Integrated auto-diagnosis based on stochastic model for rolling element bearings

Yaqiang Jin

N°d'ordre NNT : 2022LYSEI045

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
L'Institut National des Sciences Appliquées de Lyon

Ecole Doctorale N° 162
MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE

Spécialité de doctorat : Génie Mécanique

Soutenue publiquement le 25/05/2022, par :
Yaqiang JIN

Integrated auto-diagnosis based on stochastic model for rolling element bearings

---

Devant le jury composé de :

GRYLLIAS, Konstantinos	Professeur, KU Leuven	Rapporteur
WYLOMANSKA, Agnieszka	Professeur, Wroclaw University of Science and Technology	Rapporteur
EL BADAOUI, Mohamed	Professeur, Université Saint-Etienne	Examinateur
ANTONI, Jérôme	Professeur, INSA-Lyon	Directeur de thèse

Département FEDORA – INSA Lyon - Ecoles Doctorales

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **https://www.edchimie-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage secretariat@edchimie-lyon.fr	**M. Stéphane DANIELE** C2P2-CPE LYON-UMR 5265 Bâtiment F308, BP 2077 43 Boulevard du 11 novembre 1918 69616 Villeurbanne **directeur@edchimie-lyon.fr**
E.E.A.	**ÉLECTRONIQUE, ÉLECTROTECHNIQUE, AUTOMATIQUE** **https://edeea.universite-lyon.fr** Sec. : Stéphanie CAUVIN Bâtiment Direction INSA Lyon Tél : 04.72.43.71.70 secretariat.edeea@insa-lyon.fr	**M. Philippe DELACHARTRE** INSA LYON Laboratoire CREATIS Bâtiment Blaise Pascal, 7 avenue Jean Capelle 69621 Villeurbanne CEDEX Tél : 04.72.43.88.63 **philippe.delachartre@insa-lyon.fr**
E2M2	**ÉVOLUTION, ÉCOSYSTÈME, MICROBIOLOGIE, MODÉLISATION** **http://e2m2.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.e2m2@univ-lyon1.fr	**M. Philippe NORMAND** Université Claude Bernard Lyon 1 UMR 5557 Lab. d'Ecologie Microbienne Bâtiment Mendel 43, boulevard du 11 Novembre 1918 69 622 Villeurbanne CEDEX **philippe.normand@univ-lyon1.fr**
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTÉ** **http://ediss.universite-lyon.fr** Sec. : Sylvie ROBERJOT Bât. Atrium, UCB Lyon 1 Tél : 04.72.44.83.62 secretariat.ediss@univ-lyon1.fr	**Mme Sylvie RICARD-BLUM** Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (ICBMS) - UMR 5246 CNRS - Université Lyon 1 Bâtiment Raulin - 2ème étage Nord 43 Boulevard du 11 novembre 1918 69622 Villeurbanne Cedex Tél : +33(0)4 72 44 82 32 **sylvie.ricard-blum@univ-lyon1.fr**
INFOMATHS	**INFORMATIQUE ET MATHÉMATIQUES** **http://edinfomaths.universite-lyon.fr** Sec. : Renée EL MELHEM Bât. Blaise PASCAL, 3e étage Tél : 04.72.43.80.46 infomaths@univ-lyon1.fr	**M. Hamamache KHEDDOUCI** Université Claude Bernard Lyon 1 Bât. Nautibus 43, Boulevard du 11 novembre 1918 69 622 Villeurbanne Cedex France Tél : 04.72.44.83.69 **hamamache.kheddouci@univ-lyon1.fr**
Matériaux	**MATÉRIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec. : Yann DE ORDENANA Tél : 04.72.18.62.44 yann.de-ordenana@ec-lyon.fr	**M. Stéphane BENAYOUN** Ecole Centrale de Lyon Laboratoire LTDS 36 avenue Guy de Collongue 69134 Ecully CEDEX Tél : 04.72.18.64.37 **stephane.benayoun@ec-lyon.fr**
MEGA	**MÉCANIQUE, ÉNERGÉTIQUE, GÉNIE CIVIL, ACOUSTIQUE** **http://edmega.universite-lyon.fr** Sec. : Stéphanie CAUVIN Tél : 04.72.43.71.70 Bâtiment Direction INSA Lyon mega@insa-lyon.fr	**M. Jocelyn BONJOUR** INSA Lyon Laboratoire CETHIL Bâtiment Sadi-Carnot 9, rue de la Physique 69621 Villeurbanne CEDEX **jocelyn.bonjour@insa-lyon.fr**
ScSo	**ScSo*** **https://edsciencessociales.universite-lyon.fr** Sec. : Mélina FAVETON INSA : J.Y. TOUSSAINT Tél : 04.78.69.77.79 melina.faveton@univ-lyon2.fr	**M. Christian MONTES** Université Lumière Lyon 2 86 Rue Pasteur 69365 Lyon CEDEX 07 **christian.montes@univ-lyon2.fr**

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Acknowledgments

I would like to initially express my deepest gratitude to China Scholarship Council, which provides me sufficient supports to complete my Ph.D project (No.CSC201806070149) in Laboratory of vibration and acoustics (LVA) of INSA-Lyon from 2018 to 2022. The wonderful experience in France will last forever in my heart. During the Ph.D research, I believe I have gradually grasped how to become a good researcher and gained invaluable understandings about myself. This growth in my research and mindset cannot happen without the help of the professors, colleagues and friends.

First of all, I would like to thank my supervisor, Prof. Jérôme Antoni, for his invaluable guidance, patience and support in all aspects of my research. His enthusiasm for research, his creative and critical way of thinking greatly affected me during my Ph.D pursuit. I am also grateful for the excellent example he has personified as a mentor and professor. I would also like to thank my CST jury members, Prof. Didier Remond and Prof. Quentin Leclere, for their review on my research and encouraging and helpful suggestions, which really improved the quality of my Ph.D research.

I would like to thank all my current and former colleagues of the LVA for their companionship and support, making my time in the office memorable and entertaining. Especially thanks for Dr. Alice Dinsenmeyer, Dr. Fulbert Mbailassem, Dr. Xiaowen Zhu, for their kind support and help that have been of great value in my Ph.D period. A big thank for my bro Dr. Xiaokang Zhang who is not only a good colleague to work together, but also a bosom buddy to share the life. It has been a pleasure knowing you all and creating happy moments with you.

Finally my gratitude goes to my family members, relatives for their constant supports and encouragements throughout the Ph.D period. Especially to my fiancée Junli who provided emotional understanding and unconditional support during these challenging

years. Last I want to say thanks to myself for the past persistence. It is these that accompany me passing some tough times.

Life is a marathon, and the Ph.D diploma is just one milestone on the road. Step will not stop, stepping for the next one.

Abstract

Today, the most fundamental issue of condition monitoring in most industrial plants is fault diagnostics and prognostics. One of the most effective approaches to investigate this issue is condition monitoring based on vibration signal analysis. With the development of industry, multi-threaded maintenance and multi-channel acquisition are becoming more widespread in the current, which put forward higher requirements for maintenance. Based on this observation, it is proposed in this thesis one automated diagnosis framework for the rolling element bearing that integrates the successive steps of fault detection, fault type identification, fault signal reconstruction and fault size characterization. The advantage is that the complete diagnosis process is completed at once, while involving only one key hyperparameter, which improves the degree of automation of current Condition Based Maintenance (CBM) and liberating human participation.

In the presence of incipient fault, vibrations of rolling element bearings show symptomatic signatures in the form of repetitive impulses. This can be seen as a non-stationary signal whose statistical properties switch between two states. The proposed maintenance strategy models such characteristics with an explicit-duration hidden Markov model (EDHMM) and uses the estimated model parameters to perform integrated diagnosis without requiring the user's expertise. The detection of a fault is first achieved by means of a likelihood ratio test built on the EDHMM parameters. One statistical counting approach and posterior probability spectrum are then used for identifying the fault type automatically. In order to obtain the fault signal in some cases, one Bayesian filter based on the EDHMM parameters is constructed. Finally, the fault size is estimated from the duration times returned by EDHMM.

Subsequently, the capability of the integrated auto-diagnosis framework is illustrated on different experimental datasets. The first validation is forced on the vibration data

for specific conditions. The results prove the robust and accurate maintenance of the rolling element bearing. In addition, the result of accelerated degradation data also shows the effectiveness of the method, especially the ability of detecting failure occurrence and tracking quantitatively fault development. This technique has potential for using in the machine CBM.

Keywords: Intelligence Maintenance; Rolling element bearing; Integrated auto-framework; Fault detection; Fault type identification; Fault signal reconstruction; Fault size characterization.

Résumé

Aujourd'hui, le problème le plus fondamental de la surveillance d'état dans la plupart des installations industrielles est le diagnostic et le pronostic des défauts. L'une des approches les plus efficaces pour étudier ce problème est la surveillance de l'état basée sur l'analyse des signaux de vibration. Avec le développement de l'industrie, la maintenance multithread et l'acquisition multicanal se généralisent, ce qui met en avant des exigences de maintenance plus élevées. Sur la base de cette observation, il est proposé dans cette thèse un cadre de diagnostic automatisé pour le roulement qui intègre les étapes successives de détection de défaut, d'identification du type de défaut, de reconstruction du signal de défaut et de caractérisation de la taille du défaut. L'avantage est que le processus de diagnostic complet est réalisé en une seule fois, tout en impliquant un seul hyperparamètre clé, ce qui améliore le degré d'automatisation de la maintenance basée sur les conditions (CBM) actuelle et libère la participation humaine.

En présence de défaut naissant, les vibrations des roulements présentent des signatures symptomatiques sous forme d'impulsions répétitives. Cela peut être vu comme un signal non stationnaire dont les propriétés statistiques basculent entre deux états. La stratégie de maintenance proposée modélise ces caractéristiques avec un modèle de Markov caché à durée explicite (EDHMM) et utilise les paramètres estimés du modèle pour effectuer un diagnostic intégré sans nécessiter l'expertise de l'utilisateur. La détection d'un défaut est d'abord réalisée au moyen d'un test de rapport de vraisemblance construit sur les paramètres de l'EDHMM. Une approche de comptage statistique et une probabilité sont ensuite utilisées pour identifier automatiquement le type de défaut. Afin d'obtenir le signal de défaut dans certains cas, un filtre bayésien basé sur les paramètres EDHMM est construit. Enfin, la taille du défaut est estimée à partir des durées renvoyés par EDHMM.

Par la suite, la capacité du cadre d'autodiagnostic intégré est illustrée sur différents

ensembles de données expérimentales. La première validation est réalisée sur les données de vibration pour des conditions spécifiques. Les résultats indiquent une diagnostic robuste et précise du roulement à éléments roulants. De plus, le résultat sur des données de dégradation accélérée montre également l'efficacité de la méthode, en particulier la capacité à détecter l'occurrence d'une défaillance et de suivre quantitativement son développement. Cette technique a un potentiel d'utilisation en CBM.

Mots-clés: maintenance du renseignement; Roulement à élément roulant; Auto-framework intégré; Détection de fautes; Identification du type de défaut; Reconstruction du signal de défaut; Caractérisation de la taille des défauts.

Résumé étendu en français

Introduction

Contexte et motivations

En tant que composants essentiels, les roulements sont utilisés dans toutes sortes de systèmes mécaniques rotatifs, par exemple les éoliennes, l'aérospatiale, le transport, etc. Ce sont des composants d'entraînement très sollicités qui, avec le temps, deviennent sensibles à l'usure et à l'écaillage. Les performances normales des machines tournantes dépendent entièrement de l'état de santé des roulements, qui représentent près de 45 à 55% de ces défaillances d'équipement. La défaillance d'un seul roulement peut souvent provoquer l'arrêt de systèmes de fabrication entiers, réduisant ainsi la fiabilité et la disponibilité du système. Cela augmente à son tour les temps d'arrêt de la production causant une perte financière massive à l'organisation et peut même s'avérer dangereux pour la sécurité des travailleurs. Par conséquent, les stratégies de maintenance efficaces pour minimiser l'impact des défaillances des roulements sont très importantes pour l'industrie.

La stratégie de maintenance actuelle est la maintenance prédictive, qui peut être réalisée grâce à une série de tests et d'analyses sans qu'aucun dommage ne se produise. Elle est organisée et mis en œuvre en fonction de l'état de fonctionnement de l'équipement et est appelée Maintenance Based Maintenance (CBM). Cependant, ce type de maintenance repose encore principalement sur l'expertise humaine, qui sera effectuée à un coût élevé et pour une quantité limitée de données seulement. De plus, la surveillance de l'état du roulement n'est pas simple, ni la détection ou l'identification du type de défaut lorsqu'un défaut s'est produit. En réalité, le diagnostic des roulements est un sujet important et de nombreux sous-problèmes doivent être résolus. Cependant, les techniques actuelles sont presque toutes conçues pour des problèmes spécifiques, plutôt que pour l'ensemble

du diagnostic, ce qui est très exigeant pour l'utilisateur et nécessite des connaissances en algorithmes et une expérience d'expert. Tels sont les obstacles sur la route de la maintenance.

Objectifs

Sur la base de l'importance et des défis mentionnés ci-dessus de la surveillance de l'état des roulements, l'objectif de cette thèse est de présenter un cadre d'autodiagnostic intégré pour les roulements. L'importance et la contribution de cette recherche sont le développement d'un cadre d'autodiagnostic intégré grâce à la modélisation stochastique du signal pour réaliser à la fois la détection de défaut, la reconstruction du signal de défaut, l'identification du type de défaut et la caractérisation de la taille du défaut. Pour atteindre ces objectifs, un modèle stochastique, le modèle de Markov caché à durée explicite (EDHMM), est sélectionné pour modéliser dans un premier temps la distribution temps-fréquence du signal de vibration. Les paramètres estimés du modèle EDHMM sont utilisés pour effectuer ces tâches sans avoir besoin d'autres informations préalables ni de l'expertise de l'utilisateur.

Les principales contributions du présent travail sont résumées comme suit. Il introduit une méthodologie complète qui:

- réalise la détection, l'identification et la caractérisation des défauts à la fois, dans un cadre intégré qui dépend d'un seul hyperparamètre clé ; cela vient avec
  - une test du rapport de vraisemblance pour la détection d'un défaut
  - un filtre bayésien pour reconstruire le signal de défaut
  - un spectre de probabilité a posteriori qui est une alternative robuste au spectre d'enveloppe standard, exempt de tout pré-traitement, et une autre technique d'identification automatique d'un point de vue statistique sans aucun examen visuel
  - une technique simple basée sur la régression linéaire pour estimer la taille du défaut
- est non supervisé, dans le sens où elle s'applique sans besoin de données historiques et sans apprendissage,

- est automatisée, en ce sens que ces tâches sont effectuées consécutivement sans nécessiter l'intervention de l'utilisateur.

Chapitre 2

Ce chapitre décrit les caractéristiques de base des vibrations des roulements et l'état de l'art sur le diagnostic des défauts des roulements. Premièrement, les caractéristiques des roulements sont décrites dans la section 2.2, y compris la structure de base, les caractéristiques de vibration et les caractéristiques de défaut naissant. La section 2.3 décrit quelles techniques actuelles sont appliquées pour le diagnostic des défauts de roulements de différents points de vue, y compris la détection, l'amélioration du signal de défaut, l'identification et les techniques d'estimation de la taille des défauts. La littérature sur les méthodes de modélisation stochastique est passée en revue, en particulier pour le HMM et l'EDHMM. Enfin, les techniques de diagnostic automatique sont également passées en revue. Sur la base de cette recherche bibliographique, quelques points sont dignes de mention. 1). À travers la littérature disponible, comme indiqué dans la section 2.3, il est clair que diverses techniques avancées ont été proposées pour répondre aux différentes tâches de diagnostic. Cependant, ces techniques sont presque toutes indépendantes et nécessitent que les utilisateurs aient un niveau élevé d'expérience et de connaissances lorsqu'ils traitent différents problèmes de diagnostic. 2). En ce qui concerne les techniques qui tentent d'aller dans le sens d'un diagnostic intégré et automatisé dans la section 2.5, il n'existe encore aucune recherche capable d'intégrer tous les sous-problèmes de diagnostic dans une seule solution. Encore moins de travaux ont essayé d'atteindre ces objectifs de manière automatisée, c'est-à-dire sans réglage manuel des algorithmes par un expert. Donc toutes ces raisons nous poussent dans une direction, l'autodiagnostic intégré, ce qui est exactement ce que poursuit la thèse.

Chapitre 3

Ce chapitre examine le modèle stochastique et comment l'utiliser pour modéliser la distribution temps-fréquence du signal de vibration du roulement, y compris le paramétrage, l'estimation et l'analyse. La section 3.2 introduit d'abord la chaîne de Markov, ouvrant la voie aux HMM et EDHMM suivants respectivement dans les sections 3.3 et 3.4. Le cadre

proposé dépend principalement de l'EDHMM. Il vise à capturer l'évolution temporelle des coefficients STFT et à calculer la durée dans différents états. L'estimation de ces paramètres nécessite des algorithmes spécifiques, qui sont détaillés dans la section 3.5. Certains paramètres clés de l'EDHMM sont également abordés dans la section 3.5. L'efficacité de ces paramètres et les caractéristiques qu'ils révèlent sont illustrées par un signal synthétique de défaut de roulement. Enfin, l'hyperparamètre de ce modèle et l'initialisation des paramètres sont discutés. Il s'avère qu'un seul hyperparamètre, c'est-à-dire la longueur de la fenêtre N_w, doit être défini à l'avance. Cela rend l'approche potentiellement plus robuste et mieux adaptée aux applications pratiques en l'absence de connaissances préalables.

Chapitre 4

Ce chapitre présente le cadre de diagnostic intégré, y compris la détection des défauts, la reconstruction du signal de défaut, l'identification du type de défaut et la quantification de la taille du défaut. Ce chapitre est le travail de base de cette thèse. Les paramètres EDHMM obtenus dans le chapitre précedent sont utilisés dans le cadre de diagnostic proposé. La section 4.2 introduit le test du rapport de vraisemblance (LRT) et prête les paramètres EDHMM pour détecter la présence nécessitant un pré-traitement des données. Si un défaut est détecté, l'étape d'identification renvoie le spectre de probabilité postérieur (PPS), l'équivalent d'un spectre d'enveloppe, qui peut être analysé visuellement ou automatiquement. Afin d'automatiser entièrement, une méthode simple est proposée basée sur un point de vue statistique pour calculer la probabilité postérieure de différents types de défauts. De même, le signal de défaut peut être extrait grâce à une filtre de variable en temps construit par la matrice de covariance de l'observation. Indépendamment de ce dernier, l'étape de caractérisation du défaut utilise le paramètre de Poisson de l'EDHMM pour évaluer la taille du défaut par régression linéaire, offrant ainsi une solution simple à une tâche difficile. Il est à nouveau souligné que l'algorithme complet ne repose que sur l'hyperparamètre critique, N_w, l'inverse de la résolution fréquentielle. En particulier, il ne s'appuie ni sur des données historiques, ni sur des informations préalables, ni sur une intervention manuelle.

Chapitre 5

Ce chapitre présente la validation du cadre de diagnostic intégré à travers différents jeux de données. La contribution de cette partie est de déterminer les capacités du cadre d'autodiagnostic intégré proposé pour les roulements. Deux scénarios expérimentaux différents sont présentés dans cet article. La première validation expérimentale est relative à des roulements avec différents dommages, fonctionnant à différentes vitesses et sous différentes charges. Les résultats des 36 ensembles de données de roulement associés à trois cas de défaillance différents montrent l'efficacité de ce cadre. Le deuxième type de données expérimentales rapporte le comportement d'un seul roulement endommagé soumis à un long test à vitesse et charge constantes. Lors de la validation, la méthode proposée a également été comparée à l'approche de pointe, le kurtogram. Pour la première validation, quelques commentaires s'imposent 1) Il s'avère que la technique de détection est efficace pour détecter le temps d'occurrence des défauts, comme le montre la figure 5.13(c) dans le deuxième cas associé aux données de dégradation accélérée. 2) La fréquence caractéristique du roulement a été facilement modulée par la fréquence de l'arbre produisant une série de bandes latérales et d'harmoniques, ce qui a une grande influence sur l'identification. Cependant, les deux techniques d'identification proposées surmontent bien ce problème, vu la Fig. 5.3, la Fig. 5.7 et la Fig. 5.16. 3) Elles sont également capables de quantifier la taille de défaut définé dans la section 5.2, et de fournir des informations sur la propagation des fissures et suivre quantitativement l'évolution de la fissure vue à la section 5.3.

Conclusion

L'objectif principal de cette thèse est l'étude de méthodes de traitement automatique du signal pour différentes tâches de diagnostic dans un cadre intégré. Les avantages remarquables sont qu'elles fonctionnent sans données historiques, sans beaucoup d'hyperparamètres, même sans aucune intervention manuelle dans le processus. La revue de la littérature sur cette question montre que les techniques existantes se concentrent au contraire sur des tâches diagnostique spécifiques. De plus ces techniques sont presque toutes indépendantes et nécessitent que les utilisateurs aient un niveau élevé d'expérience et de connaissances. Certaines méthodes tentent de répondre à une telle aspiration, mais n'atteignent pas la dose automatisée et intégrée comme dans cette thèse. La méthode proposée surmonte ces lacunes comme démontré sur les données expérimentales du chapitre

5. Bien que la validation de la méthode proposée semble parfaite, plusieurs aspects n'ont pas été pris en compte dans la thèse. Dans cette partie, quelques-unes des directions de travail futures liées à cette recherche sont suggérées.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms and abbreviations

ANN	Artificial neural network
BPFI	Ballpass frequency of inner race
BPFO	Ballpass frequency of outer race
BSF	Ball (roller) spin frequency
CBM	Condition based monitoring
CDF	Cumulative distribution function
CDHMM	Continuous density hidden Markov model
EDHMM	Explicit duration hidden Markov model
EEMD	Ensemble empirical mode decomposition
EMD	Empirical mode decomposition
FTF	Fundamental train frequency
GLR	Generalized likelihood ratio
HMM	Hidden Markov model
LMD	Local mean decomposition
LRT	Likelihood ratio test
MED	Minimum entropy deconvolution
ML	Maximum likelihood

MMSE	Minimum mean square error
NFFT	The number of Fourier transform
NMF	Non-negative matrix factorization
PHM	Prognostics and health management
PMF	Probability mass function
PPS	Posterior probability spectrum
RMS	Root Mean Square
RUL	Remaining useful life
SNR	Signal to noise ratio
STFT	Short-time Fourier transform
STSA	Short-time spectral amplitude
SVM	Support vector machine

Operators

$\angle\cdot$	Phase
$.^{\dagger}$	Conjugate transpose
$.^{T}$	Transpose
$\hat{\cdot}$	Estimated quantity
$\Im\{\cdot\}$	Imaginary part operator
$\lfloor\cdot\rfloor$	Round-down to the closest integer
$\mathcal{CN}(\cdot)$	Complex-valued Gaussian distribution
$\bar{\cdot}$	Averaged value
$\Re\{\cdot\}$	Real part operator
$E[\cdot]$	Expected value
$\mathrm{card}(\cdot)$	The number of elements in the set
$\exp(\cdot)$	Exponential function

$\mathrm{L}(\cdot)$	The logarithmic likelihood function
$\ln(\cdot)$	Logarithmic function
$\mathrm{P}(\cdot)$	Poisson distribution

Variables physiques

α	Risk probability
$\alpha_n(i)$	Forward probability
$\beta_n(i)$	Backward probability
$\boldsymbol{\theta}$	EDHMM Parameter set
$\boldsymbol{N}_n$	nth column of STFT matrix of additive noise
$\boldsymbol{X}_n$	nth column of STFT matrix of clean signal
$\boldsymbol{Y}_n$	nth column of STFT matrix of observed signal
$\chi^2_{p,1-\alpha}$	The quantile of the chi-squared distribution with the risk α and freedom degree p
Δe	Coda of the transient pulse
Δf	Frequency resolution
Δt	Duration time in second
$\gamma_n(i)$	Posterior probability of state i
Λ	Generalized likelihood ratio
λ	Duration time in window number
$\mathbb{C}$	Covariance of the observation
$\mathbb{M}$	Pseudo-covariance of the observation
μ	Mean vector of the observation
π_i	Initial probability in the HMM
$\tau_n(i)$	Duration time sequence for state i
$\xi_n(i,j)$	Transition probability from state i to state j

D	Pitch diameter
d	Rolling element diameter
f_c	Cutoff frequency
f_r	Rotating speed
F_s	Sampling frequency
l	Fault size
L_h	Expected fatigue life
max_iter	Maximum number of iteration
N_w	Window length
R	Window shift
v	Degree of freedom
Z	Number of rolling elements
z_n	nth hidden state

Chapter 1

General introduction

Contents

1.1 Background and Motivation

As crucial components, rolling bearings serve almost in all kinds of rotating mechanical systems, e.g., wind turbine, aerospace, transportation etc. They are highly-stressed drive components which, over a period of time, become susceptible to wear and spall. The normal performance of rotating machinery is entirely dependent upon the health state of the rolling bearings, which accounts for almost 45-55% of these equipment failures [23, 108]. The failure of just a single bearing can often cause the stoppage of entire manufacturing systems thereby reducing the reliability and availability of the system. This in turn increases the production downtime causing a massive financial loss and may even prove dangerous to the safety of the workers. Therefore the effective maintenance strategy to minimize the negative impact of bearing failures is very important for the industry.

Fortunately, the maintenance theory is getting more and more advanced to meet the existing large amount of industrial monitoring data. The development of maintenance theory has gone through several important stages in the past decades, from *Run-to-Failure maintenance*, *Scheduled maintenance*, *Predictive maintenance* to *Intelligence maintenance*. The *Run-to-Failure maintenance* is the oldest definition of maintenance, which means that the maintenance only occurs after a sudden or even catastrophic failure. Therefore, such maintenance will cause a lot of time and money costs and even life-threatening. The *Scheduled maintenance* refers to schedule the maintenance plan in advance through the prior knowledge about the average service life of the crucial component, and intervene in regularly to the system based on the plan. So it implies taking the risk of replacing the healthy operational components. With the development of science and technology, it makes the *Predictive Maintenance* possible. This kind of maintenance can be achieved through a series of tests and analyses without the occurrence of damage. It is arranged and implemented based on the operating status of the equipment, and is called Condition Based Maintenance (CBM). However, human expertise is an outstanding solution but at a high cost and for a limited quantity of data only, the analysis being time-consuming. Recently, with the advent of Industry 4.0, the digital factory offers many alternatives to human monitoring. The *Intelligence Maintenance* is a promising direction, which is dedicated to improving the degree of automation of maintenance and liberating human participation. Figure 1.1 shows the main steps in the intelligence maintenance system.

In the current maintenance strategy, one of the most effective approaches is condition

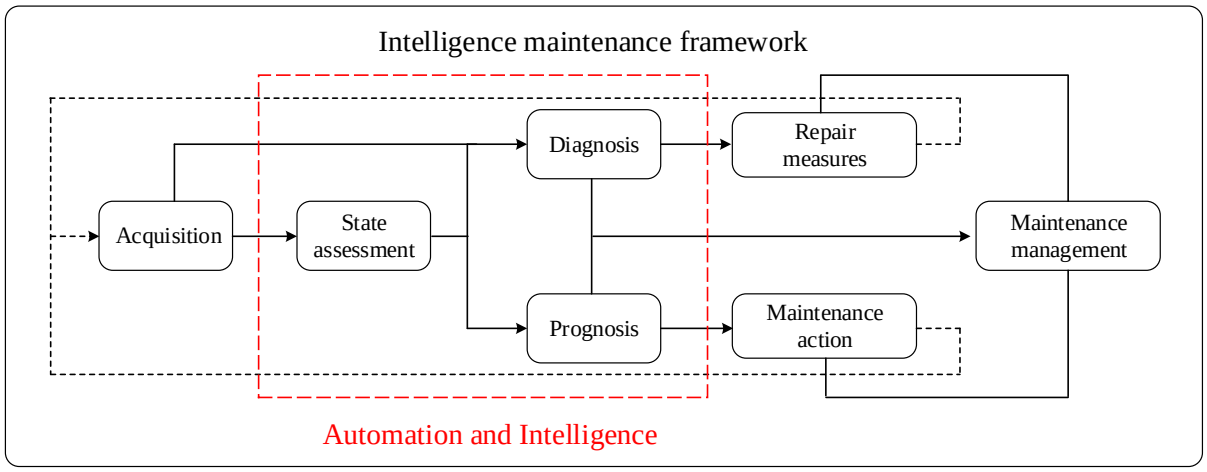

FIGURE 1.1: Intelligence maintenance framework.

monitoring based on vibration signal analysis. Compared with other common techniques, such as acoustic analysis, thermal imaging analysis, etc., vibration signal has irreplaceable advantages, for example, the well known characteristic vibration signatures, well-developed signal processing techniques, supported by various sensors commercially available for different operational conditions, etc. As the characteristic of faulty bearing, the vibration signal will appear the repetitive transient pulses contaminated with additive noise. The transient pulse is often due to faulty point passing through the load zone, which is followed by the oscillations with decaying low-frequency specified by the excited structural resonances. Whereas, the additive noise comes from a variety of interfering sources, such as, channel effects or device defects, system tremor, environmental interference, etc. Due to the structure of the bearing and additive noise interference, this kind of repetitive pattern tends to be stochastic and non-stationary, especially, in high frequency range dominantly, which brings difficulties to analyze through deterministic model or traditional techniques. Therefore recognizing that kind of transient from non-stationary signals generated from bearing needs helps from the specific signal processing tools.

In addition, as shown in Figure 1.1, in the condition monitoring of bearings, it is not simple to detect or identify the fault type when it occurs. In reality, bearing diagnosis is a big topic, and there are many sub-issues that need to be solved. However, current techniques are almost designed for specific issues, rather than toward the whole diagnosis, which is high demanding for the user and requires different algorithms knowledge and expert experience. These are the obstacles on the road to intelligence maintenance. Based on the above mentioned importance and challenges of bearing condition monitoring, the

motivation is to investigate further the use of vibration analysis in solving efficiently some key problems in condition monitoring, so as to present an integrated auto-diagnostic framework for the rolling element bearing.

1.2 Objectives and scope of the research

According to the identified research problem and the above-mentioned motivations, during the envisage phase of diagnosis solution, it is necessary to consider different factors, i.e., simplicity, practicality, automatically. The following questions can be formulated as input to this thesis.

- How to achieve integrated diagnosis?
  Is it possible to integrate all the diagnosis issues (detection, reconstruction, identification, characterization) into one-package framework? This means completing all the diagnosis issues at once without manual tuning of the algorithms.
- How to achieve unsupervised diagnosis?
  Is it possible to employ an advanced signal processing technique and stochastic modeling which requires only one measurement and does not need to be trained by a sequence of data? This means without need of historical data and without training.
- How to achieve automated diagnosis?
  Would it be possible to apply such a technique automatically? This means these tasks are completed consecutively without requiring the user's intervention, and as few hyperparameter settings as possible.

Based on the concerned questions above, this research focuses on the development of bearing intelligence CBM and proposed one integrated auto-diagnosis framework including fault detection, fault signal reconstruction, fault type identification and fault size characterization, which are as shown in Figure 1.2. The main contributions of the present work are resumed as follows. It introduces a complete methodology that:

- achieves detection, identification, and characterization of faults at once, in an integrated framework that depends on only one key hyperparameter; this comes with
  - a likelihood ratio test for detection of a fault [14]
  - a Bayesian filter for reconstruction of the fault signal [162]

  - a posterior probability spectrum that is a robust alternative to the standard envelope spectrum, free of any pre-processing, and another automatic identification technique from statistical-based view of point without any examination visually
  - a simple technique based on linear regression for estimating the fault size

- is unsupervised, in the sense that it is applied without need of historical data and without training,
- is automated, in the sense that these tasks are completed consecutively without requiring the user's intervention.

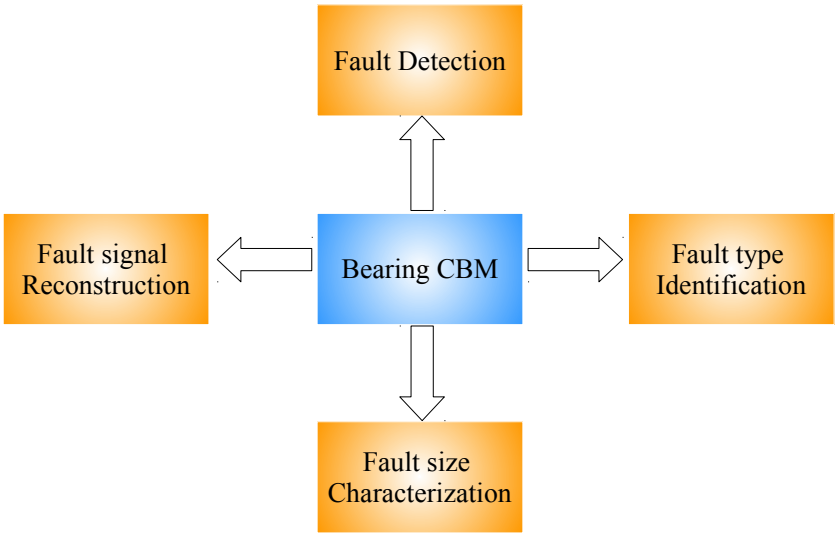

FIGURE 1.2: Components of the integrated diagnosis framework.

