-2013 grant agreement No 314311).

[57] Haug, R., Mario, E., Christien, J., Catty, J., Le, T. T., Bérenguer, C., Granjon, P., Teresa, A., Lucke, D., Adolf, T., Jiri, K., and Radek, C. (2013b). Deliverable 2.12 - state of the art and beyond the state of the art. Confidential deliverable, The SUPREME Project (FP7/2007-2013 grant agreement No 314311).

[58] Heimes, F. O. (2008). Recurrent neural networks for remaining useful life estimation. In *Prognostics and Health Management, 2008. PHM 2008. International Conference on*, pages 1–6. IEEE.

[59] Heng, A., Zhang, S., Tan, A. C., and Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, 23(3):724–739.

[60] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

[61] Hossain, M. and Jenkin, M. (2005). Recognizing hand-raising gestures using hmm. In *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*, pages 405–412. IEEE.

[62] Huang, R., Xi, L., Li, X., Liu, C. R., Qiu, H., and Lee, J. (2007). Residual life predictions for ball bearings based on self-organizing map and back propagation neural network methods. *Mechanical Systems and Signal Processing*, 21(1):193–207.

[63] Huang, W. and Askin, R. G. (2003). Reliability analysis of electronic devices with multiple competing failure modes involving performance aging degradation. *Quality and Reliability Engineering International*, 19(3):241–254.

[64] Huynh, K. T. (2011). *Quantification de l'apport de l'information de surveillance dans la prise de décision en maintenance*. PhD thesis, Université de Technologie de Troyes.

[65] Huynh, K. T., Barros, A., and Bérenguer, C. (2012). Adaptive condition-based maintenance decision framework for deteriorating systems operating under variable environment and uncertain condition monitoring. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 226(6):602–623.

[66] Iyer, R., Gish, H., Siu, M.-H., Zavaliagkos, G., and Matsoukas, S. (1998). Hidden markov models for trajectory modeling. In *Fifth International Conference on Spoken Language Processing*.

[67] Jardine, A. K., Lin, D., and Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510.

[68] Kalawoun, J., Pamphile, P., and Celeux, G. (2015). Identifiability of a switching markov state-space model. In *Gretsi 2015*.

[69] Kharoufeh, J. P., Solo, C. J., and Ulukus, M. Y. (2010). Semi-markov models for degradation-based reliability. *IIE Transactions*, 42(8):599–612.

[70] Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1):1–22.

[71] Kim, C.-J. and Nelson, C. R. (1999). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*, volume 2. MIT press Cambridge.

[72] Knoblauch, D. (2004). Data driven number-of-states selection in hmm topologies. In *INTERSPEECH*.

[73] Kujawski, D. and Ellyin, F. (1987). A fatigue crack growth model with load ratio effects. *Engineering Fracture Mechanics*, 28(4):367–378.

[74] Lawless, J. and Crowder, M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis*, 10(3):213–227.

[75] Le, T. T., Bérenguer, C., and Chatelain, F. (2015a). Multi-branch hidden semi-markov modeling for rul prognosis. In *Reliability and Maintainability Symposium (RAMS), 2015 Annual*, pages 1–6. IEEE.

[76] Le, T. T., Bérenguer, C., and Chatelain, F. (2015b). Prognosis based on Multi-branch Hidden semi-Markov Models: A case study. In *In Proc. Fault Detection, Supervision and Safety of Technical Processes Conference, Paris, France, September 2015*.

[77] Le, T. T., Chatelain, F., and Bérenguer, C. (2014). Hidden markov models for diagnostics and prognostics of systems under multiple deterioration modes. In *In Proc. European Safety and Reliability Conference, Wroclaw, Poland, September 2014*, pages 1197–1204.

[78] Le, T. T., Chatelain, F., and Bérenguer, C. (2015c). Jump markov linear systems for deterioration modeling and remaining useful life estimation. In *In Proc. European Safety and Reliability Conference, Zurich, Switzerland, September 2015*, pages 1287–1295.

[79] Le Son, K., Fouladirad, M., and Barros, A. (2012a). Remaining useful life estimation on the non-homogenous gamma with noise deterioration based on gibbs filtering: A case study. In *Prognostics and Health Management (PHM), 2012 IEEE Conference on*, pages 1–6. IEEE.