The significance and contribution of this research is the development of an integrated auto-diagnostic framework by stochastic modeling of the vibration signal to achieve fault detection, fault signal reconstruction, fault type identification and fault size characterization at once. To achieve this goal, one stochastic model, explicit duration hidden Markov model (EDHMM), is selected to model the time-frequency distribution of the vibration signal firstly. EDHMM is investigated in this research because it provides a well-defined mathematical structure for identifying non-stationary periodicities in time series, and this model returns several valuable parameters useful for the following integrated auto-diagnosis. It is worth mentioning that the proposed four components of intelligent bearing CBM depicted in Figure 1.2 are independent to each other, and any one of the tasks can be accomplished effectively without the aid of other. And they do not affect each

other because different tasks are based on different EDHMM parameters. The detailed probabilistic model and such modular integrated diagnosis framework will be discussed in the upcoming chapters.

1.3 Organization of the thesis

This thesis is divided into six chapters. The contributions and contents of each chapter are explained separately as follows:

Chapter 1 explains the background and motivations behind this research, as well as the research questions. It also describes the main objectives and contributions of this research and outlines an overview of the dissertation.

Chapter 2 outlines the state of the art on the bearing fault diagnosis. Firstly, the characteristics of rolling element bearing are described in Section 2.2, including basic structure, vibration characteristics and incipient fault characteristics. Section 2.3 describes what current techniques are applied in bearing fault diagnostics from different viewpoints, including detection, fault signal enhancement , identification and the fault size estimation techniques. Literature on stochastic modeling method are reviewed, especially the hidden Markov model and explicit duration hidden Markov model. In the final, the automatic diagnosis techniques are also reviewed. The chapter closes with one discussion based on the shortcomings of existing techniques, thereby leading to the research content of this dissertation.

Chapter 3 considers the stochastic models, and how to use them to model the time-frequency distribution of the bearing vibration signal, including the parameter setting, estimation and analysis. Section 3.2 introduces firstly the Markov chain, paving the way for the following HMM and EDHMM in Section 3.3 and Section 3.4, respectively. The proposed framework mainly depends on the EDHMM. It aims at capturing the time evolution of the STFT coefficients and computing the duration time in different states. The estimation of its parameters requires specific algorithms, which are detailed in Section 3.5. In order to have better performance, the model needs some parameters to be set in advance, which is discussed in Section 3.6.

Chapter 4 presents the integrated diagnosis framework including fault detection, fault signal reconstruction, fault type identification and fault size quantification. This chapter

is the core work of this thesis. The EDHMM parameters obtained in the previous chapter will be used in this proposed diagnosis framework. Section 4.2 introduces the likelihood ratio test (LRT) and provides the EDHMM parameters for detecting the presence of fault without requiring any data pre-preprocessing. A time-varying filter based on Bayesian theory is proposed to extract the fault signal in full-band in Section 4.3. Section 4.4 proposes a posterior probability spectrum (PPS) of the states; it is used as a robust alternative to the state-of-the-art envelope spectrum for identifying the fault frequencies. In order to avoid visual examination, a simple method based on a statistical point of view for identifying automatically the fault type is proposed. Finally, the fault size quantification through a simple idea of linear regression is described in Section 4.6. In this section, the time duration parameter of the EDHMM is used to assess the fault size, thus offering a simple solution to a challenging task.

Chapter 5 presents the validation of the integrated diagnosis framework through different bearing datasets. Two different experimental scenarios are reported in this paper. In Section 5.1, the vibration signal relative to bearings with different damages, running at different speeds and under different loads are used. The second scenario (Section 5.2) addresses the accelerated degradation data. In the validation, the proposed method is also compared with the state-of-the-art kurtogram. In the last section, results of the different experimental datasets are discussed.

The last part completes the thesis with general conclusions about the proposed integrated auto-diagnosis framework. Ideas for future research challenges are discussed together with suggestions for improvements of the proposed techniques.

The **Appendix** gives some mathematical deduction used in this thesis, including the detailed process of the stochastic model parameter estimation in Appendix A, and the algorithm convergence in Appendix B.

Chapter 2

State of the art

Contents

2.1 Introduction

This chapter provides an overview of bearing and the current vibration monitoring approaches. The chapter begins with the introduction of various characteristics of rolling element bearings. Since bearing vibration monitoring consists of predefined steps, several state-of-the-art approaches for each diagnostic module are discussed subsequently, including fault detection, fault signal enhancement, fault identification and fault size characterization. Markov process is a desired tool that models the nonstationary nature of the vibration signal and uncovers the intrinsic structure in signals. Hence, in this chapter, a review of Hidden Markov model and its variant is also given. Vibration-based monitoring of rolling element bearings is a complex topic, which involves several issues, then the integrated auto-diagnosis techniques are investigated based on current literature. Furthermore, a preliminary conclusion concerning the limitation of current research and the focus of this thesis is drawn in the last subsection.

2.2 Rolling element bearing signal characteristics

As a special structure, a rolling element bearing has its own unique characteristics, including vibration characteristics and incipient fault characteristics. These characteristics enable bearing to generate useful vibration information when rotating, which offers the possibility of diagnosing potential bearing faults. This section considers some basic but important knowledge, containing the structure of rolling element bearings, the common bearing fault types and how they occur, also the mechanism of vibration generation in rolling element bearings.

2.2.1 Basic structure of rolling element bearing

Most rolling bearings consist of rings with a raceway (inner ring and outer ring), rolling elements and a cage as shown in Figure 2.1. In reality, the outer ring is often fixed on the bearing seat or mechanical body. The inner ring is connected to the drive shaft and driven by the shaft rotation. Rolling elements geometrically contact with the raceway surfaces of the inner and outer rings at “points”, whose role is to convert motion from sliding to rolling. The cage separates the rolling elements at regular intervals, holds them in place within the inner and outer raceways, and allows them to rotate freely and reduces

the collision between them.

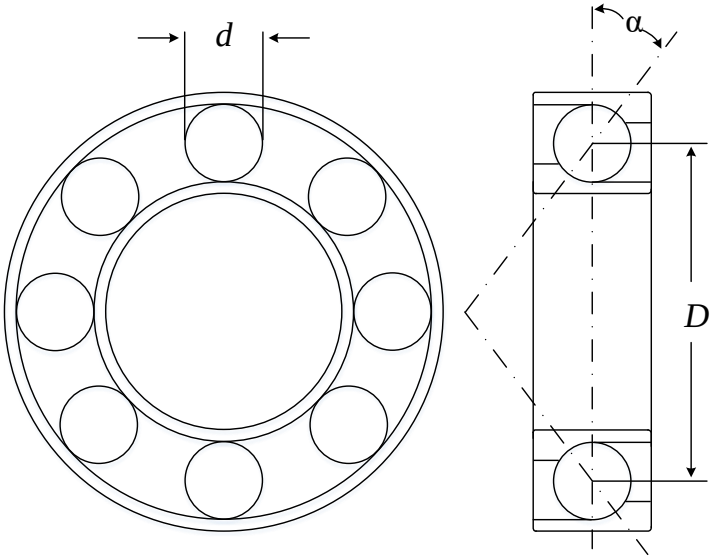

FIGURE 2.1: The diagrammatic sketch of a rolling bearing.

2.2.2 Vibration characteristics of rolling element bearings

Vibration monitoring has now become a well accepted part of many planned maintenance regimes and relies on the well known characteristic vibration signatures which rolling bearings exhibit as the rolling surfaces degrade. However, in fact, vibrations occur all the time. As a rolling element passes over a location on a race, the surface of the race is subjected to a high effective stress for a short time during element passage. This cycling of stress occurs once per element passage at the location. Simultaneously, a location on the rolling element itself is also subjected to a cycle of subsurface stress twice per revolution of the element. The vibrations are generated when surfaces interact through a combination of rolling and sliding. The bearing natural vibration is originated from elastic factors; so whether the bearing fails or not, it will produce vibration.

When the bearing experiences a localised defect during operation, each interaction will cause an instantaneous impact and excite the system to perform high-frequency free attenuation vibration response according to its natural frequency. Such impact has higher energy than the natural excitation. Due to the bearing geometrical symmetry and dynamic regularity, the impact will show typical signal signature.

For a stationary outer ring and rotating inner ring, from the bearing geometry the

fundamental frequencies are derived as follows [124]:

Bearing characteristic frequencies:

- Ballpass frequency, outer race:
  $f_{BPFO} = \frac{Zf_r}{2}[1 - \frac{d}{D}cos\alpha]$
- Ballpass frequency, inner race:
  $f_{BPFI} = \frac{Zf_r}{2}[1 + \frac{d}{D}cos\alpha]$
- Fundamental train frequency (cage speed):
  $f_{FTF} = \frac{f_r}{2}[1 - \frac{d}{D}cos\alpha]$
- Ball (roller) spin frequency:
  $f_{BSF} = \frac{Df_r}{2d}[1 - (\frac{d}{D}cos\alpha)^2]$

where, f_r is the shaft speed, Z is the number of rolling elements, and α is the angle of the load from the radial plane. These bearing equations assume that there is no sliding and that the rolling elements roll over the raceway surfaces. However, in practice this is rarely the case and due to a number of factors the rolling elements undergo a combination of rolling and sliding. As a consequence the actual characteristic defect frequencies may differ slightly from those predicted. Therefore, it is necessary to take the nearby frequency value as the fault frequency according to the actual situation.

There are some unique characteristics for the faulty bearing vibrations, such as impulsiveness [10], cyclostationarity [10] and modulation [124].

- **Impulsive characteristic**
  The impulsive waveform in the faulty bearing vibration signal is due to the response of the structure to excitation forces whenever a faulty point passes through the load zone. Such transient pulse is followed by decaying oscillations with natural frequency corresponding to structural resonances. However, the impulsive waveform caused by the fault is easily submerged in strong background noise at the incipient fault stage, resulting in a low signal-to-noise ratio, which brings great difficulty to the feature extraction of rolling bearings.

- **Cyclostationary characteristic**
  Due to the bearing vibration generation form, i.e. combination of rolling and sliding, the vibration signals from rolling-element bearings often exhibit high levels of

cyclostationarity, especially in the presence of localised faults [8]. Traditional signal processing methods based on Fourier transform encounter certain limitations or are powerless in some cases. This asks for the development of specific signal processing tools [7].

- **Modulation characteristic**

  Rolling bearing vibration signals have typical multi-modulation characteristics. Mcfadden, et al. [106] proposed firstly the vibration model of a single point defect in a rolling element bearing. It can be seen from the model that the vibration signal caused by the localised fault can be regarded as a combination of a low-frequency part (faulty impact) and a high-frequency part (natural resonance frequency). In diagnosis, the actual resonance and other modulation are not of interest, but only the fault impulse demodulated signature.

2.2.3 Incipient fault characteristics of rolling element bearing

The rolling-contact bearing is an element of machinery with a very important role, and it dominates the performance of the machine. However, a rolling element bearing will eventually fail through a fatigue mechanism that can be greatly accelerated by poor installation, overloading, improper lubrication, or contamination. If the abnormality can be detected timely in the early incipient stage, not only the maintenance cost will be reduced, but even catastrophic accidents can be effectively avoided. Therefore, the early detection of bearing faults is extremely important. When the bearing has an initial failure during operation, the vibration signal will have two important characteristics [14]; the first one can be described as an increase in the impulsiveness of the signal, i.e. a deviation from Gaussianity, and/or the second one as a shift of the statistical behaviour of the signal from stationarity to non-stationarity. However, in the incipient fault stage, the level of the vibration will often show unobvious signal features, weak impulsive energy and low signal to noise ratio, etc. These characteristics depend upon many factors, summarised by the following points,

- in the incipient stage, the impact energy of the small fault will be weak, which makes the impact easily submerged in the noise;
- the point at which the vibration is measured may correspond to complicated vibration propagation path, and noise increases during the process of propagation;

- the operating environment of the bearing and other rotating components, i.e. electric motors, gears, shaft, etc., may modify the bearing features;
- the acquisition and conversion error of the sensor is another non-negligible factor.

Therefore, in order to perform better maintenance, one good measurement environment is very important. In addition, the recovery of the fault signature in a low signal-to-noise ratio and the link to the stochastic and nonstationary nature of the fault signature are the critical problems in the incipient fault diagnosis.

2.3 General review of diagnosis of rolling element bearings

2.3.1 Bearing fault detection

During operation, the bearings are subjected to heavy and dynamic loadings generated by machines and transmitted through the components of rolling element bearings. Hence, the healthy condition of bearings is very important in the rotating machinery system. Any failure in the bearings must be detected on time to avoid increase in downtime, production time and catastrophic failure of the machinery. Thus, the detection of these defects is of vital importance for condition monitoring and quality inspection of bearings. As described in Chapter 2.2.2, an instantaneous impact will appear during each interaction. These pulses produce vibration which are the main signatures for proving the fault existence. In this subsection, two different kinds of detection techniques are reviewed, i.e. the time-domain indicators and the stochastic indicators.

2.3.1.1 Time domain indicators

Because the time domain indicators have the advantage of intuition, are easy to understand, convenient to calculate, and high efficiency, they have been widely used in the field of bearing detection. The time domain indicators can be divided into two categories, i.e. dimensional indicators and dimensionless indicators. The dimensional indicators mainly include Root Mean Square (RMS), Standard Deviation, Effective value etc. However, these indicators are usually closely related to the load and speed of rotating machines, and sensitive to the operating conditions. In practical applications, dimensionless indicators are

usually used to monitor the health condition of bearings. The time-domain dimensionless indicators are not only independent on loads and speeds of rotating machinery, but they are also able to effectively indicate early faults occurring in the rotating machinery [84]. There are six indicators usually used for the fault diagnosis of rolling element bearing, shown as follows:

Skewness (SK) [114]

$$SK = \frac{\sum_{t=1}^{L}(x_t - \overline{x})^3}{(L-1)\sigma^3}, \tag{2.1}$$

where, $x_t, t = 1, 2, ..., L$ is the tth sampling point of the raw signal x, L is the number of sampling points, $\overline{x}$ is the mean value of the sample x defined as

$$\overline{x} = \frac{1}{L}\sum_{t=1}^{L} x_t,$$

and σ is the standard deviation of the sample x defined as

$$\sigma = \sqrt{\frac{1}{L-1}\sum_{t=1}^{L}(x_t - \overline{x})^2}.$$

Kurtosis (KU) [53]

$$KU = \frac{\sum_{t=1}^{L}(x_t - \overline{x})^4}{(L-1)\sigma^4}; \tag{2.2}$$

Crest indicator (CI) [51]

$$CI = \frac{max|x_t|}{\sqrt{\frac{1}{L}\sum_{t=1}^{L}(x_t)^2}}; \tag{2.3}$$

Clearance indicator (CLI) [133]

$$CLI = \frac{max|x_t|}{(\frac{1}{L}\sum_{t=1}^{L}\sqrt{|x_t|})^2}; \tag{2.4}$$

Shape indicator (SI)

$$SI = \frac{\sqrt{\frac{1}{L}\sum_{t=1}^{L}(x_t)^2}}{\frac{1}{L}\sum_{t=1}^{L}|x_t|}; \tag{2.5}$$

Impulse indicator (IMI)

$$IMI = \frac{max|x_t|}{\frac{1}{L}\sum_{t=1}^{L}\sqrt{|x_t|}}. \tag{2.6}$$

The six dimensionless parameters are extracted from the raw vibration signals to reflect the general change of the healthy condition of bearings. Among these dimensionless indicators, the Kurtosis is more sensitive to impact signals and most widely used in the fault detection of rolling bearings. However, the above indicators can be considered as the impulsiveness measures. The cyclic behavior is very an important characteristic in the rotating machine vibration signal. Antoni proposed the indicator that used the negentropy of the square envelope spectrum to indicate the cyclostationary behavior [10]. In addition, some indicators based statistical point of view are designed to identify the cyclostationary behavior in Ref. [14],

Indicator for testing the Gaussian cyclostationary hypothesis (GCS) against the Gaussian stationary (GS)

$$I_{GCS/GS}(x) = \ln < s^2(t) > - < \ln s^2(t) >, \tag{2.7}$$

where, $s^2(t)$ is N-periodic component of the squared envelope of the signal, defined as

$$s^2(t) = \frac{1}{K} \sum_{k=0}^{K-1} |x(t+kN)|^2$$

with $K = \lfloor L/N \rfloor$ is the number of N-sample periods of in the L-sample long signal. The operation $< s^2(t) >$ stands for the time average value of $s^2(t)$.

Indicator for testing the generalized Gaussian cyclostationary (GGCS) hypothesis against the generalized Gaussian stationary (GGS) is

$$I_{GGCS/GGS}(x) = 2\kappa_0^{-1} \ln < s^{\kappa_0}(t) > -2\kappa_1^{-1} < \ln s^{\kappa_1}(t) > +C, \tag{2.8}$$

where, $s^{\kappa}(t)$, $\kappa = \kappa_0, \kappa_1$, is N-periodic component of the κ-power envelope of the signal and C stands for an additive constant. $s^{\kappa}(t)$ is defined as

$$s^{\kappa}(t) = \frac{1}{K} \sum_{k=0}^{K-1} |x(t+kN)|^{\kappa}$$

with κ_0, κ_1 are the estimates of the shape parameters of generalized Gaussian distribution.

Some researchers have also proposed band pass filter techniques based on time domain in which impulse loading excites structural resonance in the high frequency zone that can be detected by a transducer. The shock pulse method [5] is based on band pass filtering

techniques. This method is widely used in industries but cannot effectively detect defects at low speed. Dwyer [53] initially introduced the concept of spectral kurtosis and Antoni [11, 15] implemented it in the field of fault diagnosis of rolling element bearings. Spectral Kurtosis (SK) is a statistical tool that identifies the presence of a series of transients and its position in the frequency domain. Antoni [9] in his paper listed various properties of SK for the first time. More recently, Antoni [10] summarized evidence studies of transients and proposed a new spectral extraction method, namely infogram, which considers both the two characteristics and uses negentropy of the square envelope and its spectrum to depict the two characteristics of transients.

2.3.1.2 Stochastic model

As described in Chapter 2.2.2, some signal changes will appear in the process of bearing failure; one is the increase in the impulsiveness of the signal, another is a shift of the statistical behaviour of the signal from stationarity to non-stationarity, i.e. the occurrence of repetitive transients. Although the traditional time-domain indicators are effective and sensitive to impulses, they have no ability to differentiate them. The probabilistic model is another common used way to detect bearing faults. In reference [14], a statistical methodology based on the maximum likelihood ratio is introduced as a general framework to design condition indicators. The proposed indicator is optimal in the sense that it coincides with the statistics of the likelihood ratio test (LRT) which maximizes the probability of true detection given a fixed probability of false alarm (erroneous rejection of the null hypothesis). HMM is also a parametric statistical method that has the capability of pattern classification and is suitable for dynamic time series of signals that are non-stationary. The reference about bearing fault detection based on HMM can be seen in [112, 113]. In reference [167], the researchers introduced a fault detector based on the estimation of features or condition indicators from sensor data. The key attribute of the features is that they characterize a large number of targeted faults. Their selection and extraction are critical processes that affect unequivocally the success or failure of the detector. Another critical component is the fault progression model describing the degrading state of the system. In reality, however, this model is often unknown and severely hinders the application of model-based detection method. With the absence of data, it is impossible to build a data-driven model. The authors [168] therefore developed a model based on the Paris fatigue law and modified it to adapt to the fault mode of interest

and integrate it into the defect detection architecture. A parametric adaptation algorithm was then introduced. The features and fault progression patterns are then defined in a particle filter framework where the current feature distribution is compared to its baseline counterpart to detect a gap or divergence between the two.

2.3.2 Bearing fault signal enhancement

Another significant and often unavoidable problem in signal processing is the presence of background noise due to adverse recording environments, as well as convolutional noise due to sensor and signal propagation variability. Various signal processing techniques involving time, frequency and statistical methods have been used to detect incipient faults. These techniques require a high signal-to-noise-ratio (SNR), where the faulty component vibrations are higher than the background noise. Standard spectral features are highly sensitive to noise, which can decrease the following modeling effectiveness and give misleading results. It is important, therefore, to incorporate signal enhancement and/or robust spectral extraction techniques that enhance the feature while suppressing background noise. The basic goal of feature enhancement is to extract the clean signal $x(t)$ from the measurement $y(t)$ with additive background noise $n(t)$. There are various ways to denoise in the pretreatment, and this part mainly concentrates here on the class of denoising methods that capitalize on time-frequency feature. The signal model can be written as $y(t) = x(t) + n(t)$, and the corresponding Fourier transform as $\boldsymbol{Y}(f) = \boldsymbol{X}(f) + \boldsymbol{N}(f)$. The problem can be simplified as only estimating the spectral amplitude $|\boldsymbol{X}(f)|$, which is based on an underlying fact that the short-time spectral amplitude rather than phase that is important for vibration signal. The spectral amplitude estimation algorithm is one kind of the classic approaches for de-noising, which can be classified as minimum mean square error (MMSE), spectral subtraction, wiener filter based on their estimation method.

The spectral subtraction estimation approach developed by Boll [26] is designed for enhancing signals degraded by uncorrelated additive noise. It is an approach for estimating the magnitude frequency spectrum of the underlying clean signal by subtracting the noise magnitude spectrum from the noisy spectrum. The power spectrum subtraction [93] has the same idea, but instead use the power spectrum. These two basic methods are derived in Chapter 2.3.2.1. Some other variations on the basic spectral subtraction approach have been proposed. Notable work in this area is that of McAulay and Malpass [103], who

formulated the spectral subtraction approach as a maximum likelihood estimation problem of the variance of each spectral component of the original clean signal. Other popular modifications are those that involve averaging or smoothing of the sample spectrum estimator, controlling the amount of subtracted noise [93, 24].

Wiener filter [87] is applied either in the time domain or frequency domain to obtain an estimate of the undergraded signal. In the time domain the Wiener filtering operation can be represented as a convolution, and in the frequency domain as multiplication by the complex filter gain function. At a single frequency, the Wiener filter spectrum estimate [93, 103] is therefore a gain function times the corrupted spectrum $\boldsymbol{Y}(f)$. The Wiener filter is also the optimal MMSE estimator under a Gaussian assumption.

Ephraim et al. [55] proposed one algorithm assuming that the short-time spectrum follows the Gaussian distribution, and utilized MMSE on the short-time spectral amplitude (STSA) of the signal for enhancing the noisy signal. In the literature [54], Ephraim took the measure of mean-square error of the spectra into account, and extended the STSA estimator which minimizes the mean-square error of the log-spectra in enhancing noisy signal. These two STSA estimators were derived under the Gaussian assumption. Porter and Boll [118] gave a modification and improvement for the optimal MMSE estimator calculated directly from noisy data in a way that avoids the need for assuming a specific form of the distribution.

2.3.2.1 Spectral Subtraction

Spectral subtraction [26] is a simple and efficient method. The approach is to estimate the magnitude frequency spectrum of the underlying clean spectrum by subtracting the noise magnitude spectrum from the noisy spectrum. As the signal in practice tends to be non-stationary, then the STFT is applied here. The signal model can be rewritten in short time spectrum as

$$Y(n,k) = X(n,k) + N(n,k),$$

where n means the window index, and $Y(n,k)$, $X(n,k)$, $N(n,k)$ represent the spectrum at frequency k of the nth short segment of the measured signal, the clean signal and the additive noise, respectively. It is assumed that the noise $n(t)$ does not depend on the

original signal $y(t)$. Then the spectral magnitude $|N(n,k)|$ of $N(n,k)$ can be replaced by its average value $\overline{|N(n,k)|}$ taken during normal vibration (without fault). The phase $\angle N(n,k)$ of $N(n,k)$ is replaced by the phase $\angle Y(n,k)$ of $Y(n,k)$. These substitutions result in the spectral subtraction estimator $\hat{X}(n,k)$:

$$\hat{X}(n,k) = [|Y(n,k)| - \overline{|N(n,k)|}]e^{j\angle Y(n,k)}$$

or

$$\hat{X}(n,k) = \boldsymbol{H}_1 \cdot Y(n,k),$$

where,

$$\boldsymbol{H}_1 = 1 - \frac{\overline{|N(n,k)|}}{|Y(n,k)|}. \tag{2.9}$$

In the literature [93, 103], one similar idea called power spectrum subtraction is presented. The author estimated the spectral amplitude firstly, and then used the phase of the original signal $\angle Y(n,k)$ to form the estimate $\hat{X}(n,k)$. The square of the Fourier coefficient modulus can be written as

$$|Y(n,k)|^2 = |X(n,k)|^2 + |N(n,k)|^2 + X(n,k)^\dagger N(n,k) + X(n,k)N^\dagger(n,k). \tag{2.10}$$

From the observed data $y(t)$, $|Y(n,k)|^2$ can be obtained directly. The terms $|N(n,k)|^2$, $X(n,k)^\dagger N(n,k)$ and $X(n,k)N^\dagger(n,k)$ cannot be obtained exactly. In the power spectrum subtraction technique, they are approximated by $E[|N(n,k)|^2]$, $E[X(n,k)^\dagger N(n,k)]$ and $E[X(n,k)N^\dagger(n,k)]$. Because the additive noise $N(n,k)$ is statistically independent with the signal of interest $X(n,k)$, then the quantity $E[X(n,k)^\dagger N(n,k)]$ and $E[X(n,k)N^\dagger(n,k)]$ are zero. The Eq.(2.10) can be rewritten as

$$|\hat{X}(n,k)|^2 = |Y(n,k)|^2 - E[|N(n,k)|^2], \tag{2.11}$$

where, the $E[|N(n,k)|^2]$ can be replaced with $\overline{|N(n,k)|^2}$ if noise is stationary. The phase of $\hat{X}(n,k)$ is from $\angle Y(n,k)$, so that,

$$\hat{X}(n,k) = |\hat{X}(n,k)| \cdot e^{j\angle Y(n,k)}.$$

As pointed in reference [93], the estimated quantity is not guaranteed to be non-

negative since the right-hand side of Eq.(2.10) or Eq.(2.11) can become negative, and a number of somewhat arbitrary choices have been made. In many studies, the negative values are set to zero.

2.3.2.2 Wiener filtering

In the previous section, the basis for estimating the short-time spectral magnitude through a process of spectral subtraction was described. In this section, we review the wiener filter technique, in which a frequency weighting for an "optimum" filter is first estimated from the noisy observation. Here, $Y(n,k)$, $X(n,k)$, $N(n,k)$ again denote the short-time spectra associated with the windowed time functions $y(t)$, $x(t)$ and $n(t)$. The estimate $\hat{X}(n,k)$ of $X(n,k)$ that minimizes the mean-square error is obtained by filtering $y(n)$ with the noncausal Wiener filter as

$$\hat{X}(n) = \boldsymbol{H}_2 \cdot Y(n,k),$$

where, the frequency weighting $\boldsymbol{H}_2$ can be represented by the power quantity P_y, P_x and P_n as

$$\begin{aligned} \boldsymbol{H}_2 &= \frac{P_x}{P_x + P_n} \\ &= \frac{E[|X(n,k)|^2]}{E[|X(n,k)|^2] + E[|N(n,k)|^2]}. \end{aligned} \tag{2.12}$$

Based on the independence assumption and the relationship Eq.(2.10), $E[|Y(n,k)|^2]$ can be simplified as the sum of $E[|X(n,k)|^2]$ and $E[|N(n,k)|^2]$, Then Eq.(2.12) can be also written as

$$\boldsymbol{H}_2 = 1 - \frac{E[|N(n,k)|^2]}{E[|Y(n,k)|^2]}. \tag{2.13}$$

The result Eq.(2.13) derived from the MMSE estimator can be also seen as the Wiener filter. Similar with the previous section, the spectral magnitude $|N(n,k)|$ here can also be replaced by its average value $\overline{|N(n,k)|}$ taken during normal vibration (without fault). In addition, the quantities $E[|Y(n,k)|^2]$, $E[|X(n,k)|^2]$, and $E[|N(n,k)|^2]$ can also be understood as the corresponding covariance matrix. Therefore, the covariance matrix will be used to denoise thanks to Eq.(2.13), which will be introduced detailed in the Chapter 4.3.

2.3.2.3 Minimum Mean-Square Error

In this subsection, one STSA estimation through minimum mean-square error is described, and the spectral components are assumed as statistically independent Gaussian random variables with zero mean. This approach was first proposed by Ephraim in [118], and later used in various fields. Based on the formulation of the estimation problem given in the previous section, our task is to estimate the modulus $|X(n,k)|$ from the observed signal $y(t)$. The MMSE estimator $|\hat{X}(n,k)|$ of $|X(n,k)|$ is obtained through the Bayesian theory as,

$$\begin{aligned}
|\hat{X}(n,k)| &= E[|X(n,k)||y(t)] \\
&= E[|X(n,k)||Y(n,k)] \\
&= \int_{X(n,k)} |X(n,k)| p(X(n,k)|Y(n,k)) \\
&= \int_{X(n,k)} |X(n,k)| \frac{p(Y(n,k)|X(n,k))p(X(n,k))}{p(Y(n,k))} \\
&= \frac{\int_{X(n,k)} |X(n,k)| p(Y(n,k)|X(n,k))p(X(n,k))}{\int_{X(n,k)} p(Y(n,k)|X(n,k))p(X(n,k))}.
\end{aligned} \tag{2.14}$$

If one substitutes the Gaussian distribution into Eq.(2.14) and performs some manipulations (see literature [118]), the estimate $|\hat{X}(n,k)|$ can be obtained as

$$|\hat{X}(n,k)| = \frac{\xi(n,k)}{1+\xi(n,k)} |Y(n,k)|,$$

where $\xi(n,k)$ is the covariance ratio at frequency k of the nth spectral component of the clean signal and the noise $\xi(n,k) = \frac{E[|X(n,k)|^2]}{E[|N(n,k)|^2]}$. Similarly, $\angle Y(n,k)$ is used as the phase of $\hat{X}(n,k)$, and substituting $\xi(n,k)$ into the above formula as

$$\hat{X}(n,k) = \frac{E[|X(n,k)|^2]}{E[|X(n,k)|^2] + E[|N(n,k)|^2]} \cdot Y(n,k). \tag{2.15}$$

It is found that the MMSE estimator of the magnitude of the kth signal spectral component is in fact same as the Wiener estimator. For this reason, Eq.(2.15) is referred to as a Wiener amplitude estimator.

2.3.3 Bearing fault identification

The aim of the identification step is to locate the localized fault through different techniques using the data collected from the bearing. During the past few decades, a significant body of research has been carried out for addressing directly or indirectly fault identification. The state-of-the-art signal processing techniques used recently are listed in Figure 2.2, mainly including 1) signal decomposition: empirical mode decomposition and local mean decomposition, blind deconvolution, matrix decomposition techniques; 2) signal modeling: cyclostationary, hidden Markov model; 3) signal transformation: STFT, Wigner-Ville decomposition, wavelet transformation, cepstral analysis; 4) artificial intelligence, etc. In the final analysis, these advanced processing techniques all need to return back to demodulation and spectrum analysis for finding the bearing fault frequency so as to identify the fault type. These methods are reviewed as follow.

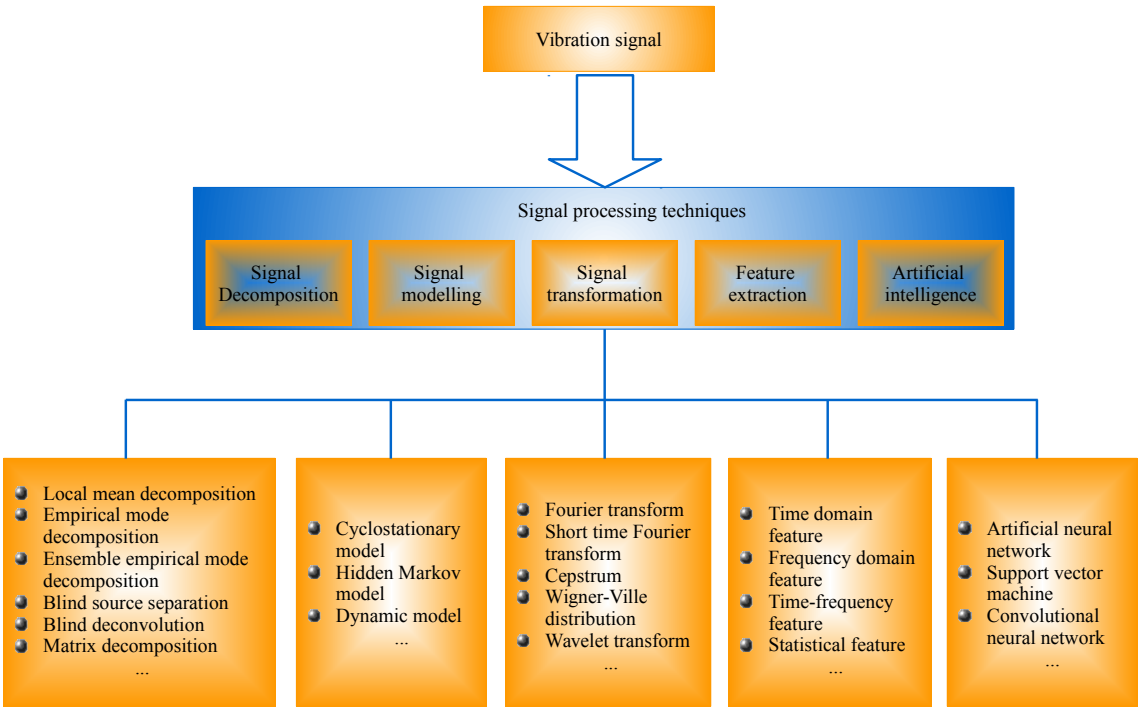

FIGURE 2.2: Vibration signal analysis techniques.

Spectral kurtosis and Kurtogram are one of the most popular methods. Antoni [11] gave a comprehensive theoretical framework of the spectral kurtosis and presented a spectral kurtosis estimation method based on the filter bank decomposition. Subsequently, Antoni [15, 9] published a further study on how to apply SK in the diagnostics of rotating machinery, and presented a tool called the Kurtogram and its fast algorithm, which shows

spectral kurtosis values in a special band-pass structure as a function of two parameters, centre frequency and bandwidth of the filtered signal. A number of improved Kurtogram-based methods have since been proposed by other researchers. Lei et al. [85] pointed out that the Kurtogram based on the STFT or FIR filters are not precise in the extraction of impulsive features of a damaged bearing signal. An improved Kurtogram method was put forward which utilized the Daubechies-wavelet based Wavelet Packet Transform filter. It was demonstrated to be more effective in de-noising the signals and taking out faulty features not present in the original kurtogram. Smith stated in reference [136] that they may fail in specific environments, such as in the presence of electromagnetic interference or other impulsive masking signals. Then the optimised spectral kurtosis was proposed for selecting the best demodulation band to extract bearing fault-related impulsive content from vibration signals contaminated with strong electromagnetic interference.

Cyclostationary characteristic widely exists in the rotating or reciprocating machinery signal, especially the bearing vibration signal, which exhibits some hidden periodicity in its energy flow. McCormick and Nandi [104] appears to be the first to apply the second-order cyclostationary statistics in the monitoring of rotating machinery, and a number of cyclostationary tools such as the spectral correlation density, the Wigner-Ville spectrum and the degree of cyclostationarity were compared against traditional stationary tools to diagnose bearings. The literature [13, 27] explores the cyclostationary characteristics of the defective bearing signals. The vibration signals of rolling bearings consist of random and periodic components. The autocorrelation function of these signals exhibits time varying, periodic and cyclostationary temperament. Sawalhi and Randall [127, 128] studied the extended inner race and outer race fault in rolling element bearings in the presence of gear interaction. The cyclostationary properties such as spectral correlation function of the system were utilized to separate out the bearing faults from the gear faults. Antoni [8] formulated systematically and comprehensively the theory of cyclostationarity, including the definition, cyclostationary tools and the application. Chen [44] used the cyclic spectral density in fault diagnosis of rolling element bearings; this method is described as the Fourier transform of the cyclic autocorrelation function obtained by modifying the time varying autocorrelation function of a Wigner-Ville distribution based on the assumption that the bearing signals possess a cyclostationary character. Abboud et al. [2] extended cyclostationary theory to nonstationary operating conditions through the interaction between time and angle, which is materialized by the angle-periodicity of the correlation

measure of two versions of the signals shifted by a constant time-lag.

Cepstrum is one of the oldest techniques used in the field of signal processing and has been improved by several authors to apply it to the rolling element bearings fault investigation. Cepstrum and logarithmic power spectrum are a pair of Fourier transforms, put differently, cepstrum is the inverse Fourier transform of the power spectrum in logarithmic coordinates. Local faults in gears give an impulsive modulation of the gearmesh signals (both amplitude and frequency modulation) resulting in large numbers of sidebands spaced at the speed of the gear on which the local fault is located [122]. The majority of such sidebands are only visible on a spectrum with a log amplitude scale, and so the cepstrum is an ideal way to extract the frequency information hidden in the sidebands. So the cepstrum can detect and quantify families of periodically spaced spectral components. Borghesani et al. [30] demonstrated the effectiveness of a cepstral based technique for prewhitening of the bearing signals thereby providing a relatively less complicated tool for identifying the bearing faults than the spectral kurtosis based methods. Park et al. [115] used a modified form of cepstrum analysis known as the minimum variance cepstrum obtained by liftering a logarithmic power spectrum by applying the minimum variance algorithm, for the early detection of faults in the ball bearings. It was observed that the minimum variance cepstrum could easily predict the fault period regardless of system frequency response or choice of optimal resonance bands subjugated mostly by the wavelet analysis methods. There are also some articles [71, 99] that provide some more insight into the cepstrum based analysis and the application on rolling element bearings fault diagnosis.

Matrix decomposition technique is usually acted as one preprocessing tool to extract the signal feature or reduce the dimension in the vibration signal analysis. As reference [88] stated, the multidimensional data reduction methods such as independent component analysis, principal component analysis, and singular value decomposition fail to fully understand the multivariate character of bearing signals data. The authors [89][88] adopted an innovative method using the generalized S-transform for time frequency representation of the signal and a two-dimensional non-negative matrix factorization (NMF) to reduce the dimension of the time frequency matrix and extract the desired features from a bearing signal. Such combination is a very common way in vibration signal analysis, as found in other similar methods such as NMF+SVM [165], NMF+kNN [164], NMF based

feature selection for classification [92], etc. The above research only took advantage of dimensionality reduction decomposition. But matrix decomposition technique is more than just this. The analysis could take a further step for recovering the time domain denoised signal. In reference [152], the improved sparse NMF was proposed to separate and extract the compound fault features of rolling bearings, and the time–frequency information was transformed into the time domain by using an inverse STFT. However, the phase recovery process was not introduced. Wodecki et al. [155] used the cyclic spectral coherence as the two dimensional representation map. Then NMF was used to analyze such map in two ways: first, it helped to initially separate cyclic components by producing a set of filters for input vibration data, and second, to identify proper damage-related frequency components in envelope spectrum.

Signal decomposition mainly include empirical mode decomposition (EMD) [69], local mean decomposition (LMD) [135], variational mode decomposition (VMD) [50]. In recent years, feature extraction based on signal decomposition technology has been widely used in fault diagnosis. Guo et al. [63] notified the EMD may be difficult to recover impulses from large noise. Further, the conventional EMD was reported to introduce mode mixing effect and the distortion of the faulty impulses. Based on these two observations, the authors therefore develop a hybrid signal processing method that combines spectral kurtosis (SK) with ensemble empirical mode decomposition (EEMD). Because the end effect and mode mixing problems also plague LMD, the authors [101, 100] proposed a soft screening stop criterion that enables LMD to automatically find an optimal number of iterations for each screening process. In the proposed method, an objective function that considers two characteristics (the root mean square and the excess kurtosis) of the target signal was defined to automatically determine the optimal number of iterations. The recently proposed VMD is to transform the signal decomposition problem into a constrained optimization problem [50]. So it overcomes the problems of end effect and mode component aliasing in EMD and LMD methods and has a more solid mathematical theory foundation.

Blind deconvolution is the technique to find an inverse filter, which can recover the original impact signal by reacting on the transmission path of the signal. At the same time, it can also enhance the impact component in the signal. Thus, blind deconvolution is very suitable for the monitoring and diagnosis of rolling bearing impact fault. There are different

blind deconvolution methods based on different optimization criteria. Minimum entropy deconvolution (MED), as a linear analysis method, was first proposed by Wiggins and applied to the fault diagnosis of bearings by Sawalhi [130]. Cheng et al. [42] proposed an improved MED method, which sloves the filter coefficients by the standard particle swarm optimization algorithm, assisted by a generalized spherical coordinate transformation. The proposed method delivered better deconvolution performance than the classical MED method. McDonald et al. [105] proposed the Maximum Correlation Kurtosis deconvolution. Based on correlation kurtosis as an evaluation index, the deconvolution was realized through an iterative process. In this way, continuous impulse sequences that are submerged by noise in the signal were highlighted. Recently, a novel blind deconvolution method rooted on the maximization of the cyclostationarity of the excitation – as typically encountered with machine faults – was proposed [34]. This method is based on the generalized Rayleigh quotient and solved by means of an iterative eigenvalue decomposition algorithm. The comparison results reveal superior other blind deconvolution methods existing in the literature.

Artificial intelligence is another powerful pattern identification tool, which has attracted great attention from many researchers and shows promise in bearing fault identification or classification applications. The step of fault identification amounts to mapping the information obtained in the feature space to bearing faults in the fault space. Numerous artificial intelligence tools or techniques have been used, including mathematical optimization, as well as classification- and probability-based methods. Specifically, classifiers and statistical learning methods have been widely used in fault diagnosis of rotating machinery, that includes, k-NN algorithms [151], Bayesian classifier [18], SVM [170] and artificial neural network (ANN). Most recently, deep learning approaches have also began to be applied in the field of fault diagnosis [86]. The effectiveness of these approaches largely depends upon the quality of features extracted from the bearing signals.

2.3.4 Bearing fault size characterization

And last but not least, bearing fault size is also an important information during the degradation process. A knowledge of when a bearing will fail – that is, its remaining useful life (RUL) – can serve as supplement to maintenance decision-making such as determining in advance the time an equipment needs to be taken out-of-service and that

can alternatively allow for sufficient lead time for maintenance planning as well. The current research are almost based on the double impact detection, and the time between them is used to estimated fault size.

The earliest research of bearing fault characterization and quantification using vibration analysis was by Epps [56]. He mainly investigated the relationship between the waveform of the vibration source and the localized defect. As he stated in the reference, the vibration characteristics associated with ball passage over a raceway fault could consist of a step response for the ball entry into the fault and an impulse response for the ball impact on the trailing edge of the fault. The entry into the fault appears like a step response, with mainly low frequency content, while the impact on exit excites a much broader band impulse response. Based on this observation, Sawalhi and Randall [129] had a further investigation with both simulated and seeded faults. In order to enable a clear separation of the two events, and produce an averaged estimate of the size of the fault, some advanced signal processing techniques, such as pre-whitening filtering, wavelet and cepstrum analyses, are used to enhance the entry event while keeping the impulse response for fault size estimation. In reference [154], Wang et al. presented a technique based on synchronous signal averaging with the bearing fault characteristic frequency obtained from raw vibration signatures. The averaged signal represents the vibration characteristics within one period of impact produced by the bearing fault. When the fault size is smaller than the diameter of the balls, the features associated with the ball's entry into and exit from the fault may be extracted and the fault size could be derived. Subsequently, in reference [131], the squared envelope of the bearing synchronous averaging is jointed with autoregressive inverse filtration, this combination gives a superior enhancement to the step response and balances it with the impulse response. Ahmadi et al. [4] conducted some experiments, and observed the vibration characteristics under different applied load and rotating speed. The experimental results show strong dependency of the duration time of the low frequency entry and exit transient events on the applied load. It is evident that changes in the speed have almost no effect on duration time of the low frequency entry and exit transient events. Therefore, a new defect size estimation method is proposed and is shown to be accurate for estimating a range of defect geometries over a range of shaft speeds and applied loads. The change in the static stiffness of the bearing and the applied load effect were also researched in [81]. However, as stated in reference [80], almost all previous researches are all based on the tested bearings with defects that have sharp

90° rectangular edges, which is inconsistent with the facts. Larizza et al. developed a numerical model for a rolling element bearing, and a new defect size estimation method was developed to improve the accuracy in estimating the size of a defect that has sloping leading and trailing edges.

2.4 Hidden Markov model and its variants

2.4.1 Hidden Markov model

In this section, one type of stochastic signal model will be focused, namely the hidden Markov model (HMM). It is a good tool to model the transient pattern in nonstationary signals. The basic mathematical theory was published in a series of classic papers by Baum and his colleagues [21, 20] in the late 1960s and early 1970s. Later, it became an important research direction in signal processing; especially in the field of speech recognition. After 2000s, HMMs have become more and more popular in various fields, such as, medical signal modeling, facial expression recognition, gene prediction, gesture recognition, musical composition and vibration signal analysis. There are two strong reasons why it has been so popular in the last decades. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Second the models, when applied properly, work very well in practice for several important applications [120]. Although the number of reported researches based on HMM's is too large to discuss in detail here, a survey on the development of modeling the time-frequency features of nonstationary signal through HMM is given here.

HMM is a statistical model of the time series, and its basic theory was founded in the early 1960s. The HMM was firstly implemented in speech processing in the 1970s, and prosperous in the 1980s. Taking into account the unique characteristics of speech signals, some mainstream features, such as, perceptual linear prediction [73], mel-frequency cepstral coefficients [39], Vector Quantization [121] etc., have been widely used to be modeled through HMMs. In addition to these features, short-time Fourier spectrum is also a good tool. This is because the short-time Fourier transformation and its inverse establish a one-to-one mapping between the time domain and the frequency domain, and FFT algorithms can be implemented efficiently [163]. Also, the short time Fourier transformation preserves information from the original signal, and ensures that important features are not lost as a

result of the transformation. Furthermore, the nonstationary signal can be considered as stationary over a short time interval. In reference [125], the authors had a comparison on these time-frequency features with their application to speech recognition. Juang and Rabiner [74] firstly used the short-time Fourier spectral vectors to characterize the sequence of sounds, and applied HMM in speech recognition and observed an accuracy rate higher than 95% in speaker-independent tasks. Flores [59] describes a scheme for robust speech recognition at poor signal to noise ratios. It consists of the spectral subtraction scheme followed by hidden markov model compensation framework. The estimated clean speech signal is then modeled through HMM for improving the recognition. Deng et al. [43] modeled the time-varying spectral information as a deterministic trend which is essentially a low order regression (state-dependent) of the spectral coefficients with parameters estimated as an integral component of the HMM training process. One kind of continuous density HMM (CDHMM) [120, 147] is very common in the speech recognition. In the CDHMM model, it is assumed that the feature vector distribution encountered at a given state can be modeled by a Gaussian distribution whose underlying mean represents the "true value" of the feature vector at that state. A more general approach is to associate with each state a weighted sum, or mixture of Gaussian distributions. A complete statement on the speech processing through observation vectorization and CDHMM is described in the paper [77]. Each observation can be modeled as the single component Gaussian probability density or a more generalized Gaussian mixture model [77]. The combination of time-frequency features and HMM approaches have also been successfully for animal vocalizations analyses [138, 149].

This analysis method is also applied in other fields, for example, robot application. In order to judge the best trajectory of the robot, Tso [148] proposed to apply STFT as the feature extraction tool for HMM technique to measure human behaviour consistency. A similar method is used in human behavior recognition and robot control [163]. Furthermore, in the paper [157], the HMM was utilized to characterize the time-frequency features of radar signals for the purpose of fall detection, and the feasibility of early warning was investigated. The transient signal detection through HMM is another interesting subject. In paper [78], the sonar signal is divided into a sequence of time-frequency frames, and each frame is represented by a feature in HMM for recognizing sonar transient signals. Chen [38] modeled the transient-present observations as an HMM and the transient-absent as independent random noise, and combined it with a sequential probability ratio

test to achieve the transient signal detection. The frequency line detection is a similar problem studied in [141, 119, 19]. Streit and Barrett [141] and Quinn et al. [119] used an HMM-based algorithm to track the peak of the STFT. In the literature [19], the author combined the FFTs feature (amplitude and phase values) and HMM tracker as frequency line detectors. The recurrence of periodicities in the airflow trace data is also covered in the literature. The authors in [64] proposed a spectrum-based HMM where the discrete latent state sequence reflects the time-varying changes as well as recurrence of periodic regimes as defined by their spectral properties.

HMM was used for fault diagnosis based on vibration signal since the 1990s. There are more than 150 paper focus on fault diagnosis and prognosis from 2000. Looking at the existing literature, the HMM-based methods for fault diagnosis can be roughly classified in two ways. The first kind of approach is concerned about modeling the total life process and describe the degradation process for prediction [79, 169, 16, 146, 134, 132]. In the other kind of approach, several HMM models for every fault mode are trained, and then the trained models are exploited to do fault mode classification [3, 32, 160, 61, 40]. In [161], Xin proposed a method based on short-time Fourier spectrum and HMM for separating the fault signal from the raw vibration signal. They modeled the signal pattern with different distribution in time-frequency domain as the different hidden states, and utilize the advantage of invertible mathematically of STFT to extract the time domain fault signal. After investigation, this is a novel idea for modeling the time-frequency futures of the vibration signal for diagnosis. Then they proposed a complete framework integrating detection, recognition, extraction and diagnosis in literature [162].

2.4.2 Explicit duration hidden Markov model

As a rich mathematical structure and good models for real world signal, HMMs have became increasingly popular in recent decades. It is an important class of models that are successful in many application areas. However, the standard HMM allows self-transition process, then the time the Markov chain spends in a state is a random variable that is statistically described by a geometric probability mass distribution [68]. Suppose the self-transition (from state i to state i) probability is denoted as a_{ii}. The probability of state occupancy decreases exponentially with time, and the probability of d consecutive

observations in state i can be written as

$$p(d|i) = a_{ii}^{d}(1 - a_{ii}). \tag{2.16}$$

$p(d|i)$ is the probability of taking the self-transition at state i for d times. Because of this property, the Markov chain has no idea of how long it has stayed in a state and when to leave the state. In many cases this property seems to be a strong limitation that results in inaccurate signal modeling [62]. Researchers have proposed a number of techniques to address these limitations. In 1980, Ferguson [60] modeled the self-transition process with an explicit duration probability distribution for addressing this problem.

The explicit duration hidden Markov model can be viewed as a special instance of the standard HMM. What makes waiting time hidden Markov model different from the standard HMM is that there exists a duration time d in each state. The duration distribution $p(d|j)$ can be either a discrete distribution or a continuous density, a non-parametric or a parametric density. When the underlying process enters state j for duration d with the probability $p(d|j)$, there are d observations produced. Therefore, the self-transitions are prohibited in the explicit duration hidden Markov model. This makes it suitable for use in a wide range of applications. The capacity and complexity of EDHMM are analyzed in [72], which points that standard hidden Markov model coupled with a moderate increase in overall topological complexity and state distribution parameter tying, are already well suited to handling nonexponential duration distributions. Since it was proposed, EDHMM has attracted the attention of scholars in various fields. In this section, only the application in fault diagnosis and prognosis is given.

The application to fault diagnosis and prognosis appeared gradually after 2000, in particular in the recent ten years. The related literature is synthesized in one table, seen in Table 2.1, following the diagnosis objects, the objective, the type of data and methodology. The table is supposed to provide a more direct view of reported studies to readers who are concerned with the use of EDHMM in prognostics and health management (PHM).

Dong et al. [48] were the first researchers to apply EDHMM to PHM in 2006. They validated the effectiveness of the EDHMM for the fault classification of UH-60A Blackhawk main transmission planetary carriers and prognosis of a hydraulic pump health monitoring application. He et al. pointed out the drawback of EDHMMs in [66]: the computational

TABLE 2.1: Summary of the use of EDHMM for fault diagnosis

Object	Reference	Objective	Type of observation	Methodology
planetary box	Dong et al. [48]	fault classification	time domain vibration signal	original EDHMM
	Fan et al. [57]	diagnosis and prognosis	time domain vibration signal	original EDHMM
hydraulic pump	Dong et al.[48]	prognosis	time domain vibration signal	original EDHMM
	Dong et al. [46]	fault classification and prognosis	time domain vibration signal	segmental EDHMM
	Dong et al. [47]	fault classification and prognosis	wavelet decomposition coefficients	segmental EDHMM
	Dong et al. [45]	fault classification and prognosis	wavelet decomposition coefficients	auto-regressive EDHMM
	Dong et al. [49]	RUL prediction	wavelet decomposition coefficients	non-stationary segmental EDHMM
	Dong et al. [97, 95]	RUL prediction	wavelet decomposition coefficients	joint EDHMM with sequential Monte Carlo method
	Dong et al. [96]	RUL prediction	wavelet decomposition coefficients	adaptive EDHMM
rotor	He et al. [66]	prognosis	frequency domain vibration signal	original EDHMM
shaft	Bechhoefer et al. [22]	prognosis	frequency domain vibration signal	original EDHMM
gearbox	Teng et al. [144]	RUL prediction	time-frequency domain of vibration signal	original EDHMM
	Li et al. [90]	fault detection and RUL prediction	time domain vibration signal	multivariate Bayesian control scheme based on EDHMM
	Li et al. [91]	fault detection and RUL prediction	time domain vibration signal	Optimal Bayesian control scheme based on EDHMM
hydraulic cylinder	Su et al.[142]	RUL prediction	cylinder wear data	original EDHMM
	Huang et al. [70]	fault diagnosis	wavelet decomposition coefficients	original EDHMM
	Xiao et al. [159]	RUL prediction	dynamic pressure signal	EDHMM combined with high-order particle filter
bearing	Wu et al. [158]	wear prediction	certain data(unspecified in the paper)	modified EDHMM
	Chen et al. [41]	prognosis	time domain vibration signal	mixtured non-gaussian based EDHMM
	Wang et al. [153]	prognosis	time domain vibration signal	Duration-dependent EDHMM
	Cartella et al. [36]	RUL prediction	time domain vibration signal	EDHMM combined with AIC
	Le et al. [82, 83]	RUL prediction	simulation vibration signal	Multi-Branch EDHMM
draught fan	Lin et al. [94]	prognosis	time domain vibration signal	original EDHMM
milling tool	Liu et al. [98]	RUL prediction	cutting force signal	modified EDHMM
	Duan et al. [52]	RUL prediction	cutting force signal	EDHMM combined with vector autoregressive degradation modeling
transformer	Hao et al. [65]	RUL prediction	vibration signal	time-varying EDHMM
fused deposition modeling	Wu et al. [156]	Condition monitoring	AE signal	original EDHMM
unknown object	Khaleghei et al. [76]	RUL prediction	simulation data	EDHMM

complexity may increase for inference and parameter estimation. But duration modeling by parametric probability distributions has been used to alleviate the computational burden. A method that does not require the estimation of the state duration time for prognosis was proposed in [66]. Dong et al. [45] presented a novel statistical learning approach based on auto-regressive EDHMM that took the correlation between observations into account for fault classification and RUL prediction. In the next few years, Dong et al. proposed a series of improved methods, for example the non-stationary EDHMM [49], aging factors integrated into EDHMM [117], joint with sequential Monte Carlo method [97, 95], etc., for equipment health prediction. Some researchers consider that the system is subjected to degradation due to different operating conditions, leading to different rates in deterioration evolution. To take into account such problem, multi-branch or multi-sensor EDHMM [47, 96, 82, 83, 41] were proposed. In addition, another strategy is the combination of EDHMM with various assistant methods, such as the sequential Monte Carlo method for decreasing the computational and space complexity [97, 95], AIC for evaluating the correct model configuration [36], particle filter for predicting the online state [159], etc. These application results show that better performance has been achieved by the combination strategy. In the articles of Dong [48, 45], the wavelet decomposition coefficients were utilized for the observations of EDHMM. But the goal is to develop trained EDHMMs to recognize N different states of a component for a given failure mode. Therefore, for the diagnosis of the fault, it is necessary that a separate EDHMM is trained for all possible fault types in addition to the EDHMM for normal conditions. After investigation, it is found that there is no article that really does fault diagnosis through the combination of EDHMM and time-frequency observations.

2.5 Integrated automatic diagnosis

The previous sections have focused mainly on the diagnosis techniques from different viewpoints. As mentioned before, the objective of this thesis is to present an integrated auto-diagnostic framework that includes fault detection, fault signal reconstruction, fault type identification, and fault size characterization. In this section, therefore, the automated diagnostic framework through the models introduced before is discussed and reviewed.

Bearing diagnosis is not simply to identify the fault type when a fault occurred. In the reality, bearing diagnosis is a big topic, and there are many sub-issues to be solved.

For example, the health status of the collected vibration signal is often unknown. In the first step, therefore, the signal needs to be checked whether it is collected from a faulty bearing or healthy one. If the bearing is determined to be faulty, the fault identification is the next subtask. In some cases, the fault signal extraction is also an important task. These are not enough, and the engineers also want to know the fault degree and whether the bearing can still be used. These problems all belong to the category of rolling element bearing CBM. Therefore, integrated diagnosis can be understood in general as "a field of research and application which aims at making use of present vibration signal in order to detect its degradation, diagnose faults, assess and pro-actively manage its failures". Based on this viewpoint, related references are reviewed.

The automation of fault diagnosis was primarily handicapped in the past due to the lack of appropriate techniques to represent the knowledge-based, symbolic reasoning of an expert diagnostician [31]. The authors of [31] presented a two-stage fault detection and diagnosis approach. The first stage is a generic and fast technique for detection using an index fusion and fuzzy logic approach; the second stage describes the development of a fault diagnosis method through the HMM 'codebook' training. In reference [111, 112], Ocak proposed one method that can realize automatically fault detection and diagnosis by training HMM. Faults can be detected online by monitoring the probabilities of the pretrained HMM for the normal case given the features extracted from the vibration signals, and the HMM for which the probability is maximum determines the condition of the bearing. The bearing prognosis issue was integrated into Ocak's PhD thesis in [110]. In reference [79], a fault diagnostics and prognostics algorithm based on HMM was proposed. The algorithm trained different HMM models, and then achieved the fault diagnostics and prognostics in a unified framework. Similar research can be found in many references [169, 16, 146, 145]. In reference [16], the authors proposed an integrated framework based on HMM to incorporate fault diagnostics and fault severity estimation. EDHMM can also be used to achieve the integrated auto-diagnosis framework. In reference [46], the proposed segmental EDHMM-based framework combines diagnostics and prognostics in an integrated manner. Reference [75] is another attempt, where the authors proposed an indicator based on the order-frequency spectral coherence to achieve self-running diagnosis, including detection, identification and classification of typical rolling element bearing faults. Zuo [107] proposed an integrated framework, and showed how to model the degradation process of a condition monitored device through the EDHMM and then presented how

to employ real-time condition monitoring data for online diagnostics and prognostics. Reference [102] offers a recent discussion on integrated and automated machine health monitoring. After investigation, the limitation of current studies is the lack of diagnosis by modeling the nonstationary vibration signal. In other words, in most works in the literature, they trained the different HMMs in advance, and then use them for detecting or identifying or classifying. The second limitation is that most of them did not integrate the sub-issues mentioned above into one diagnostic framework.

2.6 Discussion

Based on the above literature survey, there are some points of worthy concern. Through the available literature discussed in Chapter 2.3, it is clear that various advanced techniques have been proposed to address the different diagnostic tasks. However, these techniques are almost all independent, which requires users to have a high level of experience and knowledge when dealing with different diagnostic problems. Concerning the techniques that demonstrate attempts towards integrated and automated diagnosis in Chapter 2.5, there is still no one research that is able to integrate all the diagnostic sub-issues in one solution. Even fewer have tried to achieve these objectives in an automated way, i.e. without manual tuning of the algorithms by an expert. According to the ISO 13374-1:2003 at Condition monitoring and diagnostics of machines, the diagnosis process can be divided into data acquisition, data manipulation, state detection, fault identification, health assessment, prognostic assessment, and advisory generation. So all of these reasons push us in one direction, integrated auto-diagnosis technique. As we known, the ultimate goal of research is to apply in practice, which requires the method performed in a simple and automated way. This is exactly what the thesis is pursuing.

Following the goal mentioned above, this thesis is dedicated to propose one stochastic model based diagnostic framework, achieving automation and integration at the same time. As mentioned in the bearing vibration characteristics, the intrinsic uncertainties and nonstationarity which underlie the incipient fault vibration signal and its temporal sequential nature, made EDHMM-based approaches a perfect option for our objectives. In addition, as an advanced model, EDHMM comes with several valuable parameters useful for clarifying the transition law and finding temporal structures in the vibration signals. However, as investigated in Chapter 2.4, the HMMs-based approaches that are

implemented for the diagnostic task in the literature are basically used as a supervised way which needs massive data to train different models in advance, and then exploiting the trained models to process the testing signal. The computational complexity and the assumption of data available significantly impede the diagnosis applications in a practical scenario. Unlike the supervised way, in this thesis, the stochastic model is used as signal analysis tool to extract the fault information for the following integrated auto-diagnosis. The modeling and analysis process is illustrated in Figure 2.3. The repetitive series of transients passes a filter-bank analysis implemented by the STFT to generate observations. For such rotating vibration signals, the Fourier transform have proved good performances, but other transforms are also possible. And then under the theory of stochastic model, the useful information hidden in the vibration signal can be extracted by estimating the model parameters. In this example, there exist two different time-frequency distributions which indicate the transient state and the noisy state, respectively. In the meantime, the distribution parameters in the observation with different states can be estimated, and the duration time in different states can also be obtained. These important information will encourage us to achieve fault detection, fault signal reconstruction, fault identification and fault size characterization at once without historical data and manual intervention. Since this integrated framework does not impose too many prior knowledge and parameters, it is capable of more automation to address different sub-issues.

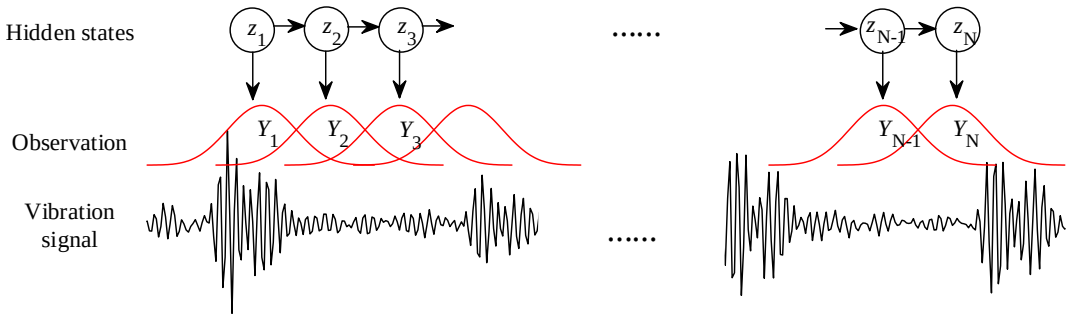

FIGURE 2.3: Illustration of HMM-based signal analysis.

2.7 Conclusion

In this chapter, the basic bearing knowledge and bearing condition monitoring techniques have been reviewed firstly. Afterward, the advantages, disadvantages, and performance of the described different types of techniques have been discussed in detail.

While it is impossible to investigate all of the existing techniques, this chapter still aims to shed some light on the strengths and drawbacks of different techniques in each diagnostic issue. Based on the investigation provided in this chapter, it is evident that effective techniques already exist in different issues, but that there is still a gap in the integrated diagnosis framework, and one in automatic way. It is found the integrated and automated diagnosis based on stochastic model of vibration signal is still a new and challenging topic. Hence, some comments have been discussed in the end, and the objectives of this thesis have been emphasized once again.

Chapter 3

Markov stochastic process

Contents

3.1 Introduction

As we know, in most situations bearing fault characteristics cannot be measured directly as the bearing vibration signature is modified by the machine structure and this situation is further complicated by other vibration coming from other equipment on the machine. Such interferences often make the interpretation of signals difficult. In probability theory, stochastic processes are the main mathematical tool for describing uncertainty. Compared with deterministic models, the stochastic models have gifted flexibility and adaptability to cope with the complexity of non-stationary phenomena. In this chapter, we now formally introduce the stochastic model, hidden Markov model and its variant. We start by reviewing the basic definitions and concepts pertaining to Markov chains, and then the hidden Markov model and its special extension, explicit duration hidden Markov model. Based on the vibration signal, a detailed estimation and analysis of the model parameters is given subsequently. Some issues with respect to the input model parameters are discussed at last.