[80] Le Son, K., Fouladirad, M., Barros, A., Levrat, E., and Iung, B. (2012b). Remaining useful life estimation based on stochastic deterioration models: A comparative study. *Reliability Engineering & System Safety*.

[81] Lee, A., Mera, Y., Saruwatari, H., and Shikano, K. (2002). Selective multi-path acoustic model based on database likelihoods. In *In Proc. of the 7th International Conference on Spoken Language Processing, ICSLP2002*.

[82] Lee, J. J., Kim, J., and Kim, J. H. (2001). Data-driven design of hmm topology for online handwriting recognition. *International journal of pattern recognition and artificial intelligence*, 15(01):107–121.

[83] Lee, M.-L. T. and Whitmore, G. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, pages 501–513.

[84] Lee, S., Li, L., and Ni, J. (2010). Online degradation assessment and adaptive fault detection using modified hidden markov model. *Journal of Manufacturing Science and Engineering*, 132(2):021010.

[85] Lehmann, A. (2009). Joint modeling of degradation and failure time data. *Journal of Statistical Planning and Inference*, 139(5):1693–1706.

[86] Lemoine, A. J. and Wenocur, M. L. (1985). On failure modeling. *Naval Research Logistics Quarterly*, 32(3):497–508.

[87] Levinson, S. E. (1986). Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech & Language*, 1(1):29–45.

[88] Li, D., Biem, A., and Subrahmonia, J. (2001). Hmm topology optimization for handwriting recognition. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 3, pages 1521–1524. IEEE.

[89] Li, J. Q. and Barron, A. R. (1999). Mixture density estimation. In *IN ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 12*. Citeseer.

[90] Li, W. and Pham, H. (2005). Reliability modeling of multi-state degraded systems with multi-competing failures and random shocks. *Reliability, IEEE Transactions on*, 54(2):297–303.

[91] Li, Y., Billington, S., Zhang, C., Kurfess, T., Danyluk, S., and Liang, S. (1999a). Adaptive prognostics for rolling element bearing condition. *Mechanical Systems and Signal Processing*, 13(1):103–113.

[92] Li, Y., Billington, S., Zhang, C., Kurfess, T., Danyluk, S., and Liang, S. (1999b). Dynamic prognostic prediction of defect propagation on rolling element bearings. *Tribology Transactions*, (December 2012):37–41.

[93] Li, Y., Kurfess, T., and Liang, S. (2000). Stochastic prognostics for rolling element bearings. *Mechanical Systems and Signal Processing*, 14(5):747–762.

[94] Liu, Q., Dong, M., and Peng, Y. (2012). A novel method for online health prognosis of equipment based on hidden semi-markov model using sequential monte carlo methods. *Mechanical Systems and Signal Processing*, 32:331–348.

[95] Lu, C. J. and Meeker, W. O. (1993). Using degradation measures to estimate a time-to-failure distribution. *Technometrics*, 35(2):161–174.

[96] Lu, J.-C., Park, J., and Yang, Q. (1997). Statistical inference of a time-to-failure distribution derived from linear degradation data. *Technometrics*, 39(4):391–400.

[97] Luo, J., Pattipati, K. R., Qiao, L., and Chigusa, S. (2008). Model-based prognostic techniques applied to a suspension system. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 38(5):1156–1168.

[98] Ma, M., Park, D.-W., Kim, S. K., and An, S. (2012). Online recognition of handwritten korean and english characters. *Journal of Information Processing Systems*, 8(4):653–668.

[99] Mallor, F. and Santos, J. (2003). Classification of shock models in system reliability. *Monografıas del Semin. Matem. Garcıa de Galdeano*, 27:405–412.

[100] Medjaher, K., Tobon-Mejia, D. A., and Zerhouni, N. (2012). Remaining useful life estimation of critical components with application to bearings. *Reliability, IEEE Transactions on*, 61(2):292–302.

[101] Meeker, W. Q. and Escobar, L. A. (1998). *Statistical methods for reliability data*. John Wiley & Sons.

[102] Mitchell, C., Harper, M., Jamieson, L., et al. (1995). On the complexity of explicit duration hmm's. *IEEE transactions on speech and audio processing*, 3(3):213–217.