3.2 Markov chains

In reality, sequential data are ubiquitous, for example, the rainfall measurement or temperature on successive days at a particular location, the vibration signature at successive time frames, etc. To express such data in a probabilistic model, one of the simplest ways is offered by the Markov model. It is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the observation attained in the previous event [25]. It can also be called first-order Markov chain, which is depicted as a graphical model in Figure 3.1. Here we consider a sequential data $Y_{1:N} \in M$ representing the specific stochastic process. The set M is the sample space, and the element in M is called an observation of the process. Then the Markov process can be expressed mathematically as

$$p(Y_{n+1} = j|Y_n = i, Y_{n-1} = i_{n-1}, ..., Y_1 = i_1) = p(Y_{n+1} = j|Y_n = i) = a_{ij}, \tag{3.1}$$

where, $i, j, i_1, ..., i_{n-1} \in M$. The observations are assumed to be independent and identically distributed. It is noted that a_{ij} is a fixed probability independent of time, representing the probability that the Markov process will make a transition to observation j given the

current observation i. Since probabilities are nonnegative and since the process must make a transition into a specific observation, clearly one has an important constraint as

$$a_{ij} \geq 0, \sum_{j} a_{ij} = 1, i, j \in M.$$

Let $\mathbb{A}$ denote the transition probability matrix of first-order Markov chain, so that

$$\mathbb{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MM} \end{bmatrix}. \tag{3.2}$$

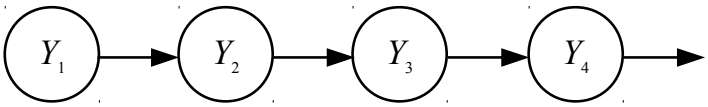

FIGURE 3.1: A first-order Markov chain of observation $\{Y_n\}$, in which the distribution $p(Y_n|Y_{n-1})$ of a particular observation Y_n is conditioned on the value of the previous observation Y_{n-1}.

According to the above description, we obtain the first-order Markov chain, which is depicted as a graphical model show in Figure 3.1. Without loss of generality, the joint distribution for a sequence of observations can be expressed through a product rule in the form,

$$\begin{aligned} p(Y_1, Y_2, ..., Y_N) &= \prod_{n=1}^{N} p(Y_n|Y_{n-1}, ..., Y_1) \\ &= p(Y_1) \prod_{n=1}^{N} p(Y_n|Y_{n-1}), \end{aligned} \tag{3.3}$$

where $p(Y_1)$ is the initial probability, and one constrain for the initial probability is $\sum_{Y_1 \in M} p(Y_1) = 1$. Thus if we use such a model to predict the next observation in a sequence, the distribution of predictions will depend only on the value of the previous observation and will be independent of all earlier observations. Although the first-order Markov chain is very simple, it provides an important modeling idea and theoretic basis for afterword probabilistic model.

Here an example related to sequential data is used to illustrate the first-order Markov

chain. Suppose that the chance of rain tomorrow (denoted as observation Y_{n+1}) depends on only whether or not it is raining today (denoted as Y_n), but not on past weather conditions. Suppose also that if it rains today, then it will rain tomorrow with probability p_1; and if it does not rain today, then it will rain tomorrow with probability p_2. If we say that the process is in state 0 when it rains and state 1 when it does not rain, then the above-mentioned process is a two-state Markov chain whose transition probability matrix is given by,

$$\mathbb{A} = \begin{bmatrix} p_1 & 1-p_1 \\ p_2 & 1-p_2 \end{bmatrix}, \tag{3.4}$$

and the transition probabilities are $p_1 = p(Y_{n+1} = 0|Y_n = 0)$, $p_2 = p(Y_{n+1} = 1|Y_n = 0)$.

3.3 Hidden Markov model

However, for some cases, the observations are complex and infinite, and the corresponding states are finite but invisible. In such cases, each state randomly generates 1 out of every k observations visible to us. We can model this by introducing additional hidden states to permit a general case. Let the hidden state sequence $\{z_n, n = 1, 2, 3, ...\}$ be a Markov chain with the transition probabilities a_{ij} and initial probabilities $\pi_i = p(z_1 = i)$, $i \in M$ defined in the previous section. Further, suppose that when the Markov chain enters state i at time n, then the observation Y_n is emitted with probability $p(Y_n|z_n = i)$. Therefore, a model of the previous defined type in which the sequence $\{Y_n, n = 1, 2, 3, ...\}$ is observed, while the sequence of underlying Markov chain states $\{z_n, n = 1, 2, 3, ...\}$ are unobserved, is called a hidden Markov model (HMM). It is an extension of a Markov chain which is able to capture the sequential relations among the hidden variables. The sequential data is represented using a Markov chain of hidden states, with each observation conditioned on the corresponding hidden state, as shown in Figure 3.2.

According to the above introduction, a HMM can be governed by the following parameters:

- **Number of states N**

  This is usually set to the total number of distinct, or elementary, stochastic events in a signal process.

- **Transition probability matrix**

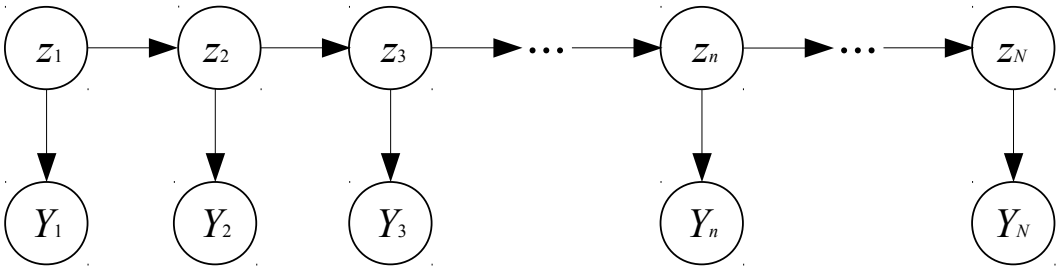

FIGURE 3.2: The graphical structure of a hidden Markov model, in which each observation Y_n is condition on the state of the corresponding hidden state and the state sequence is a first order Markov chain.

It is a constant matrix that provides a Markovian connection network between the hidden states, and models the variations in the duration of the signals associated with each state.

- **Initial probability vector**

  This is different from a transition probability as it is the first state, and so it has a marginal distribution represented by a vector of probabilities $\boldsymbol{\pi}$ with elements $\pi_i = p(z_1 = i), i \in M$.

- **Emission probability**

  This is a conditional probability from hidden state to the observed variable, $p(Y_n|z_n)$. If the observation is discrete, this probability will be a matrix that models the transition from the state to the observation. If the observation is continuous, this probability further depends on the specific parameter $\boldsymbol{\phi}$ according to a probability distribution. Because Y_n is observed, the distribution $p(Y_n|z_n)$ consists of a vector of N numbers corresponding to the N possible states.

These HMM parameters can be described as the parameter set $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbb{A}, \boldsymbol{\phi}\}$, where, $\boldsymbol{\phi}$ represents the emission probability matrix or the parameter in the observation distribution. For Gaussian observations, the distribution parameter $\boldsymbol{\phi}$ include the covariance and the mean value. Then, the joint distribution for this model is given by,

$$p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}_{1:N}|\boldsymbol{\theta}) = p(z_1)\left[\prod_{n=2}^{N} p(z_n|z_{n-1})\right]\prod_{n=1}^{N} p(x_n|z_n). \tag{3.5}$$

A useful way of interpreting and using HMMs is to consider each state of an HMM as a model of a segment of a stochastic process. One example about the vibration signal

is used to illustrate the hidden Markov model, which is depicted in Figure 3.3. The upper subplot of Figure 3.3 depicts two states Markov chain z_n, one state represents the noise interval, and another state for transient part. The lower subplot shows the noisy vibration signal observed through a sensor mounted on the rotating machine. Each specific segment of the signal has one corresponding hidden state. In this example, the observations are the continuous vibration signal. A suggested way to model it is to segment the signal into short intervals. The intervals are described by a specific distribution with different parameters. Short time windowed segments is a common way to deal with a continuous signal. In the next section, the detailed description about how to model the vibration signal will be given.

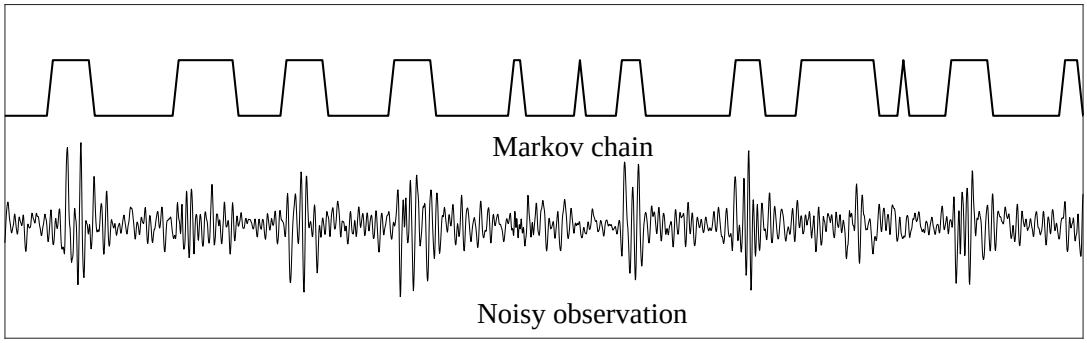

FIGURE 3.3: Example of a binary state Markov chain in a noisy signal. The continuous signal is segmented by short windows and conditioned on the corresponding hidden state.

The estimation task of prime interest for HMMs is to estimate the model parameters $\boldsymbol{\theta}$ and the hidden sequence $\{z_n\}$ from the observed sequence $\{Y_n\}$ in some optimal way. We will discuss how to estimate the parameters in the later section of this chapter, and the convergence of the algorithm in the appendix.

3.4 Vibration signal modeling

HMM is a versatile stochastic model for nonstationary phenomena that involves transitions between hidden states. It is therefore well suited for describing the time-frequency spectrum coefficients of a faulty signal, whose probability distribution switches between one state corresponding to the occurrence of impulses and another one corresponding to dead zones [161]. However, in the standard HMM the transition probability matrix is assumed constant, which implies that the time spent in a state is a random variable that

follows a geometric probability distribution [68]. For bearing signals, this is an unrealistic assumption. To fix this drawback, it is proposed to introduce an explicit duration hidden Markov model [60] (also known as "segmental HMM [126]" and "variable-duration HMM [143]", as part of the larger theory of hidden semi-Markov models (HSMM) [166]). In this section we introduce one useful variant, EDHMM. We will give a detailed introduction about EDHMM, and how to use it to model vibration signal.

3.4.1 Signal model and STFT decomposition

In order to motivate the EDHMM, let first consider the signal model. Let $y(t)$ denotes the measured vibration signal, $x(t)$ the series of impulses due to impacts on a fault and $n(t)$ the background noise, which, by definition, embodies all other sources of vibration that are not of interest for diagnosis. Clearly, in the healthy state,

$$y(t) = n(t),$$

whereas in the faulty state

$$y(t) = x(t) + n(t)$$

is a superposition of transients and noise. The faulty signal will be modeled hereafter as a nonstationary signal that switches between two states: an “active” state where an impulse occurs and an “inactive” state where noise only is observable.

As discussed above, the continuous observation need to be segmented before modeling. How to segment the vibration signal is an essential component of modeling, often having an important impact on the final result. The main concern in the segment action is the information extraction completeness. In the present work, the STFT is used as a versatile transform for decomposing nonstationary vibration signals with the ability of evidencing possible transitions between two different states. This is because the STFT decomposition and its inverse transformation establish a one-to-one mapping between the time domain and the time-frequency domain, and its oscillating basis functions are well-fitted to represent vibration signals. Also, a nonstationary signal can be considered as quasi-stationary over a short time interval. As it preserves information from the original signal, the STFT ensures that important features are not lost as a result of the transformation.

The STFT coefficients $Y(n, k)$ of a discrete-time signal $y(t), t = 0, 1, 2, ..., L - 1$ over

a time interval of length N_w is given by

$$Y(n,k) = \sum_{m=0}^{N_w-1} y(nR+m)w(m)e^{-j2\pi m\frac{k}{N_w}}, \tag{3.6}$$

where $n = 1, 2, ..., N$, $N = \lfloor (L - N_w)/R \rfloor + 1$ and $k = 1, 2, ..., N_k$, $N_k = \lfloor N_w/2 \rfloor$, represent the time and frequency indices, respectively; $w(m)$ denotes a positive and smooth N_w-long data window, and R represents the window shift.

In order to better use the STFT coefficients $Y(n,k)$, three assumptions concerning $Y(n,k)$ will be needed:

- *i*) $Y(n,k)$ and $Y(n',k')$ are independent for any $k \neq k'$ and $n \neq n'$,
- *ii*) the expected value of the STFT coefficients is zero,
- *iii*) the real $\Re\{Y(n,k)\}$ and imaginary $\Im\{Y(n,k)\}$ parts are independent and identically distributed.

The assumptions are made to simplify the model, but without losing reasonableness. The windowed signal will tend to be stationary, which means the frequency bin k and k' under one window are uncorrelated. Then the reasonableness of assumption i is only depend on the window shift R. If the shift is large enough, n and n' will be independent. Assumption *ii* and *iii* are also due to the stationary characteristic of windowed signal.

Under mild conditions, the vector of STFT coefficients $\boldsymbol{Y}_n = [Y(n,1), ..., Y(n,N_k)]$ converges in distribution to the complex-valued Gaussian distribution according to the central limit theorem [33], as expressed in Eq.(3.7). More reasons justifying the Gaussian assumption of the STFT coefficients can be found in reference [116]. As will be seen shortly, this property is advantageous in following modeling.

$$\boldsymbol{Y}_n \sim \mathcal{CN}(\boldsymbol{\mu}, \mathbb{C}, \mathbb{M}). \tag{3.7}$$

The covariance of this distribution is $\mathbb{C} = E[(\boldsymbol{Y}_n - \boldsymbol{\mu})(\boldsymbol{Y}_n - \boldsymbol{\mu})^\dagger]$, where $\dagger$ means the complex conjugate of the transpose, and the pseudo-covariance is $\mathbb{M} = E[(\boldsymbol{Y}_n - \boldsymbol{\mu})(\boldsymbol{Y}_n - \boldsymbol{\mu})^T]$. According to assumption i, the covariance of two random variables with different frequency is zero; according to assumption *ii*, the mean value $\boldsymbol{\mu}$ is zero. Therefore, the covariance $\mathbb{C}$ can be simplified as a diagonal matrix with zeros in the off-diagonal elements as given by

Eq.(3.8):

$$\begin{aligned}\mathbb{C} &= E[(\boldsymbol{Y}_n - \boldsymbol{\mu})(\boldsymbol{Y}_n - \boldsymbol{\mu})^\dagger] \\ &= E[\boldsymbol{Y}_n \boldsymbol{Y}_n^\dagger] \\ &= \begin{bmatrix} C_{11} & 0 & \cdots & 0 \\ 0 & C_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_{N_k N_k} \end{bmatrix},\end{aligned} \tag{3.8}$$

where $C_{kk} = |\Re(Y(n,k))|^2 + |\Im(Y(n,k))|^2$, $k \in \{1, 2, ..., N_k\}$ is the sum of squares of real and imaginary parts, and represents the energy of the signal at the datum n with frequency k. Similiarly, the pseudo-covariance $\mathbb{M} = E[(\boldsymbol{Y}_n - \boldsymbol{\mu})(\boldsymbol{Y}_n - \boldsymbol{\mu})^T]$ can be deduced as a zero matrix. Then, the Gaussian random variables $\boldsymbol{Y}_n$ are only determined by their covariance matrix $\mathbb{C}$, i.e. $\boldsymbol{Y}_n \sim \mathcal{CN}(0, \mathbb{C})$. It is also called a circularly-symmetric Gaussian distribution, whose distribution, $p(\boldsymbol{Y}_n)$, is given by,

$$p(\boldsymbol{Y}_n) = \frac{\exp\{-\boldsymbol{Y}_n^\dagger \mathbb{C}^{-1} \boldsymbol{Y}_n\}}{\pi^{N_k} |\mathbb{C}|}. \tag{3.9}$$

It is worth noting that the STFT is used here to pre-process the data and not as a visual diagnostic tool – i.e. as with time-frequency based approaches; therefore, the uncertainty principle to which it is subjected – which limits the time-frequency resolution – is not really an issue. It is noted here that other transforms are also possible. The columns of the coefficient matrix $\boldsymbol{Y}_n$ will be used as the observation sequence of the following modeling.

3.4.2 Explicit duration hidden Markov model

EDHMM [60, 166] is an extension of hidden Markov models that includes the case where the duration time in a state is described by a specific distribution, which is the origin of the term "explicit duration". It is used in this section to describe the temporal relationship among the STFT coefficients. Each observation $\boldsymbol{Y}_n$ belongs to one unknown state, denoted as z_n. Then the corresponding state sequence of the observations $\boldsymbol{Y}_{1:N}$ is denoted by $z_{1:N} \triangleq \{z_1, z_2, ..., z_N\}$. In the present application, the faulty bearing vibration signal comprises two states:

1. "inactive" state, hereafter labeled #1, where noise only is present,

2. "active" state, labeled #2, where an impulse occurs.

Therefore, the hidden states $z_{1:N}$ take two values, $z_n \in \{1, 2\}$. In each state i, the whole duration is denoted as d_i, which means there are d_i successive observations of the signal belonging to the state i. Here, the Poisson distribution with parameter λ_i, $d_i \sim P(\lambda_i)$, is used to model the discrete variable d_i, although other distribution choices are equally possible,

$$p(d_i = k) = \frac{\lambda_i^k}{k!} e^{-\lambda_i}, k = 0, 1, ..., \tag{3.10}$$

where, the Poisson parameter λ_i represents the average window number in state i, which will be a key parameter for the following diagnosis. One advantage of Poisson distribution is that it has only one parameter, and all the distribution information can be obtained through this parameter. In addition, from the characteristic of Poisson distribution, the duration time in different transient pulses is very similar, which makes it well suited for Poisson distribution. The simplified model is shown in Figure 3.4. This extends the standard HMM to the domain of modeling the temporal structure of hidden state sequence. From now on, one will need to define the dynamic duration time $\tau_n(i)$, the time already spent in the current state i at datum n. If the entry and exit in state i occur at time instants n_1 and n_2, respectively, then $\tau_{n_1-1}(i) = 0$, $\tau_{n_1+m}(i) = m$ for $m = 0, 1, ..., n_2 - n_1$, and $\tau_{n_2}(i) = n_2 - n_1 + 1$ is one sample of the discrete variable d_i.

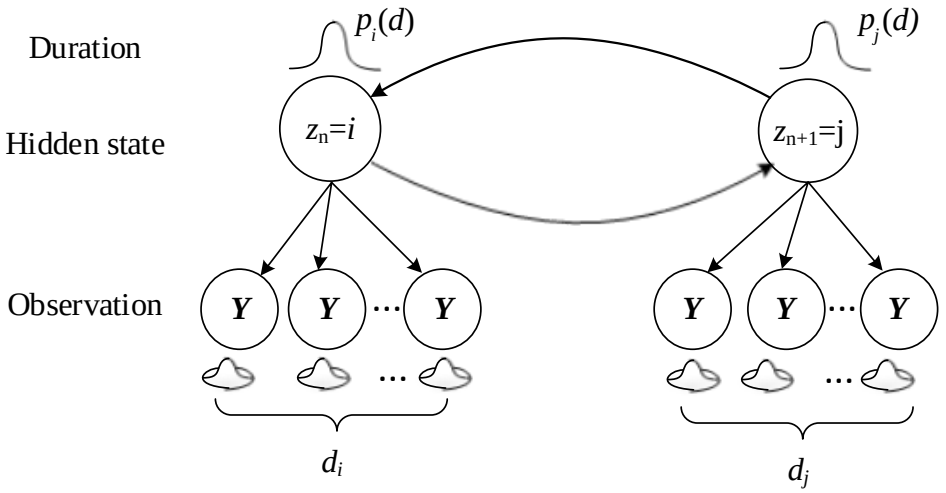

FIGURE 3.4: Explicit duration hidden Markov model, in which one state emits d_i values of observations, and each observation follows a complex-valued Gaussian distribution, $i, j \in \{1, 2\}$.

The transition probability matrix $\mathbb{A}$ in EDHMM is an 2×2 dynamic matrix that depends on $\tau_n(i)$, rather than a constant matrix as in the standard HMM. The diagonal elements, $a_{ii}(\tau_n)$, represent the recurrent state transition probabilities. These two elements are defined as the probability of remaining in the current state i at the next time step, given the time, $\tau_n(i)$, already spent in the current state i. Since the duration time follows the Poisson distribution, the recurrent probability $a_{ii}(\tau_n)$ will decrease with τ_n. Then the recurrent probability $a_{ii}(\tau_n)$ can be expressed through the Poisson cumulative distribution function (CDF) as,

$$\begin{aligned}
&a_{ii}(\tau_n(i)) = p(z_{n+1} = i | z_n = i, \tau_n(i)) \\
&= p(z_{n+1} = i | z_n = i, ..., z_{n-\tau_n(i)+1} = i, z_{n-\tau_n(i)} \neq i) \\
&= \frac{p(z_{n+1} = i, ..., z_{n-\tau_n(i)+2} = i | z_{n-\tau_n(i)+1} = i, z_{n-\tau_n(i)} \neq i)}{p(z_n = i, ..., z_{n-\tau_n(i)+2} = i | z_{n-\tau_n(i)+1} = i, z_{n-\tau_n(i)} \neq i)} \\
&= \frac{p(d_i > \tau_n(i))}{p(d_i > \tau_n(i) - 1)} \\
&= \frac{1 - p(d_i \leq \tau_n(i))}{1 - p(d_i \leq \tau_n(i) - 1)} \\
&= \frac{1 - F(\tau_n(i) | \lambda_i)}{1 - F(\tau_n(i) - 1 | \lambda_i)},
\end{aligned} \tag{3.11}$$

where $F(\tau_n(i)|\lambda_i)$ represents the CDF of Poisson distribution with parameter λ_i. The numerator, $1 - F(\tau_n(i)|\lambda_i)$, represents the probability that the explicit duration Markov process, at datum n, has stayed in state i for at least $\tau_n(i)$ samples. Figure 3.5 shows one example of Poisson distribution with the parameter $\lambda = 5$. From the CDF in the lower subplot and Eq.(3.11), it is found that the recurrent transition probability $a_{ii}(\tau_n(i))$ will gradually decrease as $\tau_n(i)$ increase. Eventually, the Markov chain will transit to state j from state i, and then go to the next Poisson process.

According to the recurrent probability defined above and the constraint $\sum_j a_{ij} = 1$, the dynamic transition probability matrix $\mathbb{A}(\tau_n(i))$ can be written as

$$\begin{aligned}
\mathbb{A}(\tau_n(i)) &= [a_{ij}(\tau_n(i))] \\
&= [p(z_{n+1} = j | z_n = i, \tau_n(i))] \\
&= \begin{bmatrix} a_{11}(\tau_n(1)) & 1 - a_{11}(\tau_n(1)) \\ 1 - a_{22}(\tau_n(2)) & a_{22}(\tau_n(2)) \end{bmatrix}.
\end{aligned} \tag{3.12}$$

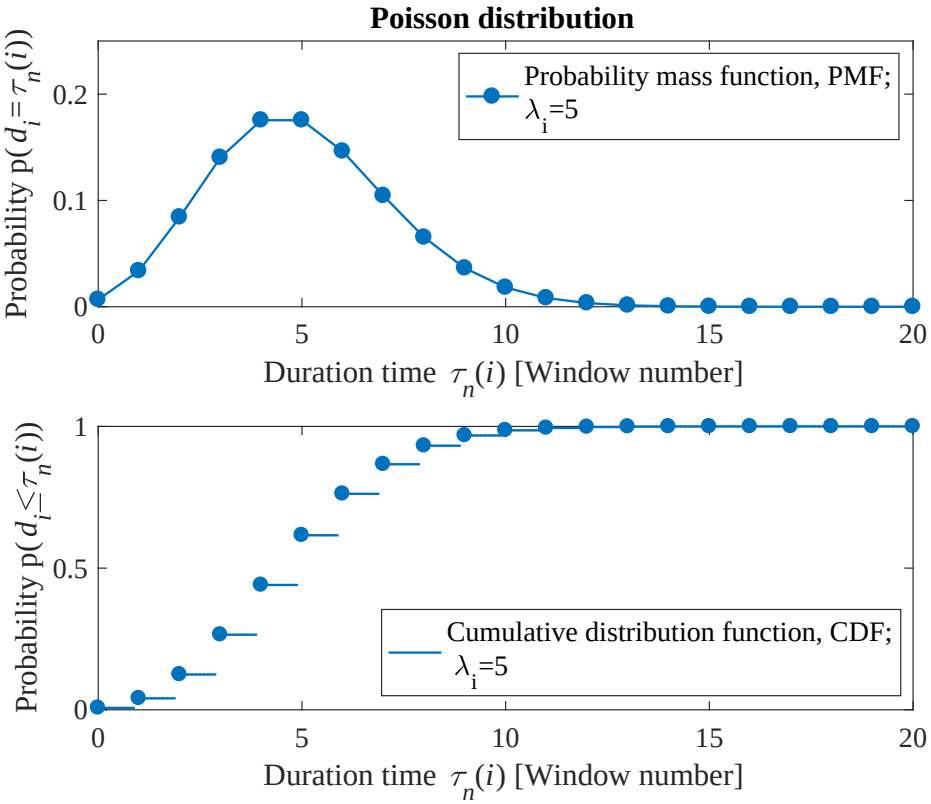

FIGURE 3.5: Illustration of Poisson distribution. The upper subplot is the PMF with the horizontal axis as the duration time $\tau_n(i)$, i.e. the number of remainings in current state i. λ_i is the expected duration time in state i; The lower subplot is the CDF showing discontinuous at the integers of $\tau_n(i)$ because a variable that is Poisson distributed takes on only integer values.

where, the two off-diagonal elements, $a_{ij}(\tau_n)(i \neq j)$, is defined as the transition probabilities between two different states, given the previous state i has lasted $\tau_n(i)$. For the two-states case, the relationship between the recurrent probability and non-recurrent probability is fixed, $a_{ij}(\tau_n) = 1 - a_{ii}(\tau_n(i))$, under the constraint $\sum_j a_{ij} = 1$.

In this definition, $\tau_n(i)$ can be adaptively optimized through the duration distribution parameters in each datum n. Therefore, the dynamic matrix $\mathbb{A}(\tau_n(i))$ will also vary regularly according to the Poisson parameter, which does not need to be estimated in the following re-estimation section. Then the state sequence generated by this dynamic matrix will have a time-varying mean and variance relating to Poisson characteristic. It will actually tend to be pseudo-cyclostationary, which is a more relevant model for the vibration signals of interest in this thesis. This is in marked contrast with the standard HMM, which involves a constant transition matrix, and therefore can only produce stationary state sequences

without noticeable cyclicity in the timing of their switches. Figure 3.6 shows a simulated example that illustrates the difference between the state sequences produced by HMM and EDHMM. The state sequence in Figure3.6(a) is generated from a dynamic transition matrix modeled by the Poisson distribution with parameters $\boldsymbol{\lambda} = (5, 20)$, while the the state sequence in Figure3.6(c) is generated from the corresponding equivalent constant transition matrix. Their spectral analysis displayed in Figure3.6(b) and Figure3.6(d) illustrates the unique ability EDHMM to produce a pseudo-cyclostationary sequence with cycle around 100 Hz.

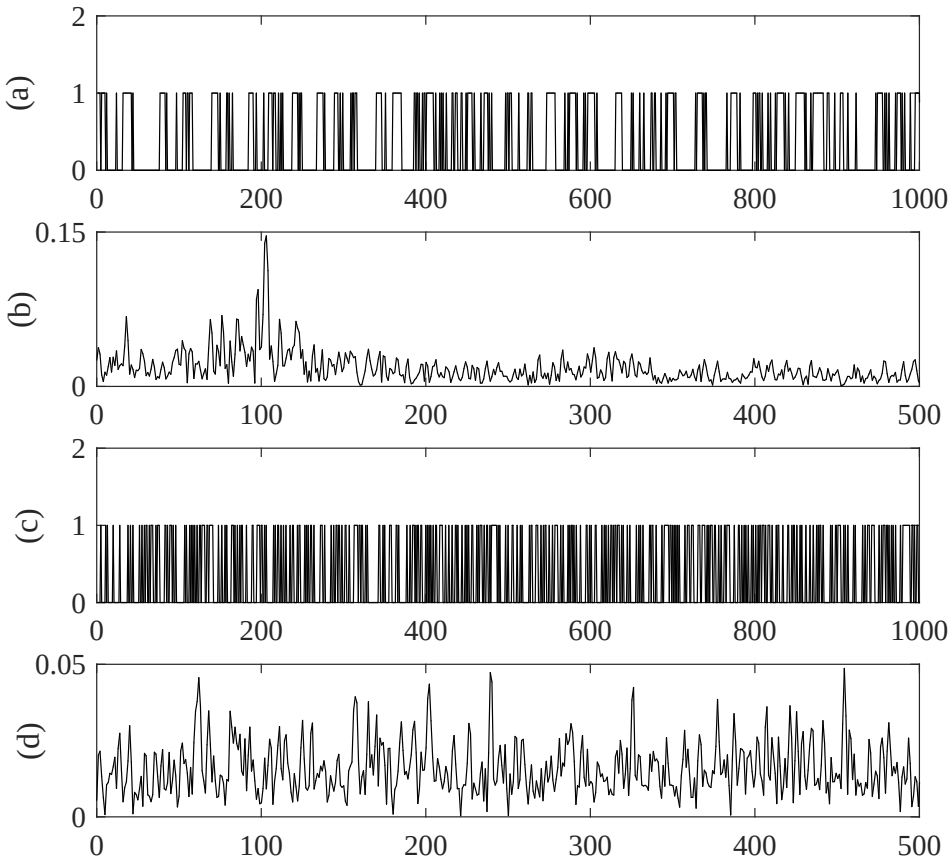

FIGURE 3.6: Illustration of the difference between a dynamic and constant transition matrix. (a) The EDHMM state sequence and (b) its Fourier spectrum; (c) the corresponding HMM state sequence and (d) its Fourier spectrum.

In EDHMM, transitions between states start from an initial probability ($n = 1$) denoted π_i, $i \in \{1, 2\}$, defined as $\pi_i = p(z_1 = i|\boldsymbol{Y}_{1:N})$ with the constraint $\sum_i \pi_i = 1$. The probabilities of different states are then updated according to transitions and observations. The modeling stage is completed up to now, and the EDHMM is completely described by the set of parameters $\boldsymbol{\theta} = \{\pi_1, \boldsymbol{\lambda}, \mathbb{C}\}$ (note that $\pi_2 = 1 - \pi_1$), where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]$ and

$\mathbb{C} = [\mathbb{C}_1, \mathbb{C}_2]$. The next section addresses the estimation of these parameters, and explains what characteristics of the signal these parameters reveal.

3.5 Parameter estimation and analysis

3.5.1 Parameter estimation

The goal is to find the optimal model parameter set $\boldsymbol{\theta}$ that maximizes the likelihood $p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})$ from the observations $\boldsymbol{Y}_{1:N}$. The forward-backward recursion formulas are first introduced to calculate the likelihood. The forward and backward variables at datum n are defined as

$$\alpha_n(i) = p(\boldsymbol{Y}_{1:n}, z_n = i|\boldsymbol{\theta}), \tag{3.13}$$

$$\beta_n(i) = p(\boldsymbol{Y}_{n+1:N}|z_n = i, \boldsymbol{\theta}). \tag{3.14}$$

The quantity $\alpha_n(i)$ represents the joint probability of observing all of the given data up to time n and the value of state i, where $\beta_n(i)$ represents the conditional probability of all future data from time $n+1$ up to N given the value of state i. Therefore, α and β are the forms of $2\times N$ matrices for the two states model.

In order to calculate the forward and backward variables, the recursion relations are derived using the conditional independence properties together with the sum and product rules. The forward quantity $\alpha_n(i)$ is expressed in terms of $\alpha_{n-1}(i)$ as,

$$\begin{aligned}
\alpha_n(j) &= p(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_n|z_n = j)p(z_n = j) \\
&= p(\boldsymbol{Y}_n|z_n = j)p(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_{n-1}|z_n = j)p(z_n = j) \\
&= p(\boldsymbol{Y}_n|z_n = j) \sum_{z_{n-1}=i} p(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_{n-1}, z_{n-1} = i, z_n = j) \\
&= p(\boldsymbol{Y}_n|z_n = j) \sum_{z_{n-1}=i} p(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_{n-1}, z_n|z_{n-1} = i)p(z_{n-1} = i) \\
&= p(\boldsymbol{Y}_n|z_n = j) \sum_{z_{n-1}=i} p(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_{n-1}|z_{n-1} = i)a_{ij}(\tau_{n-1})p(z_{n-1} = i) \\
&= p(\boldsymbol{Y}_n|z_n = j) \sum_{z_{n-1}=i} p(\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_{n-1}, z_{n-1} = i)a_{ij}(\tau_{n-1}) \\
&= p(\boldsymbol{Y}_n|z_n = j) \left[\sum_i \alpha_{n-1}(i)a_{ij}(\tau_{n-1}) \right].
\end{aligned} \tag{3.15}$$

With the same manipulation, the recursion relation for quantity $\beta_n(i)$ can be formulated as,

$$\beta_n(i) = \sum_j \beta_{n+1}(j) a_{ij}(\tau_n) p(\boldsymbol{Y}_{n+1}|z_{n+1} = j). \tag{3.16}$$

It is seen these two recursion formulas contains the dynamic duration time $\tau_n(i)$. Therefore, to calculate the formulas above, the update formula for $\tau_n(i)$ is required [17], as follows,

$$\begin{aligned} \tau_n(i) &= p(z_n = i|\boldsymbol{Y}_{1:n}, \boldsymbol{\theta})\tau_{n-1}(i) + 1 \\ &= \frac{\alpha_n(i)}{\sum_i \alpha_n(i)}\tau_{n-1}(i) + 1. \end{aligned} \tag{3.17}$$

From now on, several other posterior probabilities are defined for facilitating the subsequent re-estimation. First, the posterior probability of the current state z_n being equal to i at datum n, given the observations $\boldsymbol{Y}_{1:N}$ and parameter set $\boldsymbol{\theta}$, denoted as $\gamma_n(i)$, is expressed as

$$\begin{aligned} \gamma_n(i) &= p(z_n = i|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}) \\ &= \frac{p(\boldsymbol{Y}_{1:N}|z_n = i, \boldsymbol{\theta})p(z_n = i|\boldsymbol{\theta})}{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})} \\ &= \frac{p(\boldsymbol{Y}_{1:n}, z_n = i|\boldsymbol{\theta})p(\boldsymbol{Y}_{n+1:N}|z_n = i, \boldsymbol{\theta})}{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})} \\ &= \frac{\alpha_n(i)\beta_n(i)}{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})}. \end{aligned} \tag{3.18}$$

The sequence $\gamma_{1:N}(i)$ will be used for identifying the fault type in Chapter 4.4. According to this equation, there are some useful remarks worth noting. On one hand, if setting $n = N$ and replacing $\alpha_N(i)$, we can obtain one relationship as

$$\begin{aligned} \gamma_N(i) &= p(z_N = i|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}) \\ &= \frac{p(\boldsymbol{Y}_{1:N}, z_N = i|\boldsymbol{\theta})\beta_N(i)}{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})} \\ &= p(z_n = i|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta})\beta_N(i), \end{aligned} \tag{3.19}$$

from which it can be seen that $\beta_N(i) = \frac{\gamma_N(i)}{p(z_n=i|\boldsymbol{Y}_{1:N},\boldsymbol{\theta})} = 1$ for all the states i. This inference will be useful in the derivation of likelihood function. As we know, the likelihood function $p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})$ is the indicator for the convergence criterion of the algorithm, so it drives the

decision as whether the iterative algorithm is to be stopped or not. On the other hand, if we sum both side of Eq.(3.18), and use the fact that the left-side is a normalized distribution, the likelihood function $p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})$ can be written as

$$p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta}) = \sum_i \alpha_n(j)\beta_n(i). \tag{3.20}$$

Then the likelihood function can be estimated for any convenient choice of n. In order to simplify the algorithm, $n = N$ is selected for removing $\beta_n(i)$ according to the inference form Eq.(3.19), i.e, $\beta_N(i) = 1$. The likelihood function can be simplified as

$$p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta}) = \sum_i \alpha_N(i). \tag{3.21}$$

Next, another useful posterior probability, denoted as $\xi_n(i,j)$, is the probability of a transition from state i to state j given the observations $\boldsymbol{Y}_{1:N}$, which can be expressed in terms of the forward-backward variables and model parameters as,

$$\begin{aligned}
&\xi_n(i,j) = p(z_n = i, z_{n+1} = j|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}) \\
&= \frac{p(\boldsymbol{Y}_{1:N}|z_n = i, z_{n+1} = j, \boldsymbol{\theta})p(z_n = i, z_{n+1} = j|\boldsymbol{\theta})}{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})} \\
&= \frac{\alpha_n(i)\beta_{n+1}(j)a_{ij}(\tau_n(i))p(\boldsymbol{Y}_{n+1}|z_{n+1} = j, \boldsymbol{\theta})}{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})}.
\end{aligned} \tag{3.22}$$

According to the above formulas, the parameters $\boldsymbol{\theta} = \{\pi_1, \boldsymbol{\lambda}, \mathbb{C}\}$ can be estimated as follows. The initial state probability π_i estimate is

$$\hat{\pi}_i = \gamma_1(i)/\sum_i \gamma_1(i), \tag{3.23}$$

where the numerator $\gamma_1(i)$ can be understood as the probability of occurrence of $z_1 = i$, and the denominator, $\sum_i \gamma_1(i)$, as the possibility of z_1 over all states. The estimate of the covariance $\mathbb{C}_i$ is (See details in appendix A),

$$\hat{\mathbb{C}}_i = \frac{\sum_{n=1}^N \gamma_n(i)\boldsymbol{Y}_n\boldsymbol{Y}_n^\dagger}{\sum_{n=1}^N \gamma_n(i)}. \tag{3.24}$$

In the case of two-state EDHMM, one has the useful relationships

$$\hat{\pi}_1 + \hat{\pi}_2 = 1 \quad \text{and} \quad \hat{\mathbb{C}}_1 + \hat{\mathbb{C}}_2 = \frac{\sum_{n=1}^{N} \boldsymbol{Y}_n \boldsymbol{Y}_n^{\dagger}}{N}. \tag{3.25}$$

In the Poisson distribution, the maximum likelihood estimate of the parameter from the samples is the sample mean. Therefore, the parameter $\boldsymbol{\lambda}$ in the duration distribution is estimated as

$$\hat{\lambda}_i = \sum_{n=1}^{N-1} \xi_n(i,j)\tau_n(i) \qquad (i \neq j), \tag{3.26}$$

where the weight $\xi_n(i,j)$ represents the probability of transition from state i to state j, and the condition $(i \neq j)$ means the weighted average of $\tau_n(i)$ is calculated only at the transition time. Eqs.(3.23) to (3.26) complete one updating process for estimating the parameters $\boldsymbol{\theta} = \{\pi_1, \boldsymbol{\lambda}, \mathbb{C}\}$, which is then iterated. The iteration stop criterion depends on the increase rate of the likelihood function $p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})$, shown in Eq.(3.21). The proof of this iterative convergence is demonstrated in Appendix B. The full estimation procedure is summarized as follows,

1) Assume an initial parameter set $\boldsymbol{\theta}^{old}$.

2) Compute the forward-backward variables $\alpha_n(i)$ and $\beta_n(i)$ through Eqs.(3.15) and (3.16). Next, calculate the forward-backward variables and the related probability $\gamma_n(i)$ and $\xi_n(i,j)$ through Eqs.(3.18) and (3.22). Finally, etimate the new parameter set $\boldsymbol{\theta}^{new}$ through Eqs.(3.23) to (3.26).

3) Check for the stop criterion of either the logarithmic likelihood or the parameters. If the stop criterion is not satisfied, the let $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$ and return to step 2).

The pseudo-code is summarized in Algorithm 1. It is noted that it involves two parameters: *max_iter* for the maximum number of iterations and *Threshold* for the expected relative tolerance between the logarithmic likelihood probabilities of two consecutive times. These two parameters can be easily set by default. Also, the STFT parameters, N_w, R, and the EDHMM parameter $\boldsymbol{\theta} = \{\pi_1, \boldsymbol{\lambda}, \mathbb{C}\}$ will be discussed in detail in Chapter 3.6.

3.5.2 Parameter analysis

The parameters of EDHMM be estimated in the previous section, will be used to construct the integrated diagnosis framework of the next chapter. The parameters of a

Algorithm 1 EDHMM

Input: $y(t)$, N_w, R, max_iter, Threshold
STFT:$\longrightarrow \boldsymbol{Y}_{1:N}$
EDHMM Initialization: $\boldsymbol{\theta} = \{\pi_1, \boldsymbol{\lambda}, \mathbb{C}\}$
for k <max_iter **do**
→ Pre-process
$\tau_n(i) = \frac{\alpha_n(i)}{\sum_i \alpha_n(i)} \tau_{n-1}(i) + 1$
$\gamma_n(i) = \frac{\alpha_n(i)\beta_n(i)}{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})}$
$\xi_n(i,j) = \frac{\alpha_n(i)\beta_{n+1}(j)a_{ij}(\tau_n(i))p(\boldsymbol{Y}_{n+1}|z_{n+1}=j,\boldsymbol{\theta})}{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})}$
→ Update the parameter $\boldsymbol{\theta} = \{\pi_1, \boldsymbol{\lambda}, \mathbb{C}\}$
$\hat{\pi}_i = \gamma_1(i) / \sum_i \gamma_1(i)$
$\hat{\mathbb{C}}_i = \frac{\sum_{n=1}^{N} \gamma_n(i) \boldsymbol{Y}_n \boldsymbol{Y}_n^{\dagger}}{\sum_{n=1}^{N} \gamma_n(i)}$
$\hat{\lambda}_i = \sum_{n=1}^{N-1} \xi_n(i,j)\tau_n(i) \qquad (i \neq j)$
→ **Stop criteria**
if $\frac{\log p^k(\boldsymbol{Y}_{1:N}) - \log p^{k-1}(\boldsymbol{Y}_{1:N})}{|\log p^k(\boldsymbol{Y}_{1:N})| + |\log p^{k-1}(\boldsymbol{Y}_{1:N})|}$ <Threshold **then**
Stop
else
$k \leftarrow k+1$; $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$
Endif
Endfor
Output: output result

random process determine the characteristics of the signals generated by the process. As the characteristics of the signals change, so do the corresponding parameters. So in this section, a synthetic signal,

$$y(t) = x(t) + n(t),$$

is used to demonstrate the function of the output parameters and what characteristics they reveal.

The fault signal $x(t)$ is a simulated bearing inner race fault, and white noise $n(t)$ is set to a noise-to-signal-ratio of 0 dB. The signal is gendered with a resonance frequency $f_0 = 5000Hz$, and modulated by a low fault frequency $f_{BPFI} = 163Hz$. The rotating frequency, f_r, is set as $30Hz$ and the signal length is $L = 48000$ samples in one second, as shown in Figure 3.7. The easy signal is generated in order to better understand the meaning of the EDHMM parameters.

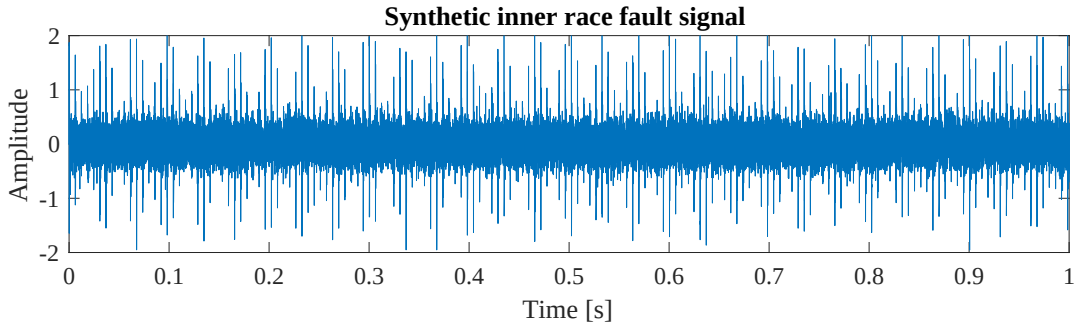

FIGURE 3.7: The synthetic signal with SNR=0 dB and sampling frequency $f_s = 4.8kHz$ over one second.

Covariance $\hat{\mathbb{C}}$

The positive diagonal matrix $\hat{\mathbb{C}}_i$ is the covariance of the observations in state i, as described in Eq.(3.8). The diagonal elements, $[\hat{\mathbb{C}}_i]_{kk} = |\Re(Y(n,k))|^2 + |\Im(Y(n,k))|^2$, $k \in \{1, 2, ..., N_k\}$, are the estimates of the variance at frequency k. Therefore, this parameter is a good indicator for displaying the energy distribution in the frequency domain [162]. It can also be used as a criterion for deciding whether the signal is healthy (see next Chapter 4.2). It also determined the frequency location of the transient pulses. In this sense, it can also be used for frequency band selection. Figure 3.8 shows the spectrogram of the synthetic signal; the frequency resolution is set to $\Delta f = f_s/N_w = 375Hz$ and a window shift $R = 16$ is used.

From the spectrogram, it is found that the covariance of the transient state indicated by a solid line has a high energy in the band centred on $5000Hz$, exactly where the resonance frequency is, and low energy in other areas. By contract, the noise state indicated by dashed line has low energy in all frequency bands.

Sequence τ_n and posterior probability $\gamma_{1:N}(i)$

As mentioned earlier, each time a defect strikes a mating surface, a pulse of short duration is generated that excites the resonances periodically at the characteristic frequency related to the fault location. The above simulated signal is observed in the time interval [0, 0.1] s shown in Figure 3.9. The posterior probability sequence $\gamma_{1:N}(i)$ shown in Figure 3.9(b) represents the possibility of different states. By some manipulation of this sequence $\gamma_{1:N}(i)$, the pulse frequency and the information of the type fault in the signal can be obtained.

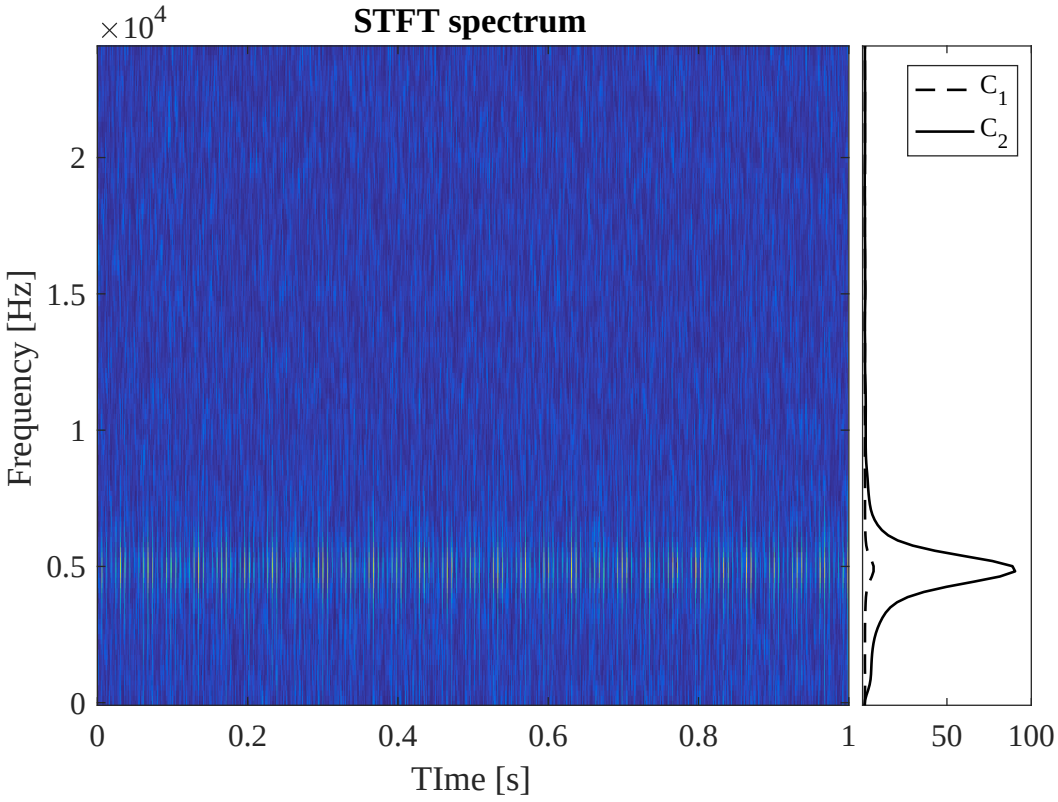

FIGURE 3.8: The spectrogram of the synthetic signal with resonance frequency $f_0 = 5kHz$ and fault frequency $f_{BPFI} = 161Hz$.

The parameter $\tau_n(i)$ is like a counter recording the number of Fourier window during state i. It describes well the temporal structure of the vibration signal. Figure 3.9(c) and (d) display the duration time sequence in state 2 and state 1, respectively. From the figure, it is found that these transient pulses last about 3 or 4 windows, and the inactive ones last about 18 windows. A variation in these duration values is probably due to the fact that when a roller goes to the non-load zone, the magnitude of the transient will be low and covered by noise. In practice, this may also be due to the difficulty of maintaining a constant speed during data collection. After obtaining the parameter τ_n, the single time period (interval between two impacts) of the simulated signal can be easily calculated through the summation of $\tau_n(1)$ and $\tau_n(2)$. The usage of it for identifying the fault type based on a statistical point of view will be illustrated in the next chapter. These two model parameters are proposed first time in vibration signal analysis.

Another interesting point is that the rotating frequency f_r, drawn in the red line of Figure 3.9(a), is removed in the posterior probability shown in Figure 3.9(b). This

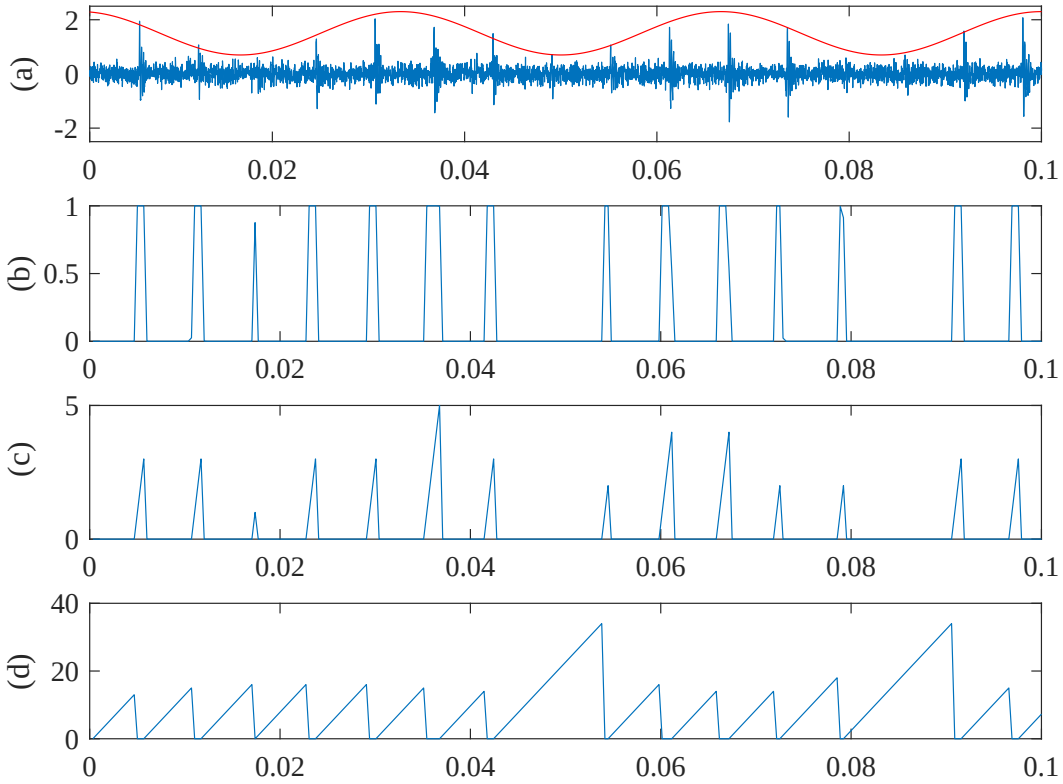

FIGURE 3.9: Zoom of simulated signal in interval [0 0.1] s. (a) Raw vibration signal; (b) posterior probability $\gamma_n(2)$; (c) duration time sequence $\tau_n(2)$ of the active state; (d) duration time sequence $\tau_n(1)$ of the inactive state.

is because the posterior probability is not sensitive to signal amplitude, so it will avoid the influence of signal modulation and lead to a clear spectrum. This advantage will be demonstrated in the validation chapter 5. In addition, it is found that there are two transient pulses wrongly estimated in Figure 3.9(b) as the roller was in the non-load zone so that the amplitude was small and covered by the noise. But the state sequence incompleteness will affect less the diagnosis result, identification and characterization, which will be discussed in Chapter 4.6.

Poission parameter $\boldsymbol{\lambda}$

λ_i is the parameter of a Poisson distribution that models the number of STFT windows covering the state i, d_i. It can be estimated as the expected value of the duration time in different states. The parameter λ_2 represents the average number of windows that cover a transient pulse. It will increase as the fault size increases, so it will be a good indicator to assess the fault size. This parameter is proposed first time to used in fault

size estimation. The fault size quantification based on λ_2 will be addressed in Chapter 4.5.

In order to illustrate the discrepancy of this parameter, 16 different sets of inner race fault data from the CWRU bearing data center are used here, including 4 different rotating speeds and 2 fault sizes. For statistical purposes, it is required to divide each dataset into several short signals through sliding segmentation with a overlap, to obtain more dataset. Then we can obtain a point (λ_1, λ_2) for each segment of the signal, as shown in Figure 3.10.

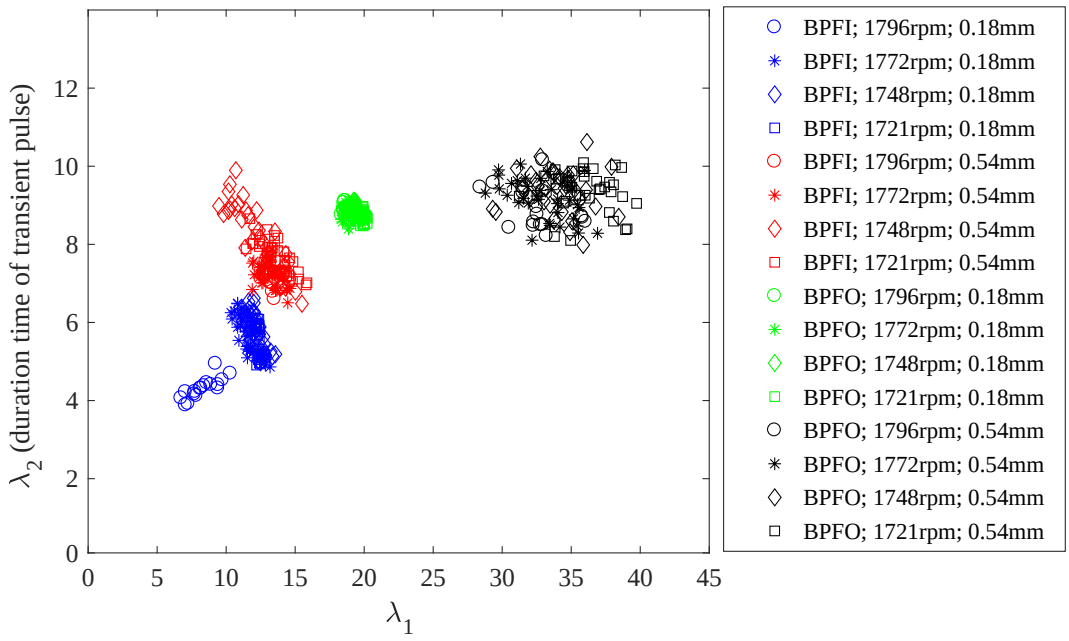

FIGURE 3.10: Illustration of the discrepancy of the parameters $\boldsymbol{\lambda}$ in different data sets.

From this figure, some remarks with respect to the parameter $\boldsymbol{\lambda}$ are worthy of attention. 1) Both λ_1 and λ_2 increase when the fault size increases, which is counter-intuitive. As the time period (intervals between the impacts) is fixed, λ_1 should theoretically decrease as the fault increases. This may be due to the transient pulses not being detected, and being identified wrongly as noise states. 2) It is found that, under the same fault size, λ_2 of the outer race is greater than the λ_2 of the inner race. It is also found that the rotating speed has an impact on duration time, but not much. From the figure, we can see that the features (λ_1, λ_2) are good at separating different samples (different fault type, different fault size). In this sense, they could be used as the features for classification.

3.6 Input parameter settings

In order to be fully automated, a methodology should rely on a minimum of hyperparameters. The hyperparameters introduced so far relate to the STFT and the EDHMM.

On the one hand, the STFT is characterized by the window type, its length N_w, and its shift R. It is known to be relatively insensitive to the window type, provided the window is a smooth function. Given a window type, the value of the shift should be small enough to prevent any loss of information. A typical choice is a Hann window with $R = N_w/4$. Hence, the window length N_w is the only critical parameter. It directly controls the frequency resolution $\Delta f = F_s/N_w$. As discussed in the reference [162], the window length should be shorter than the fault period (interval between two impacts). Hence, this condition can possibly be met given a prior knowledge of the potential faults and their associated characteristic frequencies. A lower bound is surely more difficult to find; it is data-dependent in general as it corresponds to the coarsest frequency resolution required to analyze the signal.

On the other hand, the EDHMM involves initial values for the unknown paramaters π_1, $\mathbb{C}_1$, and λ_i, $i = 1, 2$ (remember that π_2 and $\mathbb{C}_2$ are easily deduced from π_1 and $\mathbb{C}_1$ according to formula (3.25)). There are various possible initialization strategies. An ad hoc strategy is discussed in [162] for a two-state HMM of the STFT coefficients, which also applies here. Another popular strategy is to first cluster the data with the K-means algorithm and then get estimates from the obtained clusters. A last possibility is to resort to random initialization. Intensive numerical experiments have shown that the proposed EDHMM is robust to initialization, whatever the strategy used, because it involves only two states. It is therefore concluded that initialization of the EDHMM will not significantly affect the finial results. In conclusion, N_w remains the sole critical hyperparameter of the proposed methodology.

3.7 Conclusion

This chapter introduced and reviewed firstly the basic theory of Markov model, including Markov chain and hidden Markov model. It then mainly introduced the used stochastic model, explicit duration hidden Markov model, for modeling the vibration

signal, from vibration signal segmentation, stochastic modeling, parameter estimation to parameter analysis.

First, the STFT was formulated as a decomposition tool and it was explained how to apply it to the time-frequency transform of vibration signals, and how the observed time-frequency coefficients are modeled. Then, the iterative estimation algorithm was used to find the optimal parameter $\boldsymbol{\theta}$ to maximize the likelihood $p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})$. Afterward, some valuable parameters provided by EDHMM were introduced, which are useful for the vibration signal diagnosis, such as the posterior probability sequence $\gamma_{1:N}(i)$, covariance $\hat{\mathbb{C}}$, duration time sequence $\tau_{1:N}$, Poisson parameter $\boldsymbol{\lambda}$, etc. The effectiveness of these parameters and what characteristics they reveal were illustrated through a synthetic bearing fault signal. For example, the elements in the covariance $\hat{\mathbb{C}}$ measure the energy distribution in the frequency domain, which will be a good parameter for fault detection; and the probability sequence $\gamma_{1:N}(i)$ removes the influence of other high amplitude components like resonances and amplitude modulation; it is for some occasions able to increase the effect of spectrum analysis; the parameter $\boldsymbol{\lambda}$ in the Poisson distribution estimates the expected value of the duration time of different states, which is able to assess the fault severity. Finally, the hyperparameter of this model and the parameters initialization were discussed. It is found that only one hyperparameter, i.e. window length N_w, needs to be set in advance. This makes it potentially suited for practical applications in the absence of prior knowledge.

Chapter 4

Integrated auto-diagnostic framework

Contents

4.1 Introduction

In modern applications, the vibrations are often acquired in continuous flow on multiple channels and machines simultaneously, which brings the difficulty to process and analyze them. So it is not practical or possible anymore to have an expert dedicated to only looking at analysis results and condition indicators continuously. Such practical problems demand for the development of integrated and automatic diagnosis techniques. In this Chapter, to achieve this goal, an automated diagnosis framework is proposed, including detection, identification, fault size characterization and fault signal reconstruction. The estimated parameters of the EDHMM model are used to complete these tasks without need of other prior information and user's expertise.

This Chapter is organized as follows. Section 4.2 describes the statistical hypothetical test algorithm called likelihood ratio test for detecting the vibration signal. In Section 4.3, a time-varying filter is designed to recover the fault signal from corrupted signal. Subsequently, the identification of the fault type is achieved through one statistical counting method, and the posterior probability spectrum is also introduced. Section 4.5 introduces the characterization of the fault size using the duration time in different states. One discussion based on this integrated framework is given in Section 4.6.

4.2 Fault detection

Early detection is capable for examination of the fault occurrence and prediction of the fault evolution of the failing component before the fault progresses to a state that endangers the system's operational integrity. Timely and correct decision making is important for the real-time monitoring of crucial system. Therefore, the first step is to check whether the recorded vibration signal belongs to a healthy bearing or if it indicates the presence of a fault, in which case subsequent fault identification is needed.

As seen in the investigation of Chapter 2.3.1, the current techniques, either cannot accurately find the failure occurrence point as there is no corresponding alarm threshold, or take up a lot of resources for data training. In this section, one technique based on the likelihood ratio test (LRT) is introduced. According to Neyman-Pearson, the LRT is the most powerful test, which means it has the largest possible test power (probability of detection) given the significance level α (probability of false alarm). Unlike other

indicators, the likelihood ratio naturally follows a chi-squared distribution, which provides the mathematical basis for finding the alarm threshold. The detailed process is formulated as follows. First, the two alternative hypotheses, H_0 and H_1, are characterized by the following likelihood functions:

$$\begin{cases} \text{Null hypothesis } H_0 \text{ (healthy signal with state 1 only):} \\ p(\boldsymbol{Y}_{1:N}|H_0) = \prod_{n=1}^{N} \mathcal{CN}(\boldsymbol{Y}_n; 0, \mathbb{C}^{H_0}) \\ \text{Alternative hypothesis } H_1 \text{ (faulty signal with intermittent states 1 and 2):} \\ p(\boldsymbol{Y}_{1:N}|H_1) = \prod_{n=1}^{N} [\gamma_n(1) \cdot \mathcal{CN}(\boldsymbol{Y}_n; 0, \mathbb{C}_1^{H_1}) + \gamma_n(2) \cdot \mathcal{CN}(\boldsymbol{Y}_n; 0, \mathbb{C}_2^{H_1})]. \end{cases}$$

Since the theoretical covariance matrices are unknown, they have to be estimated under the two alternative hypotheses. On the one hand, under H_1, the covariance matrices $\mathbb{C}_i^{H_1}$, $i = 1, 2$, are estimated by Eq.(3.23) as explained in Chapter 3.5. On the other hand, under H_0, the estimate of $\mathbb{C}^{H_0}$ is

$$\hat{\mathbb{C}}^{H_0} = \frac{\sum_{n=1}^{N} \boldsymbol{Y}_n \boldsymbol{Y}_n^H}{N} = \hat{\mathbb{C}}_1^{H_1} + \hat{\mathbb{C}}_2^{H_1}, \tag{4.1}$$

where the last equality results from Eq.(3.24). As explained in Chapter 3.5.2, the covariance associated to one state is a good indicator for displaying the energy distribution in the frequency domain. Therefore, the covariances $\mathbb{C}_i^{H_1}$ and $\mathbb{C}^{H_0}$ will be similar for the null hypothesis, while very different under the alternative hypothesis. Based on this feature, the statistical indicator, generalized likelihood ratio (GLR), can be constructed by plugging in these estimates into the probabilities $p(\boldsymbol{Y}_{1:N}|H_i)$, $i = 0, 1$ as,

$$\begin{aligned} \Lambda &= \ln \frac{p(\boldsymbol{Y}_{1:N}|H_1)}{p(\boldsymbol{Y}_{1:N}|H_0)} \\ &= \ln \frac{\prod_{n=1}^{N} [\gamma_n(1) \cdot \mathcal{CN}(\boldsymbol{Y}_n; 0, \hat{\mathbb{C}}_1^{H_1}) + \gamma_n(2) \cdot \mathcal{CN}(\boldsymbol{Y}_n; 0, \hat{\mathbb{C}}_2^{H_1})]}{\prod_{n=1}^{N} \mathcal{CN}(\boldsymbol{Y}_n; 0, \hat{\mathbb{C}}^{H_0})}. \end{aligned} \tag{4.2}$$

The principle of the test is to accept the hypothesis H_1 if the GLR Λ is greater than a given alarm threshold. According to Wilk's theorem, twice the GLR asymptotically follows a chi-squared distribution with number of degrees of freedom, denoted v, equal to the difference between the number of unknown parameters under hypotheses H_1 and H_0,

that is

$$2\Lambda \sim \chi^2_v. \tag{4.3}$$

In the null hypothesis H_0, there are only N_k unknown elements in the covariance $\mathbb{C}^{H_0}$, whereas for hypothesis H_1, $2N_k$ for the two covariance matrices and N for the posterior probability γ_n (noting that $\gamma_n(1) + \gamma_n(2) = 1$). So the difference in degrees of freedom is $v = N_k + N$. Therefore, the null hypothesis H_0 is rejected if $\Lambda > \chi^2_{v,1-\alpha}/2$ at the risk α, where $\chi^2_{v,1-\alpha}$ is the quantile of the chi-squared distribution with probability $1-\alpha$. Figure 4.1 shows one example of Chi-squared distribution with the freedom degree $v = 5$ and alarm $\alpha = 0.05$. Therefore, any confidence interval can be obtained based on distribution with known parameters, which is able to provide a mathematical basis for fault detection.