[103] Mitchell, C. D. and Jamieson, L. H. (1993). Modeling duration in a hidden markov model with the exponential family. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 331–334. IEEE.

[104] Mobley, R. K. (2002). *An introduction to predictive maintenance*. Butterworth-Heinemann.

[105] Murphy, K. (1998a). Hidden markov model (hmm) toolbox for matlab. *online at http://www. ai. mit. edu/˜ murphyk/Software/HMM/hmm. html*.

[106] Murphy, K. P. (1998b). Switching kalman filters. Technical report, Citeseer.

[107] Murphy, K. P. (2002). *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley.

[108] Myotyri, E., Pulkkinen, U., and Simola, K. (2006). Application of stochastic filtering for lifetime prediction. *Reliability Engineering & System Safety*, 91(2):200–208.

[109] Nakagawa, T. (2007). *Shock and damage models in reliability theory*. Springer Science & Business Media.

[110] Nakamura, S. (2009). *Stochastic reliability modeling, optimization and applications*. World Scientific.

[111] NASA Prognostics Center of Excellence. Prognostics data repository. http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/. [Online; accessed 19-October-2015].

[112] Nelson, L. S. (1964). The sum of values from a normal and a truncated normal distribution. *Technometrics*, 6:469–471.

[113] NF-EN-13306-X-60-319 (2001). Terminologie de la maintenance. Standard, Association Française de Normalisation (AFNOR).

[114] Nicolai, R. P., Frenk, J., and Dekker, R. (2009). Modelling and optimizing imperfect maintenance of coatings on steel structures. *Structural Safety*, 31(3):234–244.

[115] Ocak, H., Loparo, K. A., and Discenzo, F. M. (2007). Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling: A method for bearing prognostics. *Journal of sound and vibration*, 302(4):951–961.

[116] Orchard, M. E. and Vachtsevanos, G. J. (2009). A particle-filtering approach for on-line fault diagnosis and failure prognosis. *Transactions of the Institute of Measurement and Control*, 31(3-4):221–246.

[117] Orsagh, R., Roemer, M., Sheldon, J., and Klenke, C. J. (2004). A comprehensive prognostics approach for predicting gas turbine engine bearing life. In *ASME Turbo Expo 2004: Power for Land, Sea, and Air*, pages 777–785. American Society of Mechanical Engineers.

[118] Pandey, M., Yuan, X.-X., and Van Noortwijk, J. (2009). The influence of temporal uncertainty of deterioration on life-cycle management of structures. *Structure and Infrastructure Engineering*, 5(2):145–156.

[119] Park, C. and Padgett, W. (2005). Accelerated degradation models for failure based on geometric brownian motion and gamma processes. *Lifetime Data Analysis*, 11(4):511–527.

[120] Pavlovic, V., Rehg, J. M., Cham, T.-J., and Murphy, K. P. (1999). A dynamic bayesian network approach to figure tracking using learned dynamic models. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 94–101. IEEE.

[121] Peel, L. (2008). Data driven prognostics using a kalman filter ensemble of neural network models. In *Prognostics and Health Management, 2008. PHM 2008. International Conference on*, pages 1–6. IEEE.

[122] Pellerey, F. (1994). Shock models with underlying counting process. *Journal of applied probability*, pages 156–166.

[123] Peng, H., Feng, Q., and Coit, D. W. (2010). Reliability and maintenance modeling for systems subject to multiple dependent competing failure processes. *IIE transactions*, 43(1):12–22.

[124] Pham, H. T. and Yang, B.-S. (2010). Estimation and forecasting of machine health condition using arma/garch model. *Mechanical Systems and Signal Processing*, 24(2):546–558.

[125] Ponchet, A., Fouladirad, M., and Grall, A. (2010). Assessment of a maintenance model for a multi-deteriorating mode system. *Reliability Engineering & System Safety*, 95(11):1244–1254.

[126] Proschan, F. (1975). Shock models with underlying birth process. *Journal of Applied Probability*, pages 18–28.

[127] Qiu, J., Seth, B. B., Liang, S. Y., and Zhang, C. (2002). Damage mechanics approach for bearing lifetime prognostics. *Mechanical systems and signal processing*, 16(5):817–829.