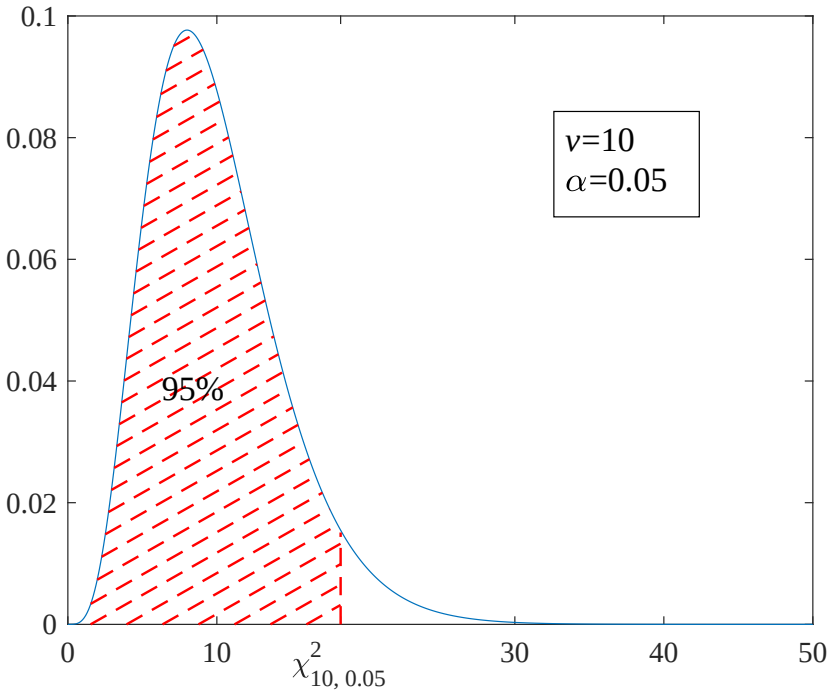

FIGURE 4.1: Illustration of fault detection based on chi-squared distribution with parameter $v = 10$ and $\alpha = 0.05$.

4.3 Fault signal reconstruction

Spectral amplitude estimation forms the basis of many signal restoration systems, such as for vibration signal denoising. For restoration of time-domain signals, an estimate of the instantaneous magnitude spectrum is combined with the phase of the noisy signal and then transformed via an inverse discrete Fourier transform to the time domain. The

related methods are covered in Chapter 2.3.2. Bayesian spectral amplitude estimation methods offer substantial performance improvements on spectral subtraction by utilising the probability density functions of the signal and noise process. The hidden states associated with signal spectra provide useful information for the evolution of signal spectra and the correlation between those spectra. Therefore, HMM can provide a mathematically tractable model basis for signal enhancement. In this section, the EDHMM based denoising method is formulated in the following.

The basic goal of feature enhancement is to extract the clean signal $x(t)$ from the measurement $y(t)$ with additive background noise $n(t)$. Firstly, according to the Bayesian theory and EDHMM theory, the posterior probability distribution of the nth observation $\boldsymbol{X}_n$ can be written as

$$p(\boldsymbol{X}_n|\boldsymbol{Y}_n,\boldsymbol{\theta}) \propto p(\boldsymbol{Y}_n|\boldsymbol{X}_n,\boldsymbol{\theta})p(\boldsymbol{X}_n|\boldsymbol{\theta}),$$

where, $\boldsymbol{X}_n$ represents the Fourier coefficients of the fault signal at nth window. For the specific frequency k, the fault signal spectrum posterior probability is

$$\begin{aligned}
&p(X(n,k)|Y(n,k),\boldsymbol{\theta}) \\
&\propto \frac{\exp\{\frac{-|Y(n,k)-\gamma_n X(n,k)|^2}{\hat{\mathbb{C}}_1(k)}\}\exp\{\frac{-|X(n,k)|^2}{\hat{\mathbb{C}}_2(k)-\hat{\mathbb{C}}_1(k)}\}}{\pi^2\hat{\mathbb{C}}_1(k)(\hat{\mathbb{C}}_2(k)-\hat{\mathbb{C}}_1(k))} \\
&\propto \frac{\exp\{-\frac{[\gamma_n^2(\hat{\mathbb{C}}_2(k)-\hat{\mathbb{C}}_1(k))+\hat{\mathbb{C}}_1(k)]X(n,k)^2-2\gamma_n(\hat{\mathbb{C}}_2(k)-\hat{\mathbb{C}}_1(k))Y(n,k)X(n,k)+(\hat{\mathbb{C}}_2(k)-\hat{\mathbb{C}}_1(k))X(n,k)^2}{\hat{\mathbb{C}}_1(k)(\hat{\mathbb{C}}_2(k)-\hat{\mathbb{C}}_1(k))}\}}{\pi^2\hat{\mathbb{C}}_1(k)(\hat{\mathbb{C}}_2(k)-\hat{\mathbb{C}}_1(k))},
\end{aligned} \tag{4.4}$$

where, $\hat{\mathbb{C}}_1$ and $\hat{\mathbb{C}}_2$ are the covariances of the observation in different states, and they can be estimated through Eq.(3.23). Then, the estimation of covariance of the signal of interest $\boldsymbol{X}_n$ is obtained as $\hat{\mathbb{C}}_2 - \hat{\mathbb{C}}_1$. From another way, the STFT coefficients of the fault signal $x(t)$ converge in distribution to the complex-valued Gaussian distribution according to the central limit theorem [33]. According to the certain assumptions, the STFT coefficients $X(n)$ at a given datum n follow the complex Gaussian distribution. For the specific frequency point k, the posterior probability distribution can be written as,

$$p(X(n,k)|Y(n,k),\boldsymbol{\theta}) = \frac{\exp\{-\frac{|X(n,k)-\mu_x(k)|^2}{\mathbb{C}_x(k)}\}}{\pi\mathbb{C}_x(k)}. \tag{4.5}$$

Combining the two formulas (4.4) and (4.5), the corresponding mean $\mu_x(k)$ and

covariance matrix $\mathbb{C}_x(k)$ of the posterior probability distribution can be given as

$$\begin{cases} \mathbb{C}_x(k) = \dfrac{\hat{\mathbb{C}}_1(k)(\hat{\mathbb{C}}_2(k) - \hat{\mathbb{C}}_1(k))}{\gamma_n^2(\hat{\mathbb{C}}_2(k) - \hat{\mathbb{C}}_1(k)) + \hat{\mathbb{C}}_1(k)} \\ \mu_x(k) = \dfrac{\mathbb{C}_x(k)}{\hat{\mathbb{C}}_1(k)}\gamma_n Y(n,k). \end{cases}$$

If one substitutes the covariance $\mathbb{C}_x(k)$ into the mean $\mu_x(k)$, the expectation of the fault signal spectrum at datum n can be obtained as,

$$E\{\boldsymbol{X}_n|\boldsymbol{Y}_n, \boldsymbol{\theta}\} = \boldsymbol{\mu}_x = \frac{\gamma_n(\hat{\mathbb{C}}_2 - \hat{\mathbb{C}}_1)}{\gamma_n^2(\hat{\mathbb{C}}_2 - \hat{\mathbb{C}}_1) + \hat{\mathbb{C}}_1}\boldsymbol{Y}_n. \tag{4.6}$$

Finally, the time-domain signal of interest, $x(t)$, is obtained from Eq.(4.6) by using the inverse STFT. Compared with the Wiener filter as Eq.(2.10), this can be seen as a time-varying filter, as which takes account of the posterior probability γ_n.

The technique of fault signal reconstruction is assessed based on its ability to recover the fault signal when the signal is disturbed by noise. Two different SNRs were chosen, 0 dB and -5 dB. For higher SNRs, the bearing fault signal impacts are more prominent and for lower SNRs the bearing fault impacts are easily masked by background noise.The noise immunity of the method will be discussion in the last section. Figure 4.2(a) depicts a simulated bearing inner race fault signal at rotating speed of 30 Hz, and the simulated fault characteristic frequency $f_{BPFI} = 161$ Hz. The simulated signal was further corrupted by Gaussian noise with SNR= 0 dB which is shown in Figure 4.2(b). Figure 4.2(c) shows that this technique has the capability of removing Gaussian noise with a 0 dB SNR, recovering a clean signal as close as possible to the simulated signal (bearing fault signal).

The same simulated signal is corrupted by the Gaussian noise with SNR=-5 dB, shown in Figure 4.3(b). Figure 4.3(c) shows that this technique has the capability of removing Gaussian with a -5 dB SNR and recovering a clean signal as close as possible to the source signal. The bottom subplot shows a consistent impulsive signal of a damaged bearing with an average time interval of 6.23 ms very close to the characteristic defect. The kurtosis of the corrupted signal was found to be 4.19, which is very close to a Gaussian distribution; while the kurtosis of the reconstructed fault signal was 20.73 and it can be seen that the recovered signal was improved.

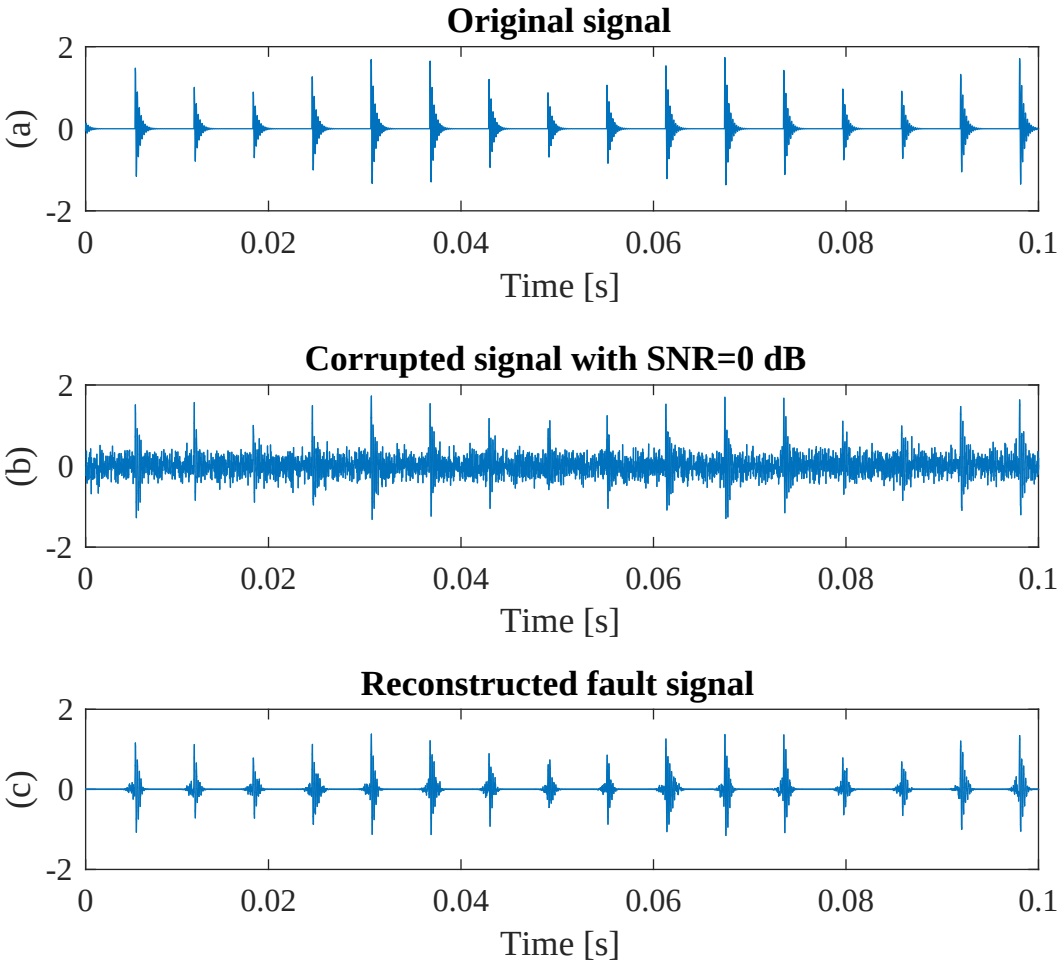

FIGURE 4.2: Illustration of the capability of fault signal reconstruction technique. (a) Simulated signal with 1800 RPM; (b) Corrupted signal with SNR= 0 dB Gaussian noise; (c) Reconstructed fault signal.

4.4 Fault type identification

Once a fault is detected, the type and severity of the fault will become the primary issue. Spectral analysis of vibration signal is widely used in bearing fault analysis. Most advanced techniques all rely on demodulation and spectrum analysis for finding the bearing fault type. However, apart from evaluation of vibration signal spectrum to identify specific bearing component frequencies, vibration signals can also be analysed by other means to identify the fault type. In this section, two different approaches are proposed based on the EDHMM parameters, i.e, posterior probability spectrum and statistical analysis-based identification.

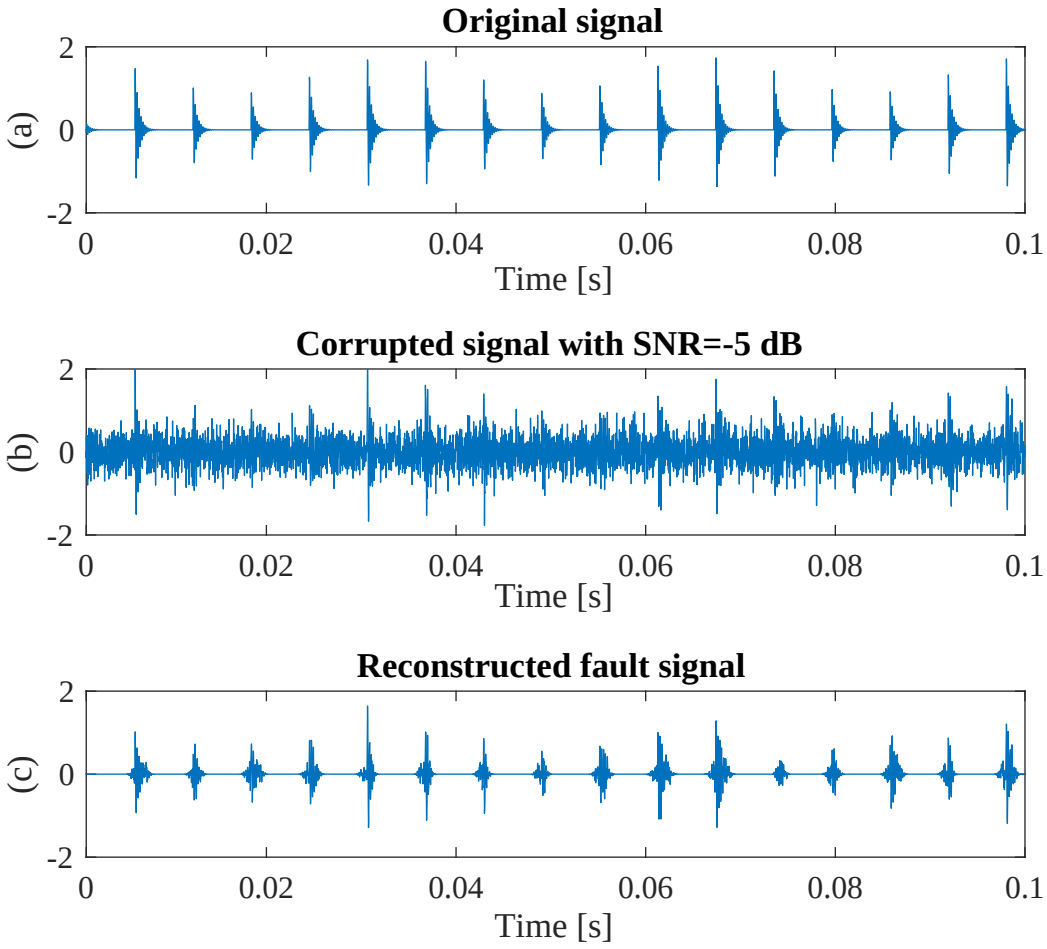

FIGURE 4.3: Illustration of the capability of fault signal reconstruction technique. (a) Simulated signal with 1800 RPM; (b) Corrupted signal with SNR= -5 dB Gaussian noise; (c) Reconstructed fault signal.

4.4.1 Posterior probability spectrum

The components in the vibration signal of a faulty bearing mainly consist of damped and repetitive impulses on the top of stationary background noise. This has been modeled in Chapter 3.4.2 as statistical properties switch between two states (the active and inactive states). The transition frequency between these two states relates to the bearing fault frequency, which is itself expressed in terms of the geometrical parameters of the bearing and is specific to each type of fault. Therefore, the idea is to identify the type of fault by performing the spectral analysis of the posterior probability of the active state, $\gamma_n(2) = p(z_n = 2|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta})$, with respect to time datum n. This defines the posterior

probability spectrum (PPS) $S(k)$ as

$$S(k) = |\mathcal{F}\{\gamma_n(2)\}|^2 = |\sum_{n=1}^{N} \gamma_n(2) e^{-j2\pi k \frac{n}{N}}|^2, \tag{4.7}$$

where $\mathcal{F}\{\cdot\}$ represents the discrete Fourier operator. The advantage of the PPS is that it is only related to the state information of the vibration signal, which is more conducive to spectrum analysis than the original signal. In addition, the PPS is not sensitive to the amplitude of the transient pulses, so it can avoid the influence of modulation, thereby reducing the sidebands in the spectrum. This advantage had been illustrated by an exemplary signal with inner race fault shown in Fig.4.4. It is shown in the subplot (a) that the characteristic frequency f_{BPFI} and its harmonic are very clear. Compared to the squared envelop spectrum in subplot (b), most sidebands or amplitude modulation have been removed in the PPS. This advantages will be further explained in Chapter 3.5.2. In respect of spectrum analysis, it is similar to the envelope spectrum, a cutting-edge tool widely used in vibration-based condition monitoring [124]. However, contrary to the latter, the PPS does not rely on manual pre-processing for the selection of informative bands in the signal, as typically achieved by means of finely tuned bandpass filters (see e.g. [6],[67],[109]). The advantage of the PPS will be demonstrated through validation in the next chapter.

The spectral analysis of the state sequence is a valuable tool for identifying the bearing fault type, yet it still requires a visual inspection. If this task is to be automated as well, one can still resume the spectrum by scalar indicators (see e.g. [29]) or use automatic spectral analysis methods as introduced in [58]. Another strategy is proposed hereafter in order to automatically identify the fault type among a given set of candidates.

4.4.2 Statistical analysis

Spectrum analysis is one common and effective method for identifying the bearing fault type. But the disadvantage is that it is not automatic. Finally, it still depends on the visual examination of the end-user. Although some methods can realize automatic scanning, they are difficult to implement. In order to completely escape the manual intervention, this section introduces one simple method based on a statistical point of view for identifying automatically the fault type.

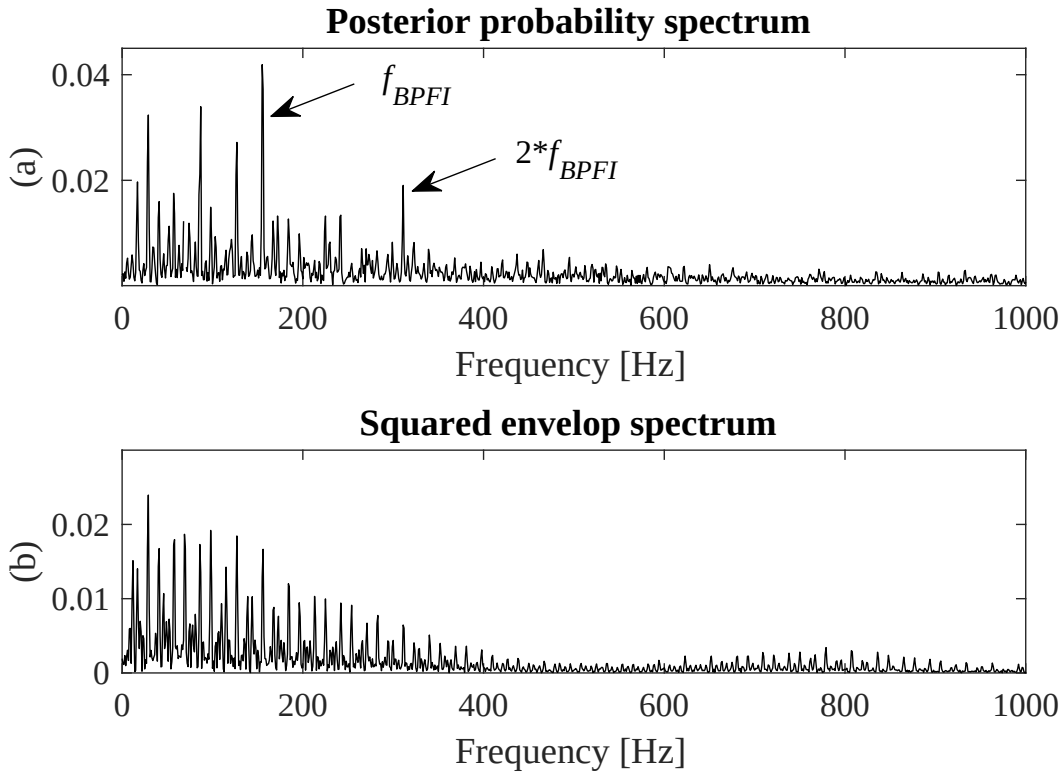

FIGURE 4.4: Illustration of the capability of PPS. (a) Posterior probability spectrum S(k); (b) Envelope spectrum of original vibration signal.

Since the EDHMM returns the time $\tau(i)$ already spent in the current state i at each instant n, the average period between two successive impulses can also be estimated, thus giving an indication of the characteristic fault frequency. More specifically, let t_s be the elapsed time between the s-th and the $(s+1)$-th impulses (see Figure 4.5) and let f_c be the expected characteristic frequency of a fault type c. Then, accounting for the possible presence of uncertainties in the actual value of f_c (unpredictable dependence on load, presence of slippage, etc.) and in the measurement of t_s (estimation errors, effect of noise), one can count the number of times t_s falls in a narrow frequency interval $[f_c - \sigma, f_c + \sigma]$ of width 2σ centered on f_c:

$$N_c = \text{card}(\{t_s : 1/t_s \in [f_c - \sigma, f_c + \sigma], s \in \{1, 2, ..., S\}\}), \tag{4.8}$$

where card(A) means the number of elements in the set A. If a number C of exclusive fault types are considered (e.g. inner-race fault, outer-race fault, ball-fault), the posterior

probability of observing fault c, $c = 1, ..., C$ given $t_{1:S}$ is then estimated by

$$p(f_c|t_{1:S}) = \frac{N_c}{\sum_{n=1}^{C} N_n}. \tag{4.9}$$

Since the equations to calculate the characteristic frequencies are based on bearing geometry and speed alone, variations due to loading and slipping are not considered. σ is the frequency error, which extends the characteristic frequency to a characteristic interval; it is typically about 5% of the fault characteristic frequency. Therefore, automatic identification of the type can be achieved by selecting the highest empirical posterior probability, i.e. finding the maximum probability is to identify the fault type. In addition, the samples $1/t_{1:S}$ can be also represented in the form of a histogram, and then the fault identification issue is equivalent to finding the highest bin.

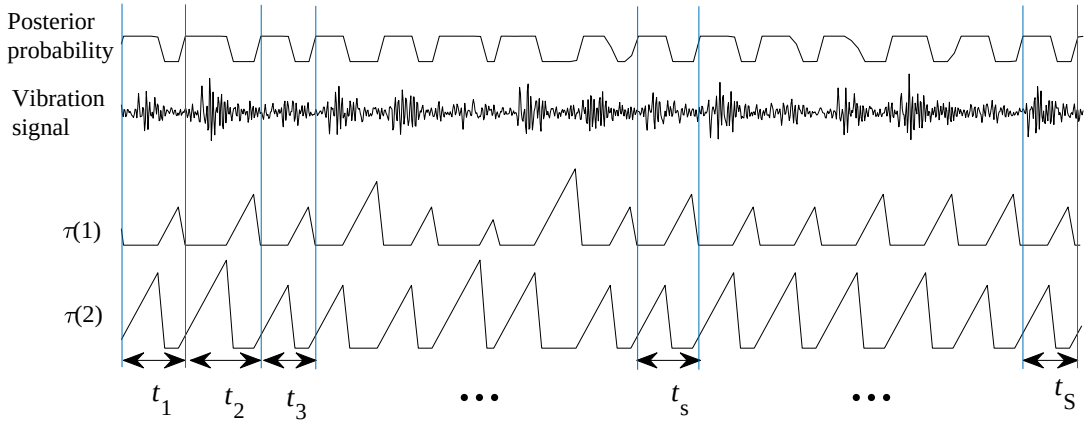

FIGURE 4.5: Illustration of the impulse cycles. The first sequence is the state posterior probability; second is the vibration signal; The last two subplots are the duration time sequence for different states.

4.5 Fault size characterization

The fault type can be determined through the techniques defined in the previous section, but the fault severity is still unknown. In the characterization stage, the size of the bearing fault is of concern. A proportional relationship between the transient pulse duration and fault size seems a reasonable assumption. In reference [129], Randall made a very thorough study of the vibration signatures due to the entry and exit of the rolling elements from the fault region and the duration time between them was used to estimate the fault size. Similarly, in EDHMM the duration time of the transient pulse state can

be used as an indicator for the fault size. The parameter λ_2 of the Poisson distribution (shown in the Eq.(3.25)) represents the average number of windows that cover the transient pulse. The corresponding average duration of the transient (in second) is

$$\Delta t \approx \frac{\lambda_2 R}{F_s}, \tag{4.10}$$

with F_s the sampling frequency of the signal. The value of Δt reflects the duration of the passage of the fault in the contact zone of the bearing. In practice, it is likely to be slightly over-estimated since it includes the coda, say Δe (in seconds), of the impulse response of the structure. However, if needed, the latter can possibly be measured by other means and subtracted from Δt. For instance, a reasonable estimate of the coda can be obtained by measuring the response of the structure after being impacted by a hammer. Another simple method – used hereafter – is to deduce the value of Δe from measurement taken at different rotation speeds. Besides, the precision of Δt is on the order of the window shift, i.e. R/F_s . This is surely less accurate than the precision that could be achieved with other methods such as blind deconvolution [130, 123, 34, 35], yet the latter usually comes with a high demand on the user's expertise. Sacrificing a degree of precision is the price to pay for having an automated method.

For a faulty bearing with stationary outer race, the fault size l (in meter) is related to the transient duration Δt as follows,

Outer race fault:

$$l \approx \pi f_{FTF}(\Delta t - \Delta e) \cdot D \tag{4.11}$$

Inner race fault:

$$l \approx \pi f_{FTF}(\Delta t - \Delta e) \cdot D(1 + d/D) \tag{4.12}$$

Rolling element fault:

$$\begin{aligned} l &\approx \pi f_{FTF}(\Delta t - \Delta e) \cdot D(1 + d/(2D)) \\ &\approx \pi f_{BSF}(\Delta t - \Delta e) \cdot d \end{aligned} \tag{4.13}$$

where D and d stand for the pitch and element diameters of the bearing, respectively, and f_{FTF} for the fundamental train frequency. It is worth noting that a fault on a rolling

element actually produces impacts of slightly different durations on the outer and inner races. Formula (4.11) applies to the former case and formula (4.12) to the latter case. Formula (4.13) is an average, when both races are impacted. Noting that when several rotation speeds are available, say f_r^j, $j = 1, 2, ...$, the above formulas can be cast in the general form $\Delta t^j = l \cdot c(f_r^j) + \Delta e$, where $c(f_r^j)$ is a function of f_r^j. Therefore, the fault size l and the coda error Δe can be estimated as the slope and the intercept, respectively, of the fitted lines between the duration times Δt^j and $c(f_r^j)$, as shown in Figure 4.6. One advantage of this technique is that regression will decrease the error produced in the estimation and transformation of Δt, and this advantage will become more obvious as the fitted samples increase. Although this technique still cannot accurately quantify the size, it will be useful for condition monitoring as demonstrated hereafter in the next chapter.

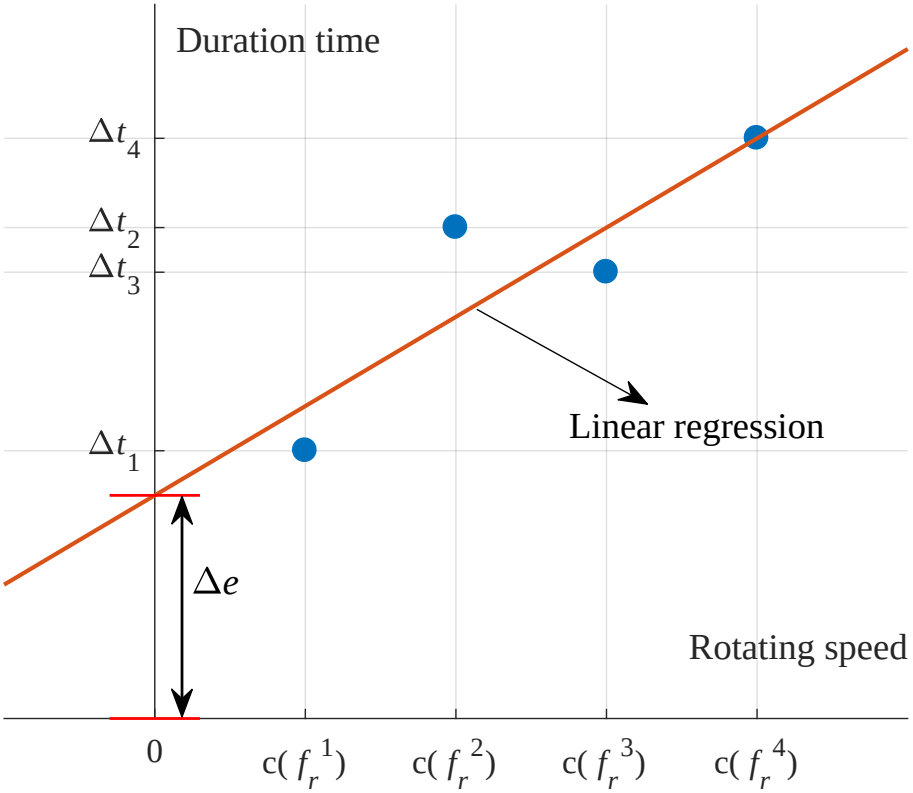

FIGURE 4.6: The illustration of fault size characterization through linear regression, with the horizontal axis as the function $c(f_r^j)$ and vertical axis as the duration time Δt^j, in which the estimated fault size $\hat{l}$ and the coda error Δe correspond to the slope and the intercept of the fitting line.

4.6 Discussion

Figure 4.7 summarizes the integrated diagnostic algorithm and resumes its essential steps introduced in this chapter. The algorithm takes the recorded signal as an input and returns elements of information on 1) detection, 2) reconstruction, 3) identification and 4) size characterization of a potential fault. The detection step returns a binary output, which can be possibly substantiated by the likelihood ratio Λ (or a $p-$value). The higher the value of Λ with respect to a threshold, the lower the risk of false alarm (see Eq.(4.3)). If a fault is detected, the identification step returns the PPS, the equivalent of an envelope spectrum, which can be analyzed visually or automatically. In order to be fully automatic, a simple method that calculates the posterior probability of different fault types for identifying the fault type is also introduced. Eventually, the fault signal can be extracted through a time-varying filter constructed from the covariance matrix of the observations. Independently of the latter, the fault characterization step returns an estimate of the fault size.

It is emphasized again that the complete algorithm relies only on one critical hyperparameter, N_w, the inverse of the frequency resolution. In particular, it does neither rely on historical data, nor on finely tuned pre-filters, nor on trial-and-error manipulation of time-frequency distributions. In addition, these four tasks are independent of each other, in the sense that any one of the tasks can be accomplished effectively without the aid of other. Therefore, it can be understood as a modular integrated framework. However, one high risk in this diagnosis framework is the strong dependency on the stochastic model. Only if the model is well established, the integrated diagnosis will perform well.