[128] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[129] Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of speech recognition*, volume 14. PTR Prentice Hall Englewood Cliffs.

[130] Rauch, H. E., Striebel, C., and Tung, F. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450.

[131] Rausand, M. and Høyland, A. (2004). *System reliability theory: models, statistical methods, and applications*, volume 396. John Wiley & Sons.

[132] Robinson, M. E. and Crowder, M. J. (2000). Bayesian methods for a growth-curve degradation model with repeated measures. *Lifetime Data Analysis*, 6(4):357–374.

[133] Russell, N. and Moore, R. (1985). Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pages 5–8. IEEE.

[134] Saxena, A., Goebel, K., Simon, D., and Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *Prognostics and Health Management, 2008. PHM 2008. International Conference on*, pages 1–9. IEEE.

[135] Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*, 3(4):253–264.

[136] Si, X.-S., Wang, W., Hu, C.-H., and Zhou, D.-H. (2011). Remaining useful life estimation–a review on the statistical data driven approaches. *European Journal of Operational Research*, 213(1):1–14.

[137] Sikorska, J., Hodkiewicz, M., and Ma, L. (2011). Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, 25(5):1803–1836.

[138] Singpurwalla, N. D. (1995). Survival in dynamic environments. *Statistical Science*, pages 86–103.

[139] Sophie, S.-Z., Thomas, A., Dominik, L., Ralf, H., Pierrick, B., and Javier, G.-S. (2014). Supreme sustainable predictive maintenance for manufacturing equipment. In *Euromaintenance conference.*

[140] Sun, J., Zuo, H., Wang, W., and Pecht, M. G. (2012). Application of a state space modeling technique to system prognostics based on a health index for condition-based maintenance. *Mechanical Systems and Signal Processing*, 28:585–596.

[141] Tang, L., Kacprzynski, G. J., Goebel, K., and Vachtsevanos, G. (2009). Methodologies for uncertainty management in prognostics. In *Aerospace conference, 2009 IEEE*, pages 1–12. IEEE.

[142] Tobon-Mejia, D. A., Medjaher, K., Zerhouni, N., and Tripot, G. (2012). A data-driven failure prognostics method based on mixture of gaussians hidden markov models. *Reliability, IEEE Transactions on*, 61(2):491–503.

[143] Tran, V. T., Thom Pham, H., Yang, B.-S., and Tien Nguyen, T. (2012). Machine performance degradation assessment and remaining useful life prediction using proportional hazard model and support vector machine. *Mechanical Systems and Signal Processing.*

[144] Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A., and Wu, B. (2006). *Intelligent Fault Diagnosis and Prognosis for Engineering Systems.* Wiley Online Library.

[145] Van Noortwijk, J. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1):2–21.

[146] van Noortwijk, J. M., van der Weide, J. A., Kallen, M.-J., and Pandey, M. D. (2007). Gamma processes and peaks-over-threshold distributions for time-dependent reliability. *Reliability Engineering & System Safety*, 92(12):1651–1658.

[147] van Noortwijk, J. M. and van Gelder, P. H. (1996). Optimal maintenance decisions for berm breakwaters. *Structural Safety*, 18(4):293–309.

[148] Vidal, R., Chiuso, A., and Soatto, S. (2002). Observability and identifiability of jump linear systems. In *Decision and Control, 2002, Proceedings of the 41st IEEE Conference on*, volume 4, pages 3614–3619. IEEE.

[149] Vlok, P.-J., Wnek, M., and Zygmunt, M. (2004). Utilising statistical residual life estimates of bearings to quantify the influence of preventive maintenance actions. *Mechanical systems and signal processing*, 18(4):833–847.

[150] Walter, E. (2013). *Identifiability of state space models: with applications to transformation systems*, volume 46. Springer Science & Business Media.

[151] Wang, M. and Wang, J. (2012). Chmm for tool condition monitoring and remaining useful life prediction. *The International Journal of Advanced Manufacturing Technology*, 59(5-8):463–471.

[152] Wang, P. and Coit, D. W. (2007). Reliability and degradation modeling with random or uncertain failure threshold. In *Reliability and Maintainability Symposium, 2007. RAMS'07. Annual*, pages 392–397. IEEE.

[153] Wang, P. and Vachtsevanos, G. (2001). Fault prognostics using dynamic wavelet neural networks. *AI EDAM*, 15(04):349–365.

[154] Wang, T., Yu, J., Siegel, D., and Lee, J. (2008). A similarity-based prognostics approach for remaining useful life estimation of engineered systems. In *Prognostics and Health Management, 2008. PHM 2008. International Conference on*, pages 1–6. IEEE.

[155] Wang, W.-b. (2000). A model to determine the optimal critical level and the monitoring intervals in condition-based maintenance. *International Journal of Production Research*, 38(6):1425–1436.

[156] Wang, W. Q., Golnaraghi, M. F., and Ismail, F. (2004). Prognosis of machine health condition using neuro-fuzzy systems. *Mechanical Systems and Signal Processing*, 18(4):813–831.

[157] Wang, Y. and Pham, H. (2011). Imperfect preventive maintenance policies for two-process cumulative damage model of degradation and random shocks. *International Journal of System Assurance Engineering and Management*, 2(1):66–77.

[158] Whitmore, G. and Schenkelberg, F. (1997). Modelling accelerated degradation data using wiener diffusion with a time scale transformation. *Lifetime data analysis*, 3(1):27–45.

[159] Williams, B. C., Hofbaur, M., and Jones, T. (2001). Mode estimation of probabilistic hybrid systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

[160] Wu, W., Hu, J., and Zhang, J. (2007). Prognostics of machine health condition using an improved arima-based prediction method. In *Industrial Electronics and Applications, 2007. ICIEA 2007. 2nd IEEE Conference on*, pages 1062–1067. Ieee.

[161] Yam, R., Tse, P., Li, L., and Tu, P. (2001). Intelligent predictive decision support system for condition-based maintenance. *The International Journal of Advanced Manufacturing Technology*, 17(5):383–391.

[162] Yan, J., Koc, M., and Lee, J. (2004). A prognostic algorithm for machine performance assessment and its application. *Production Planning & Control*, 15(8):796–801.

[163] You, M.-Y., LIN, L., GUANG, M., and JUN, N. (2010). Two-zone proportional hazard model for equipment remaining useful life prediction. *Journal of manufacturing science and engineering*, 132(4).

[164] Yu, J. (2012). Health condition monitoring of machines based on hidden markov model and contribution analysis. *Instrumentation and Measurement, IEEE Transactions on*, 61(8):2200–2211.

[165] Yu, S.-Z. (2010). Hidden semi-markov models. *Artificial Intelligence*, 174(2):215–243.

[166] Yu, S.-Z. and Kobayashi, H. (2003). An efficient forward-backward algorithm for an explicit-duration hidden markov model. *Signal Processing Letters, IEEE*, 10(1):11–14.

[167] Yu, S.-Z. and Kobayashi, H. (2006). Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden markov model. *Signal Processing, IEEE Transactions on*, 54(5):1947–1951.

[168] Yu, W. K. and Harris, T. A. (2001). A new stress-based fatigue life model for ball bearings. *Tribology transactions*, 44(1):11–18.

[169] Zhang, X., Xu, R., Kwan, C., Liang, S. Y., Xie, Q., and Haynes, L. (2005). An integrated approach to bearing fault diagnostics and prognostics. In *American Control Conference, 2005. Proceedings of the 2005*, pages 2750–2755. IEEE.

[170] Zhang, Y., Zhao, X., Liu, W., Zhang, J., Jia, Y., and Feng, T. (2011). Research on gearbox wearing prognosis based on gamma-state space model. In *Reliability, Maintainability and Safety (ICRMS), 2011 9th International Conference on*, pages 279–283. IEEE.

[171] Zhou, Y., Ma, L., and Mathew, J. (2008). A non-gaussian continuous state space model for asset degradation.

[172] Zhou, Y., Ma, L., Mathew, J., Kim, H., and Wolff, R. (2009). Asset life prediction using multiple degradation indicators and lifetime data: a gamma-based state space model approach. In *Reliability, Maintainability and Safety, 2009. ICRMS 2009. 8th International Conference on*, pages 445–449. IEEE.

[173] Zuo, M. J., Jiang, R., and Yam, R. (1999). Approaches for reliability modeling of continuous-state devices. *Reliability, IEEE Transactions on*, 48(1):9–18.