In practical applications, the vibration signal is always distorted by all kinds of noise. For example, when a roller goes to the non-load zone, the magnitude of the transient pulse will be low and easily covered by noise, which may be estimated wrongly as noise state. As shown in Figure 3.9(b), it is seen that two transient pulses with low energy that have been estimated wrongly. But the wrong estimation of a state or the state sequence incompleteness affect little the final result. The reason is because the statistical analysis is based on statistics comparison, it is rather unaffected by the hidden state incompleteness. The example in Figure 3.10 is used here to explain. In this synthetic signal, there are some transient pulses that are totally covered by noise. However, for the statistical analysis based identification, the probability of inner race fault is the highest in the posterior

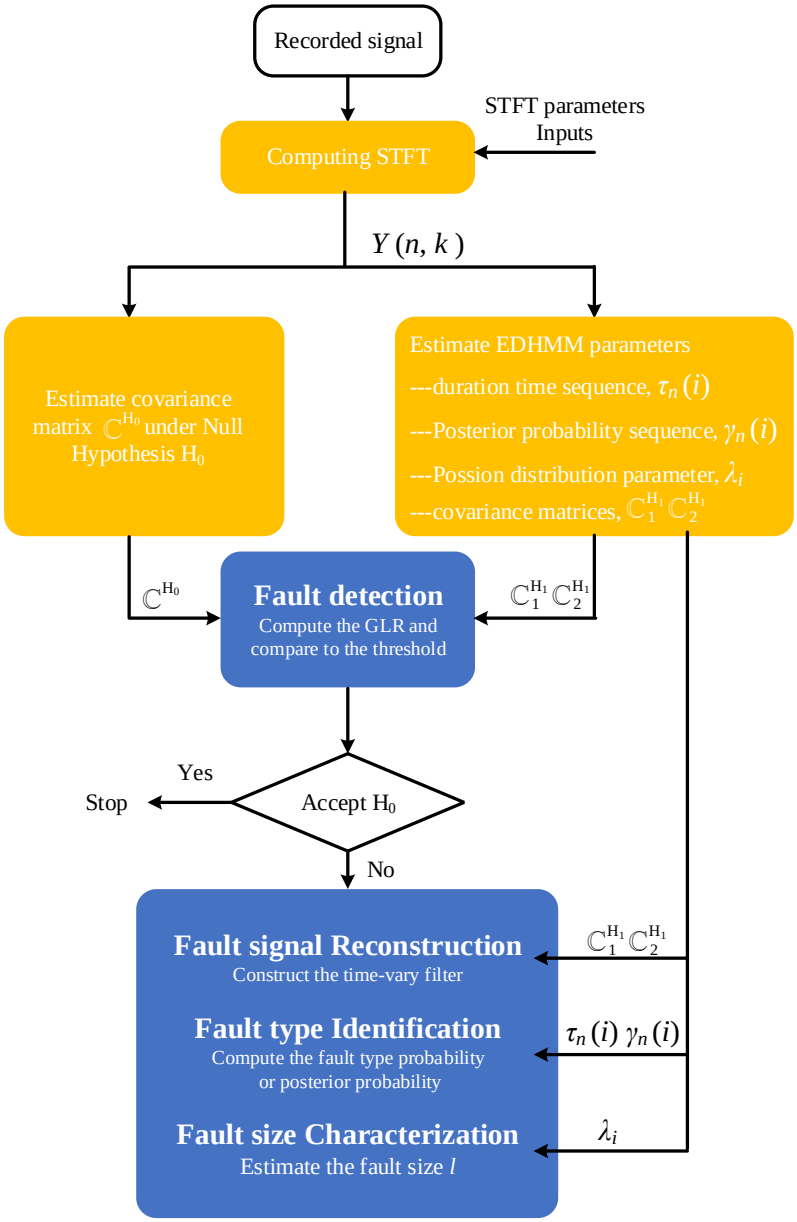

FIGURE 4.7: The flowchart of the integrated diagnosis framework.

probability Table 4.1. Meantime, Figure 4.8 shows the histogram of the frequency data of the incipient fault signal, $f = 1/t_s, s \in \{1, 2, ..., S\}$. Histogram is a probability density representation, which will in general more intuitively reflect the distribution of the fault impulsive cycle. It shows that the majority periods fall into the frequency bin [157 167] Hz, which includes the characteristic frequency f_{BPFI}. The second highest bin is [76 86] Hz, which is due to the doubling period because of the wrong estimation of the state. With respect to the characterization, the missing of some transient pulses will also not affect the estimation of λ_2. Because the estimated $\hat{\lambda}_2$ (the expected number of Fourier

windows over transient state) is only related to $\tau_n(2)$ according to Eq. 3.26. Therefore, hidden state incompleteness will not affect the estimation of λ_2. But overcompleteness (the noise state is estimated wrongly as transient state) will affect the estimation. Such influence will be decreased after the linear regression. In conclusion, wrong estimation of a state or the state sequence incompleteness affect little the finial result.

TABLE 4.1: The posterior probabilities of fault type given $t_{1:S}$

Fault type	BPFI	BPFO	BSF	FTF
Posterior probability	$p(f_{BPFI}\|t_{1:S})$	$p(f_{BPFO}\|t_{1:S})$	$p(f_{BSF}\|t_{1:S})$	$p(f_{FTF}\|t_{1:S})$
	1	0	0	0

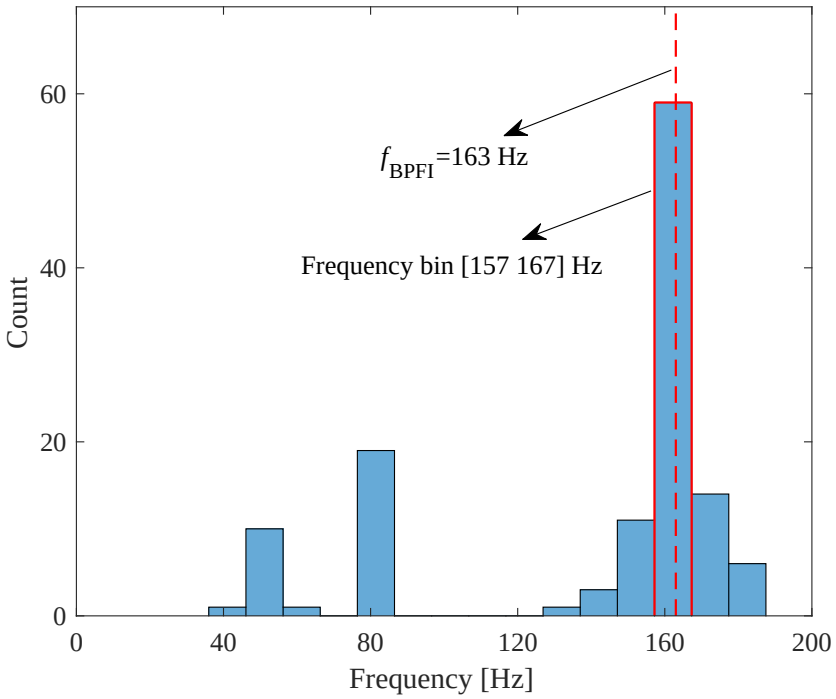

FIGURE 4.8: Histogram of the transient pulse cycle frequency, showing the red dash line relating to the fault characteristic frequency located in the highest bin.

Two aspects are important regarding the effectiveness of the proposed method against different noises, i.e. the resonance frequency band and the noise energy (amplitude). If the noise and the signal of interest locate in different frequency bands, e.g. the noise from environment or component outside the system. No mater how strong of the noise, it won't affect the result. The same synthetic signal $x(t)$ in 3.5.2 are used here, as shown in the yellow signal in Figure 4.9 (a). The resonance frequency is 5000 Hz. And the noise $n(t)$ is filtered by a high pass filter with the cutoff frequency f_c as 8000 Hz as the blue signal in Figure 4.9 (a). It is found the amplitude of the noise is higher than the signal of interest

$x(t)$. The composite signal is shown in Figure 4.9 (b), in which, the fault signal $x(t)$ is totally covered by the noise. However, in Figure 4.10, the PPS of the composite signal displays the obvious information of signal $x(t)$. This example proves the proposed method is immune to noise that locates in different resonance frequency band. In another case, when they have the same resonance frequency, the effectiveness of the method depends on the magnitude of the noise. If the signal of interest $x(t)$ is totally covered by the noise in the same resonance frequency band, for example, the strong noise from the same mechanical system, the proposed method will fail.

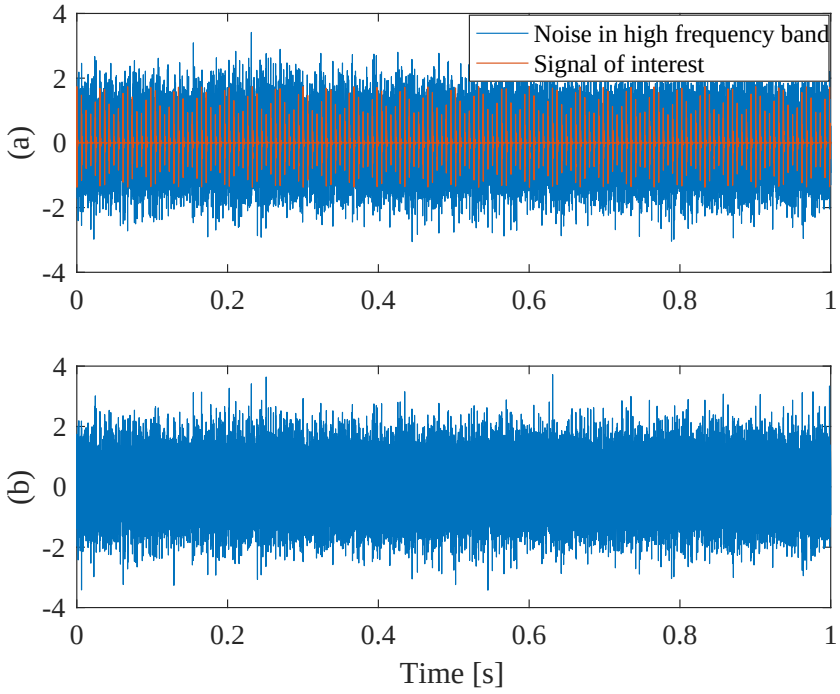

FIGURE 4.9: Illustration of the synthetic signal. (a) The noise $n(t)$ and the signal of interest $x(t)$; (b) the composite signal.

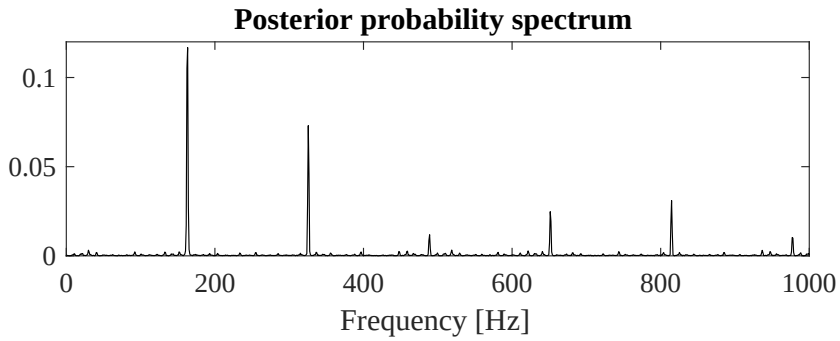

FIGURE 4.10: The posterior probability spectrum of the composite signal.

4.7 Conclusion

This chapter has presented the diagnosis framework based on the parameters from the EDHMM. All the diagnosis issues can be addressed at once in an automated way, i.e. without manual intervention in the process. The contribution of this part is the realization of integrated auto-diagnosis.

Chapter 5

Experimental validation

Contents

5.1 Introduction

So far, the stochastic model, the diagnostic framework and its automatization have been discussed in the previous chapters only at the theoretical level. Brief illustrations of the model parameters on simulated data have been given, highlighting their performance. In this chapter, the proposed integrated auto-diagnostic framework is tested on experimental data. The experimental validations of this chapter give a better insight into the proposed method in a practical setting. For this validation, two different kinds of experimental data associated to bearing signal are used. The first experimental data is relative to bearings with different damages, running at different speeds and under different loads; the second kind of experimental data reports the behaviour of a single bearing undergoing a long test at constant speed and load until fault occurrence. Based on the validation results, the final section gives a general discussion of the achieved diagnostic performances of the proposed integrated framework.

5.2 Variable condition data

In order to illustrate the benefit of the integrated auto-diagnostic framework, experimental data from the Case Western Reserve University Bearing Data Center [37] are examined. In Cases 1 and 2, 24 sets of data associated to the drive end with three bearing fault sizes are selected: 12 sets for the inner race fault and 12 sets for the outer race fault. There are four different motor speeds and motor loads in each fault size. In Case 3, the rolling element fault data are from Technology Research Group of the Polytechnic University of Madrid [140]. The dataset comprises vibration signals from a double-row spherical roller bearing (FAG22205E1KC3) with four different fault depths and three speeds of rotation.

5.2.1 Case 1: Inner race fault data

In this case, 12 sets of inner race fault data acquired from drive end are used, and the sampling frequency is 48 kHz. Table 5.1 gives a summary of the data number and the corresponding information. According to the frequency sampling and rotation frequency, the window length and window shift are set to $N_w = 128$ samples and $R = 16$, which gives a frequency resolution of 375 Hz. Therefore, there are $N = \frac{48000-N_w}{R} + 1 = 2993$ windows

covering one second signal. The number of Fourier transform (NFFT) is set as 256, then $N_k = NFFT/2 = 128$. The quantile in Eq.(4.3) can be calculated as $\chi^2_{v,1-\alpha} = 3253.1$, with risk $\alpha = 0.05$ and $v = N + N_k = 3121$ degrees of freedom. The bearing details and its characteristic frequencies are listed in Table 5.2.

TABLE 5.1: List of the inner race fault datasets and their corresponding basic information

Dataset	Accelerometer location	Fault Size [mm]	Motor load [hp]	Motor Speed [rpm]
#097	Drive end	0	0	1796
#098			1	1772
#099			2	1750
#100			3	1730
#109		0.18	0	1796
#110			1	1772
#111			2	1750
#112			3	1730
#174		0.36	0	1796
#175			1	1772
#176			2	1750
#177			3	1730
#213		0.54	0	1796
#214			1	1772
#215			2	1750
#217			3	1730

TABLE 5.2: The basic information bearing details and fault frequencies

Bearing type	SKF 6205-2RS JEM	
Position	Drive end	
Characteristic frequencies (×rotation frequency f_r)	BPFO	3.585
	BPFI	5.415
	BSF	2.357
	FTF	0.3983

Table 5.3 summarize the inner race fault diagnosis results, including fault detection, fault identification, and fault size quantification. In order to demonstrate the effectiveness of the detection algorithm, four healthy data records (#097 – #100) are used as a basis of comparison. In the "Detection" column, it is found that the test statistic Λ of the healthy data are all less than the quantile $\chi^2_{v,1-\alpha}/2$, whereas the test statistic Λ of the faulty data are greater than this threshold except for #174. The "Identification" column lists the posterior probabilities of four fault types given the impulse cycle $t_{1:S}$, where,

$p_1 = p(f_{BPFI}|t_{1:S})$, $p_2 = p(f_{BPFO}|t_{1:S})$, $p_3 = p(f_{BSF}|t_{1:S})$, $p_4 = p(f_{FTF}|t_{1:S})$. From the results, it is seen that the inner race fault probability, p_1, is maximum for almost all the records, proving the effectiveness of the statistical analysis for automated identification method. In the "Characterization" column, the fault size quantification results are given according to the algorithm of Chapter 3.5. Note that the results for the record #174 are meaningless because this dataset is likely to have been corrupted due to some unknown reasons, e.g. mechanical looseness as suggested in the reference [137].

TABLE 5.3: Inner race fault diagnosis results

Dataset	Detection		Identification				Characterization	
	Λ	Result	p_1	p_2	p_3	p_4	λ_2	l [mm]
#097	503.4	Accept H_0	——				——	——
#098	491.7	Accept H_0	——				——	——
#099	727.1	Accept H_0	——				——	——
#100	529.6	Accept H_0	——				——	——
#109	83606	Reject H_0	1	0	0	0	5.23	0.14
#110	107329	Reject H_0	1	0	0	0	6.11	0.19
#111	85730	Reject H_0	0.98	0.02	0	0	6.18	0.20
#112	98226	Reject H_0	0.96	0.04	0	0	5.99	0.15
#174	NAN	NAN	——				——	——
#175	84173	Reject H_0	0.70	0.30	0	0	6.26	0.27
#176	85226	Reject H_0	0.57	0.09	0.34	0	6.55	0.37
#177	73326	Reject H_0	0.55	0.06	0.39	0	6.59	0.39
#213	71309	Reject H_0	1	0	0	0	7.29	0.68
#214	49428	Reject H_0	0.97	0.03	0	0	7.10	0.57
#215	51570	Reject H_0	1	0	0	0	7.28	0.66
#217	26051	Reject H_0	0.93	0	0.07	0	7.11	0.56

The assessment of the fault size is illustrated in Figure 5.1. As explained in the previous chapter, the parameter λ_2 returns the average duration of the transient pulse (in window length) augmented by the coda of the impulse response of the structure. As explained in Chapter 4.5, the latter can be measured by the vertical-intercept of the fitted lines between the duration time Δt and the rotation frequency f_r. The results of coda Δe are reported in Table 5.4. The corrected duration time of the fault is eventually obtained by subtracting the coda error Δe, i.e. $\Delta t - \Delta e$, with Δt calculated from Eq.(4.9). Figure 5.1 shows that the estimate of the fault size varies slightly with motor speed, but is close to the actual size within an acceptable error. In this figure, the estimated fault sizes are in the 90% confidence intervals of the actual values obtained by assuming a uniform error with standard deviation equal to R/F_s.

TABLE 5.4: The coda error Δe [ms]

Fault size (mm)	0.18	0.36	0.54
Δe (ms)	1.92	1.96	2.03

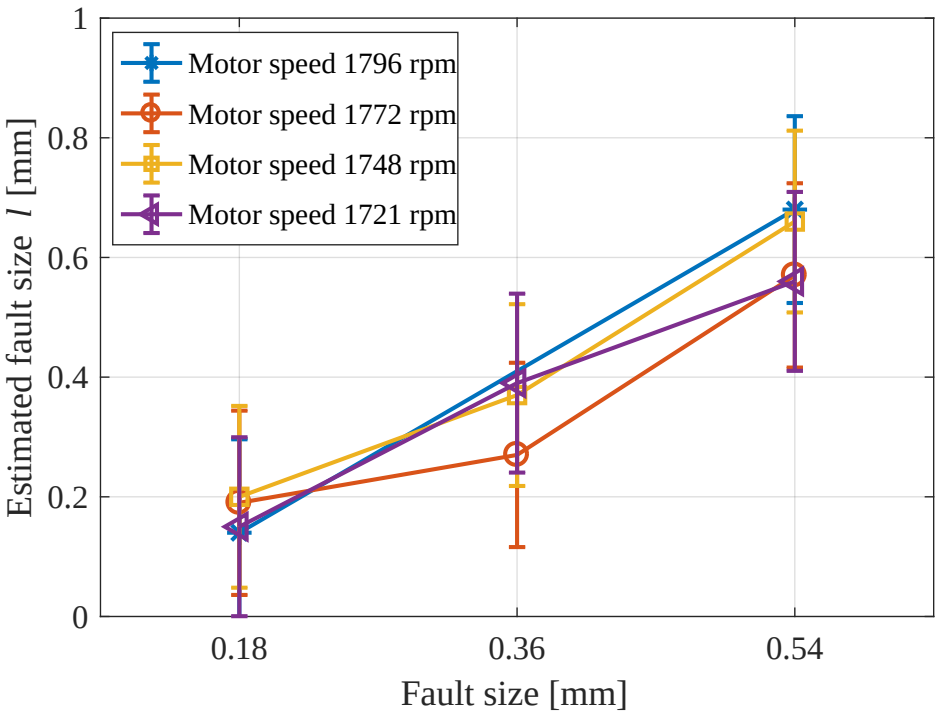

FIGURE 5.1: Quantification of the fault size together with 90% confidence intervals based on the assumption of a uniform error with standard deviation equal to R/F_s.

The capability of the proposed integrated framework to perform automated diagnosis is now validated on record #175. The spectrogram of the raw signal is displayed in Figure 5.2. $\mathbb{C}_1$ and $\mathbb{C}_2$ are the covariances of two different states, representing the energy distribution. It is seen that there exist nonstationary components in the frequency band [0,5000] Hz. The posterior probabilities are summarized in Table 5.3. It is seen that the probability $p(f_{BPFI}|t_{1:S})$ is maximum, thus indicating an inner-race fault. The identification is performed totally automatically, without the need of the decision of the end-user. In this record, it is found that the outer fault also accounts for a non-negligible proportion. This is likely due to the modulations by the shaft and the cage rotations, resulting in some low-frequency components incidentally falling into the outer fault characteristic interval.

In order to assess its performance for bearing fault identification, the PPS $S(k)$

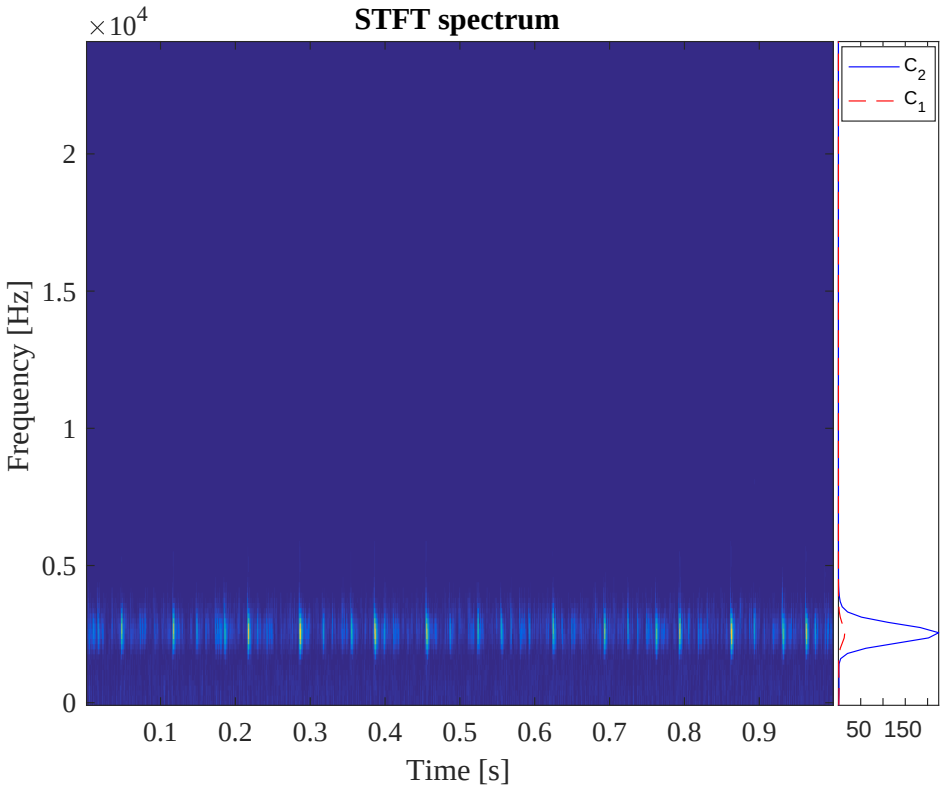

FIGURE 5.2: Spectrogram of the record #175 and the corresponding covariance of two states, $\mathbb{C}_1$ and $\mathbb{C}_2$, in the right subplot.

is illustrated in Figure 5.3(a). It is seen that the inner fault frequency f_{BPFI} and its harmonics are clearly dominant in the spectrum, thus indicating an inner-race fault. This is compared in Figure 5.3(b) with the state-of-the-art approach, which consists in computing the envelope spectrum after bandpass filtering the signal in the most informative frequency band, here selected as the maximum of the kurtogram [9, 137]. The standard envelope spectrum also reveals the BPFI, but accompanied with several sidebands due to modulations by the shaft and the cage rotations. These truly pertain to the inner-race fault signature. The reason why they are not present in the PPS is because the latter is, by construction, less sensitive to amplitude modulation. Incidentally, the PPS also shows strong sidebands at three time the speed rotation (present too in the envelope spectrum), which are likely due to frequency modulations.

In order to better assess the performance of this framework, the results are observed in the time interval [0.5, 0.6] s shown in Figure 5.4. The expected value λ_2 of the duration time in the transient pulse state is 6.26 (window lengths), which means that the transient pulse lasts about 2.09 ms. This includes the coda of the impulse response that lasts about

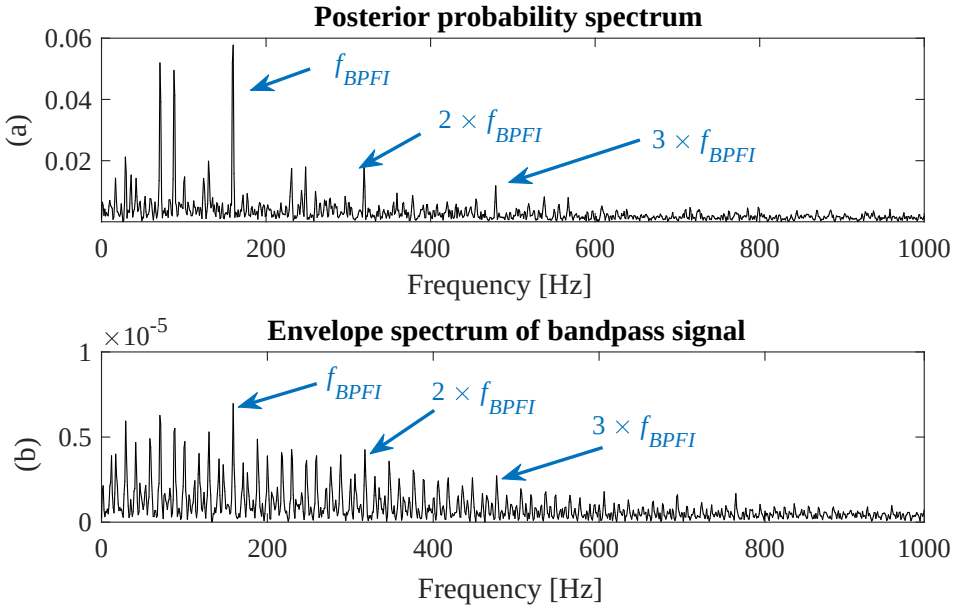

FIGURE 5.3: Spectral analysis of record #175. (a) Posterior probability spectrum $S(k)$; (b) Envelope spectrum of bandpass signal with maximum kurtosis in band [5; 6] kHz selected from kurtogram.

1.96 ms. Therefore, after correction, the fault size estimate is $l \approx 0.27$ mm according to Eq.(4.10). The estimated value deviates from the actual size 0.36 mm, which might be due to the precision of duration time Δt. Although it is hard to quantify exactly the fault size, it can still provide a meaningful guide for bearing diagnosis and prognosis.

Finally, the capability of the proposed method to reconstruct the fault signal is now demonstrated through record #175. Compared to Figure 5.5(a), it can be seen from Figure 5.5(b) that many low-energy transient pulses have been recovered after signal reconstruction. It can be seen that the background noise has been eliminated, and the original signal has been enhanced. The impulses from the defect on the inner race, especially the impulses produced in the non-load zone, can be clearly seen and the SNR has been improved. The reconstructed signal shows a typical sequence of impulsive signal, and few frequency components such as 29 Hz, 160 Hz were detected.

In order to illustrate the sensitivity of the proposed method with respect to the window length, the effect of taking different values of N_w is investigated when estimating the empirical posterior probability p_1 – see Figure 5.6. It is seen that only the case with $N_w = 128$ has probability greater than 0.5, yet the probability remains significant for at least three octaves of N_w. This demonstrates that the temporal structure of the signal can

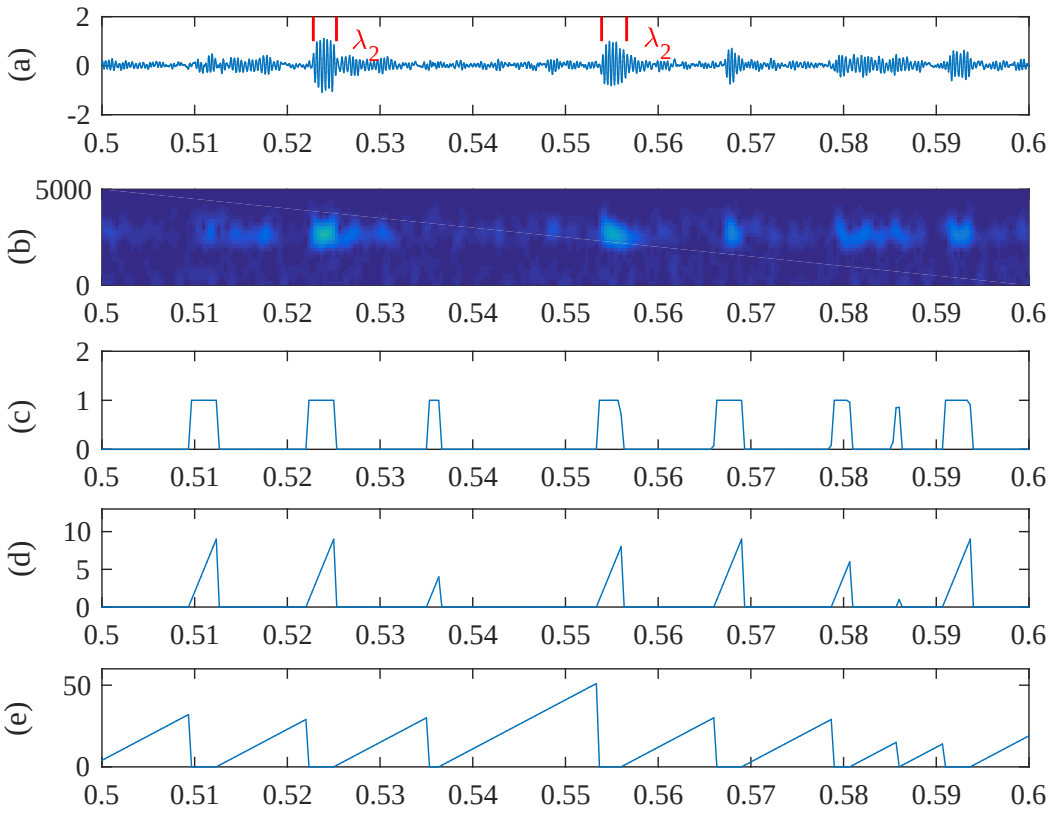

FIGURE 5.4: Zoom of #175 in interval [0.5, 0.6] s. (a) Raw vibration signal; (b) the corresponding spectrogram; (c) posterior probability $\gamma_n(2)$; (d) duration time sequence $\tau_n(2)$ of the active state; (e) duration time sequence $\tau_n(1)$ of the inactive state.

be revealed in a reasonably large range of values of the hyperparameter N_w. In addition, it is found that the window shift R almost has no effect on the result.

5.2.2 Case 2: Outer race fault data

In this case, the outer race fault datasets from the Case Western Reserve University are used to validate the integrated diagnostic method. Table 5.5 lists the basic information of the used 12 data sets. The sampling frequency is 48 kHz.

The same STFT parameters are used, so the quantile in the fault detection is also same with the first case, i.e. $\chi^2_{v,1-\alpha} = 3253.1$. Table 5.6 displays the diagnosis results in the case of an outer race fault. It shows that faults of sizes 0.18 mm ($\#135 - \#138$) and 0.54 mm ($\#238 - \#241$) are detected correctly. They are also correctly identified as outer-race faults with the maximum probability $p(f_{BPFO}|t_{1:S})$. In addition, the last column of the table shows that the estimated fault sizes are reasonable. However, the

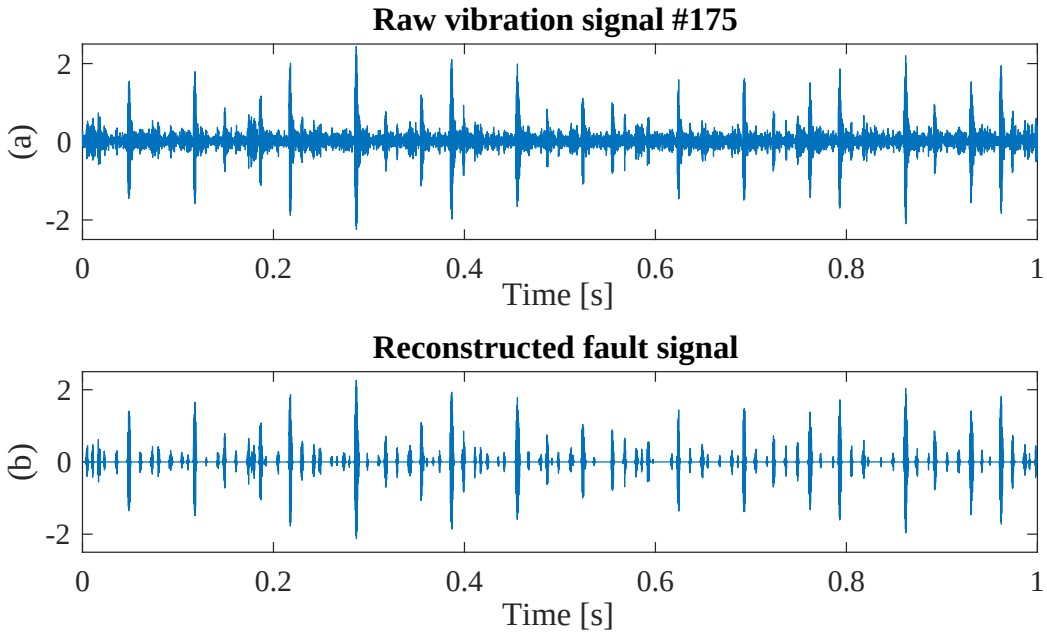

FIGURE 5.5: Fault signal reconstruction. (a) Raw vibration signal #175; (b) Reconstructed fault signal.

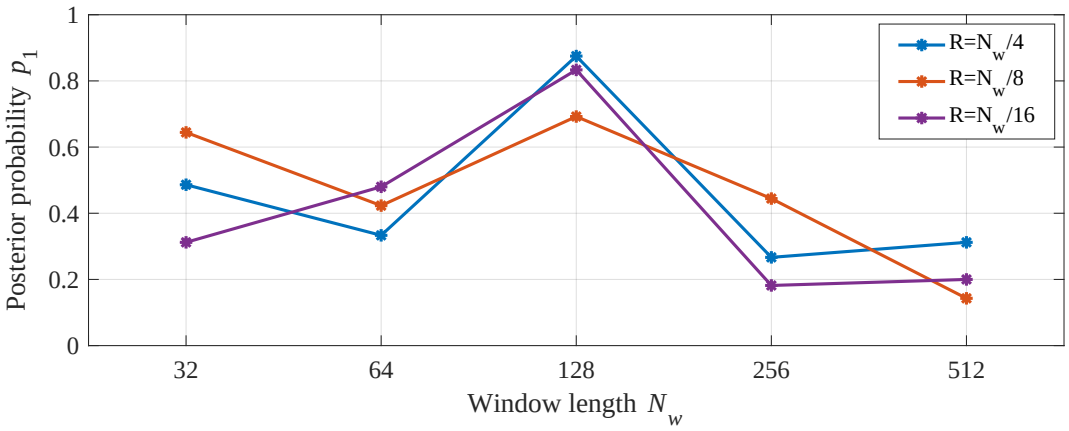

FIGURE 5.6: Empirical posterior probability p_1 for different values of the hyperparameter N_w.

records corresponding to the fault size 0.36 mm ($\#201 - \#204$) have poor diagnosis results. For record #201, the presence of a fault is correctly detected, but the probability of outer race fault is not the maximum. Records $\#202 - \#204$ are detected wrongly as healthy signal. It is noted here that these signals were already recognized as difficult to diagnose in the reference [137]. A visual inspection of these signals actually confirms that they strongly resemble healthy signals. They have been analyzed here to fairly assess the limits of the proposed method. It is emphasized that the latter has not been designed to improve

TABLE 5.5: List of the outer race fault data datasets and their corresponding basic information

Dataset	Accelerometer location	Fault Size [mm]	Motor load [hp]	Motor Speed [rpm]
#135	Drive end	0.18	0	1796
#136			1	1772
#137			2	1750
#138			3	1730
#201		0.36	0	1796
#202			1	1772
#203			2	1750
#204			3	1730
#238		0.54	0	1796
#239			1	1772
#240			2	1750
#241			3	1730

the detection capability of state-of-the-art methods (e.g. as given in [137]), but rather to cover the four consecutive tasks of diagnosis in an automated way. The fact that it does so nearly as well as finely hand-tuned methods is remarkable.

TABLE 5.6: Outer racer fault diagnosis results

Dataset	Detection		Identification				Characterization	
	Λ	Result	p_1	p_2	p_3	p_4	λ_2	l [mm]
#135	321543	Reject H_0	0	1	0	0	8.65	0.15
#136	284131	Reject H_0	0	1	0	0	9.24	0.21
#137	268814	Reject H_0	0	1	0	0	8.99	0.20
#138	265902	Reject H_0	0	1	0	0	9.40	0.23
#201	2797.8	Reject H_0	0.76	0.24	0	0	——	——
#202	385.2	Accept H_0	——				——	——
#203	275.2	Accept H_0	——				——	——
#204	153.8	Accept H_0	——				——	——
#238	210946	Reject H_0	0.03	0.97	0	0	9.26	0.45
#239	221400	Reject H_0	0	1	0	0	10.10	0.62
#240	200886	Reject H_0	0	1	0	0	9.85	0.61
#241	206363	Reject H_0	0	0.82	0.18	0	10.37	0.68

5.2.3 Case 3: Rolling element fault data

In this subsection, the proposed automated diagnosis framework is illustrated on the dataset from Technology Research Group of the Polytechnic University of Madrid [140]. The dataset comprises vibration signals from a double-row spherical roller bearing (FAG

22205E1KC3), with different fault depths, recorded by accelerometers in three directions. There are four different fault depths and three speeds of rotation. The sampling frequency is 40 kHz. Each combination was recorded three times during 30 s to ensure that small differences due to uncontrollable variables were distributed evenly across all records. Table 5.7 reports the rolling element fault data sets and their corresponding basic information, and Table 5.8 lists the potential fault frequencies. More details can be found in reference [139]. This case aims at analyzing all signals recorded by the accelerator at 6 o'clock on the bearing casing.

TABLE 5.7: List of the rolling element fault datasets and their corresponding basic information

Dataset	Accelerometer location	Fault Size [mm]	Motor Speed [rpm]
#10	6 o'clock of the bearing caeing	F1(0.006)	200
#11			350
#12			500
#19		F2(0.014)	200
#20			350
#21			500
#31		F3(0.019)	200
#32			350
#33			500
#40		F4(0.027)	200
#41			350
#42			500

TABLE 5.8: Bearing details and fault frequencies

Bearing type	FAG 22205E1KC3	
Characteristic frequencies ($\times$ rotation frequency f_r)	BPFO	6.1852
	BPFI	8.8148
	BSF	5.4030
	FTF	0.4123

The duration of the records is one second (40000 samples). According to the sampling frequency and rotating frequency, the window length N_w and NFFT are set to 512 and 1024 respectively. Due to the large difference between the rotation speeds in this case, different window shifts R are used, 64 for 200 rpm and 350 rpm and 32 for 500 rpm. Therefore, we can obtain the number of windows in STFT $N = \frac{40000-N_w}{R} + 1 = 618$ and the $N_k = NFFT/2 = 512$. So the degrees of freedom for 200 rpm and 350 rpm data is

$p_1 = N + N_k = 1130$. Similarly, the degrees of freedom for 500 rpm can also be obtained as $p_2 = 1747$. Accordingly, the statistical threshold for fault detection can be calculated as $\chi^2_{1130,0.95}/2 = 605.2$ and $\chi^2_{1747,1-\alpha}/2 = 923.2$. The results for fault detection, fault identification and fault size quantification are reported in Table 5.9. These faults are correctly detected and identified in all cases.

TABLE 5.9: Diagnosis results

Dataset	Detection		Identification				Characterization	
	Λ	Result	p_1	p_2	p_3	p_4	λ_2	l [mm]
#10	7975	Reject H_0	0.03	0.17	0.80	0	4.44	1.70
#11	10535	Reject H_0	0.03	0.16	0.81	0	3.71	1.58
#12	20109	Reject H_0	0.06	0.05	0.89	0	6.76	1.76
#19	8102	Reject H_0	0.02	0.12	0.86	0	4.31	1.66
#20	10825	Reject H_0	0	0.14	0.86	0	4.34	1.91
#21	25138	Reject H_0	0.11	0.06	0.83	0	6.72	1.75
#31	9599	Reject H_0	0.04	0.08	0.88	0	5.15	1.91
#32	17264	Reject H_0	0	0	1	0	4.23	1.85
#33	25955	Reject H_0	0	0	1	0	7.51	2.04
#40	3598	Reject H_0	0	0	1	0	5.40	1.99
#41	10361	Reject H_0	0	0.10	0.90	0	4.43	1.95
#42	23130	Reject H_0	0.15	0	0.85	0	7.60	2.08

Following similar lines as in the first case, the coda Δe is assessed from the relationship between parameter λ_2 and rotation frequency f_r. The estimated fault size based on the corrected duration time is displayed in the "Characterization" column. Figure 5.7 illustrates the fault size quantification through the model parameter λ_2. From this figure, it is found that the estimated value is slightly influenced by the motor rotation speed, thus illustrating the overall uncertainty of the method. The errors caused by fluctuations are located in the 90% confidence intervals based on the assumption of a uniform error with standard deviation equal to R/F_s. From the figure, it is found the error of the fault size quantification is still acceptable.

Next, the bearing data #12 associated to rolling element fault F1 (fault depth = 0.006 mm, fault area = 11.05 mm^2) and speed 500 rpm is selected to illustrate the diagnosis process. The LRT introduced in Chapter 4.2 is $\Lambda = 20109$, which is greater than the statistical threshold. Therefore, this signal is considered as symptomatic. The posterior probability spectrum $S(k)$ of the posterior probability sequence is displayed in Figure 5.8(a). It is seen that the fault frequency f_{BSF} and its harmonics dominate the spectrum,

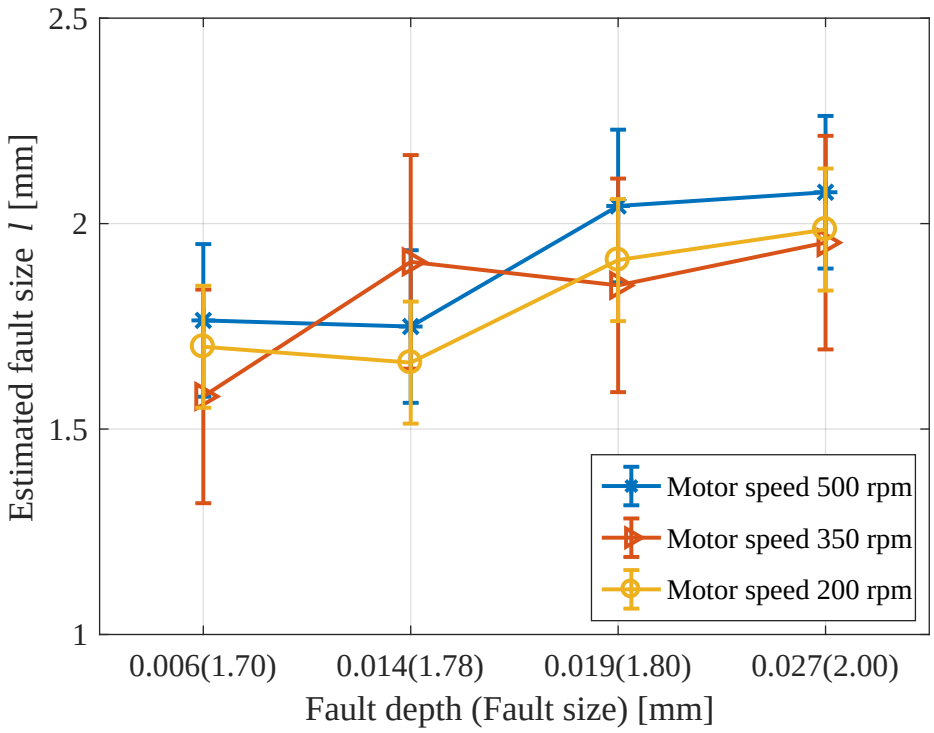

FIGURE 5.7: Quantification of the fault size together with 90% confidence intervals based on the assumption of a uniform error with standard deviation equal to R/F_s.

thus indicating a ball fault. For comparison, the state-of-the-art envelope spectrum – computed as in Case 1 – is displayed in Figure 5.8(b). In this case, it is less efficient than the PPS. The reason is that the signal contains one very strong impulse, much higher than the other ones (an observation quite typical of ball faults), which jeopardizes its spectral analysis (note that this could be fixed on an expert level by using the robust envelope analysis technique described in Ref. [28]). By comparison, the PPS is robust to outliers. Similarly, the posterior probabilities $p(f|t_{1:S})$ of the fault type displayed in Table 5.9. It is seen that $p(f_{BSF}|t_{1:S})$ is the highest, thus pointing out that the fault comes from the rolling element. One advantage of the latter is that it performed automatically, without visual examination.

The zoomed output of the EDHMM is shown in Figure 5.9. According to the duration time sequence $\tau_n(1)$ and $\tau_n(2)$ shown in Figure 5.9(c)-(d), the corresponding mean values λ_1 and λ_2 are 20.50 and 6.76 windows. The corresponding duration time are 16.40 ms and 5.40 ms, respectively. The corrected impulse duration time is calculated by subtracting the coda error Δe from Δt, equal to 3.79 ms. Following the lines of Chapter 4.5, the fault size l estimate returned by EDHMM is about 1.76 mm. According to the indication given

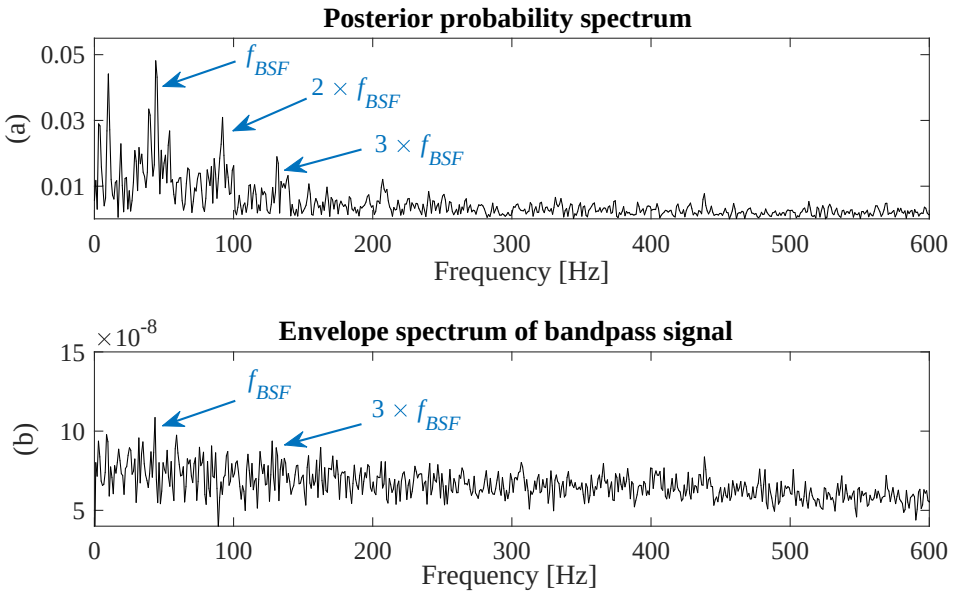

FIGURE 5.8: Spectral analysis of dataset # 12. (a) Posterior probability spectrum $S(k)$; (b) Envelope spectrum of bandpass signal with maximum kurtosis in band [8.8; 10] kHz selected from kurtogram.

in Ref. [139], the fault covers an area of 11.05 mm^2 and extends over the full length of the roller, that is 6.5 mm, the fault width is roughly 1.70 mm. It is found that the estimated value is not too far from the actual value.

Now, the capability of the proposed method to reconstruct the fault signal is validated on the record #12. Figure 5.10(a) shows the faulty signal with frequency $f_{BSF} = 45.025Hz$. The result of a reconstructed signal is shown in Figure 5.10(b). The observed signal in Figure 5.10(a) has a kurtosis level of 6.98, which indicates a bearing damage. The kurtosis of the recovered signal was 23.02, which is even a stronger indication of a damaged bearing. In the signal of subplot (a), it is not possible to identify all the spikes of a rolling element fault and some transient pulses with low energy are masked by background noise. After signal reconstruction, subplot (b) shows a typical sequence of spikes with a time interval of about 22.2 ms, very close to the characteristic defect frequency.

5.3 Accelerated degradation data

The previous section has focused on the failure resulting from an artificial damage. In this section, the life time dataset of rolling element bearings is used to gauge the performance of the proposed method in detecting a growing potential failure of a new

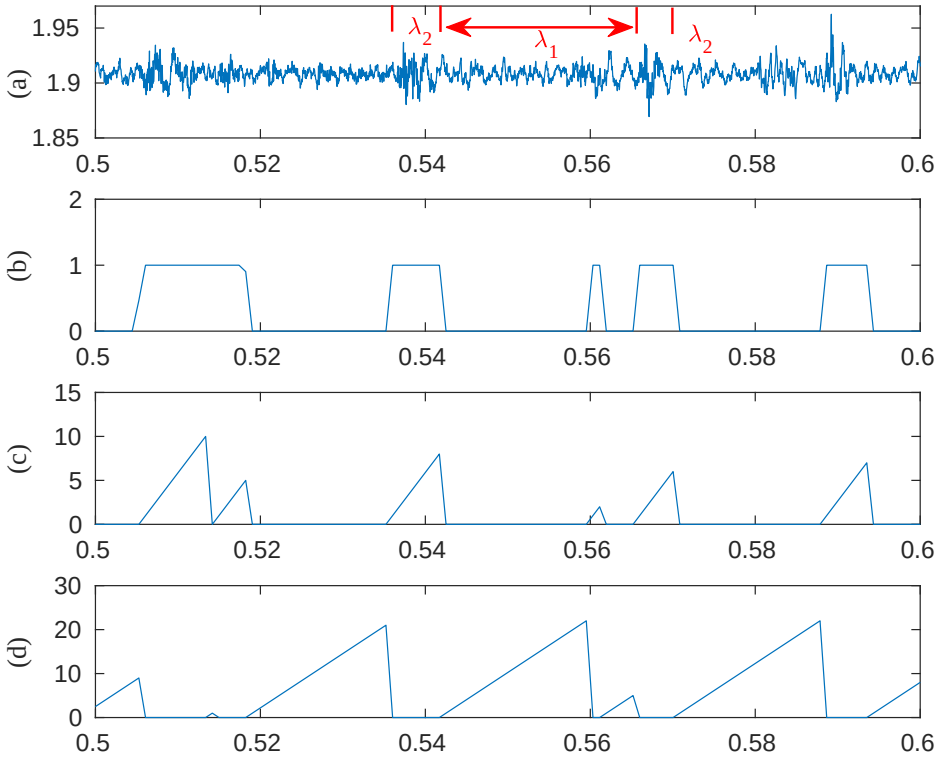

FIGURE 5.9: Zoom of #12 in interval [0.5 0.6] s. (a) Vibration signal; (b) posterior probability $\gamma_n(2)$; (c) duration time sequence $\tau_n(2)$ of active state; (d) duration time sequence $\tau_n(1)$ of inactive state.

bearing, which is eventually run to failure, and also in tracking fault size evolution. Results from naturally occurring fault of new bearings are more difficult to analysis than artificial defects.

The used datasets are provided by the Institute of Design Science and Basic Component at Xi'an Jiaotong University (XJTU) [150]. This datasets contain complete run-to-failure data of 15 rolling element bearings, and one dataset named Bearing 3_1 with the end of outer race fault is selected to demonstrate the effectiveness of the proposed integrated diagnosis framework. The type of tested bearings is LDK UER204, and the detailed parameters are given in Table 5.10. The sampling frequency is 25.6 kHz. As shown in Figure 5.11, a total of 32768 sampling points (i.e. 1.28 s) are recorded for each sampling, and the sampling period is equal to 1 min.

The total operating life is 42 h 18 min. The rotating speed is constant at 2400 rpm, and the radial load is 10 kN. Because the load is applied in the horizontal direction, the accelerometer placed in this direction is able to capture more degradation information of

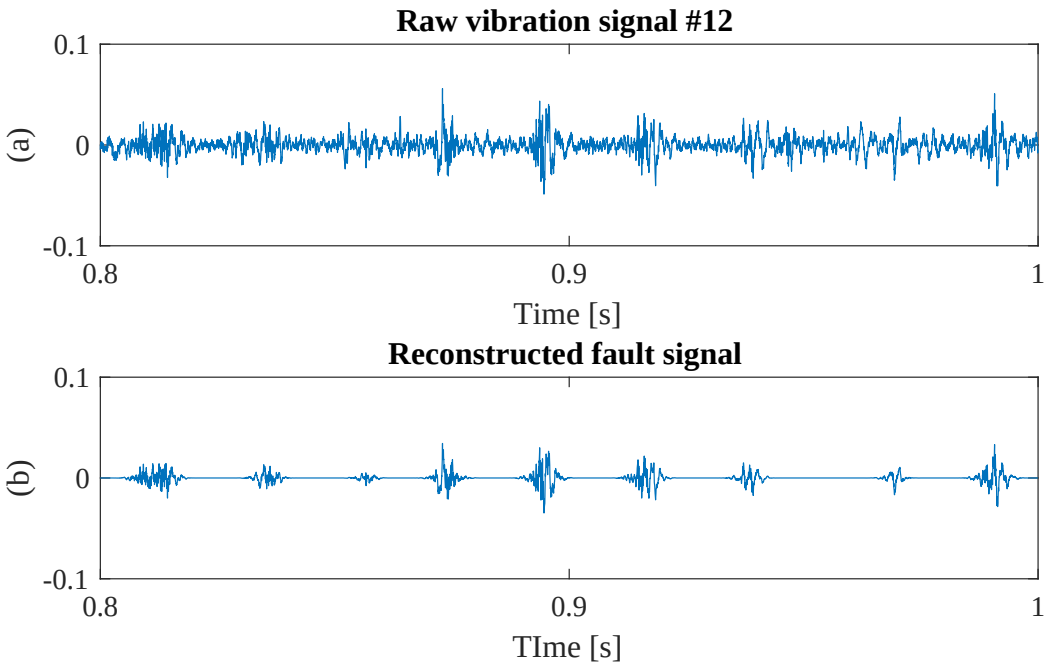

FIGURE 5.10: Fault signal reconstruction. (a) Raw vibration signal #12; (b) Reconstructed fault signal.

TABLE 5.10: Bearing details and fault frequencies

Bearing type		LDK UER204	
Outside diameter	39.80mm	Inside diameter	29.30 mm
Pitch diameter	34.55mm	Ball diameter	7.92 mm
Number of balls	8	Contact angle	0°
Dynamic load rating	12.82 kN	Static load rating	6.65 kN
Characteristic frequencies ($\times$ rotation frequency f_r)		BPFO	3.0831
		BPFI	4.9169
		BSF	2.0666
		FTF	0.3854

the tested bearings. Therefore, the horizontal vibration signal is selected to demonstrate the effectiveness of the proposed integrated diagnosis framework. Figure 5.12 shows the typical horizontal vibration signal during the whole operating life.

In common engineering applications, the selection process and expected fatigue life of rolling bearing are determined by the ISO standard 281:2007 [1]. Here, the bearing's basic fatigue life rating is defined as using the number of rotations for which 90% of all bearings in a specific group achieve or exceed a calculated time without failure (probability

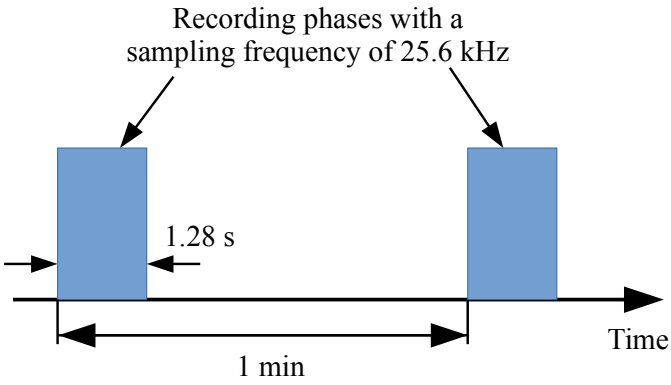

FIGURE 5.11: Signal acquisition settings.

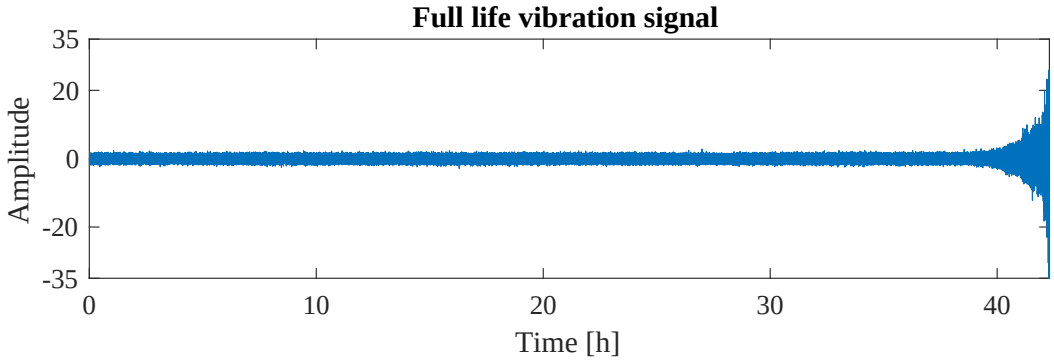

FIGURE 5.12: Full life vibration signal in the horizontal direction of the bearing 3_1.

of failure: 10%). According to [1], the theoretical expected fatigue life can be calculated using the following equation,

$$L_h = \frac{10^6}{60n}\left(\frac{C}{P}\right)^{\epsilon}, \tag{5.1}$$

where, L_h is the expected fatigue life (h), n is the rotating speed (rpm), C and P are the dynamic load rating (kN) and the applied dynamic load (kN) respectively. The exponent of the life equation ϵ is 3 for a ball bearing. The actual life time tends to be longer than the rating life. For this example, the expected life time can be calculated as 14.6 h, while the the actual situation is 42.3 h.

Some features were extracted from the time domain data such as the Kurtosis (KU), the RMS, the crest indicator (CI) and shape indicator (SI) of the vibration signal, shown in Figure 5.13. Definitions of these features have been described in Chapter 2.2. Trends of these features are plotted to indicate the progressive failure of the bearing over time. From Figure 5.13, it is seen that there is no clear indication of failure in the first 40 h, where the bearing operation is stable as seen in Figure 5.12. Data after about 40 h shows increasing

vibration energy. It can also be observed from Figure 5.13 that the features significantly increase in last several hours, in particular RMS. From the kurtosis plot, it is seen that values of about 3 indicate no failure condition. With fault initiation, the impacts of rolling elements generate impulses, leading to increasing kurtosis values. The failure appears to happen at approximately 40 hours. But there is no way to get the exact failure occurrence time, because there is no corresponding level or threshold for these time domain features. Therefore, the time domain indicators are effective for tracking the development of failures, but unable to indicate when the failures happen to appear.

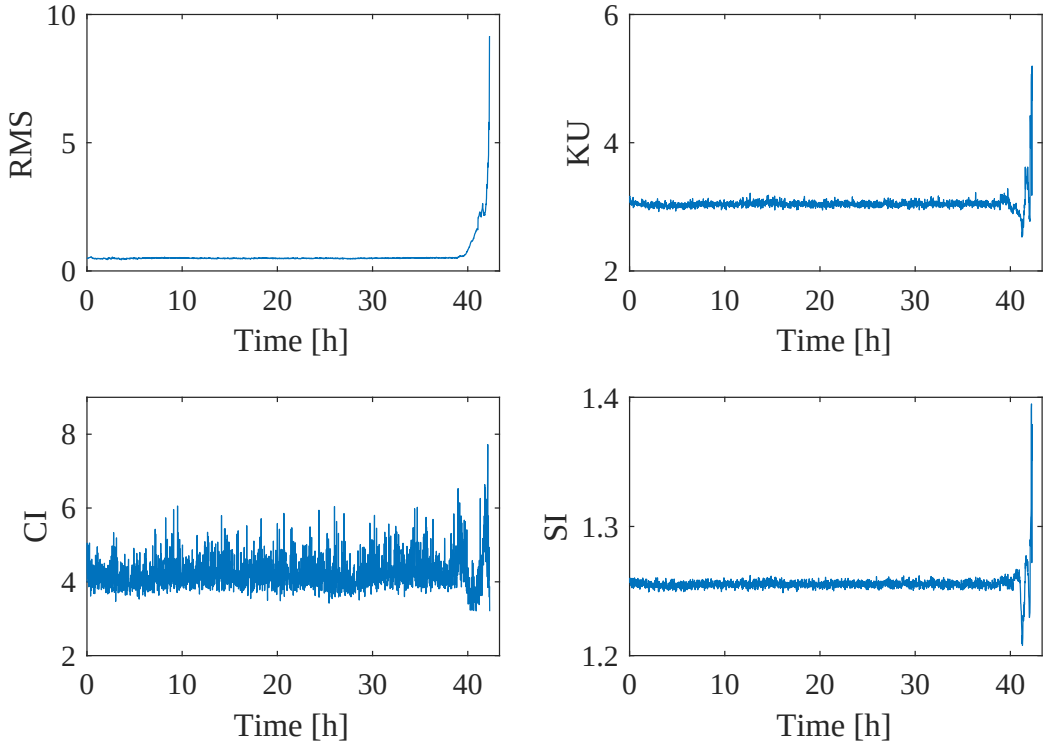

FIGURE 5.13: Bearing 3_1 vibration signal indicators; RMS: Root mean square; KU: Kurtosis; CI: Crest indicator; SI: Shape Indicator.

In order to figure out the exact failure occurrence time and track the progressive failure process, the integrated diagnosis framework is performed here to analyze these degradation data. According to the sampling frequency and rotating frequency, the window length N_w and window shift R are set to 256 and 16 samples, respectively.

First the fault detection algorithm is used to calculate the GLR Λ. Figure 5.14 (b) shows the vibration signal from 39 h, which is the interval of interest where the fault appear. The GLR Λ is calculated and shown by the blue line in Figure 5.14 (c). Each data point in the blue line represents the Λ for 1.28 seconds duration of the signal. The

time data is taken at 1 minute interval. Thus for 3.3 hours a total of 198 Λ points are calculated. The red dashed line represents the quantile level $\chi^2_{v,1-\alpha}/2$. The higher the value of Λ with respect to the threshold $\chi^2_{v,1-\alpha}/2$, the higher the probability of detection, and the lower the risk of false alarm. It is seen in Figure 5.14 (c) that the two lines cross around 39.6 h, which reflects the starting point of the failure. Compared with the time domain indicators plotted in Figure 5.13, GLR Λ has the ability to provide a threshold for comparison and deciding the health status. Figure 5.14 (d) displays the evolution of the Poisson parameter λ_2. This parameter controls the distribution of the transient pulse duration time, which can be estimated as the average number of windows that cover the transient pulse. A proportional relationship between the transient pulse duration and fault size seems a reasonable assumption. So λ_2 can be used as a quantity to characterize the fault size. In Figure 5.14 (d), each asterisk represents λ_2 for 1.28 seconds duration of the signal every two minutes. Then there are 81 obtained values for tracking quantitatively, rather than qualitatively like the other time domain indicators, the development of the failure. From the figure, it is seen that the λ_2 sequence gradually increases with the development of the failure, from the first value at about 5.0 to the end value at about 7.9. From the time the fault is found in the vibration signal to the end of the degradation test, the fault size increases by about a factor 1.6.

Compared to the traditional time domain indicators (Kurtosis, Skewness, RMS, crest, etc.), the advantage of the GLR Λ is that it follows the chi-squared distribution, which provides a theoretical quantile that can be used to make a decision. This is extremely important in engineering applications. In addition, in the proposed framework, the stochastic model also provides the mean transient pulse duration time, λ_2, which is able to track the development of faults in a proportional manner. This enables to characterize quantitatively the fault size, not just a qualitative analysis like the traditional time domain indicators.

After the fault occurrence point is detected, the subsequent signal can be saved and analyzed in details. Figure 5.15 displays the signal at 39.6 h, which shows bearing failure occurrence. Since the fault is incipient, the impulses are still weak, and strongly masked by background noise. This signal has a kurtosis level of 3.16, which complies with a Gaussian characteristic. This shows that this signal is challenging and difficult to diagnose. The identification approach of the proposed integrated framework is applied to identify the

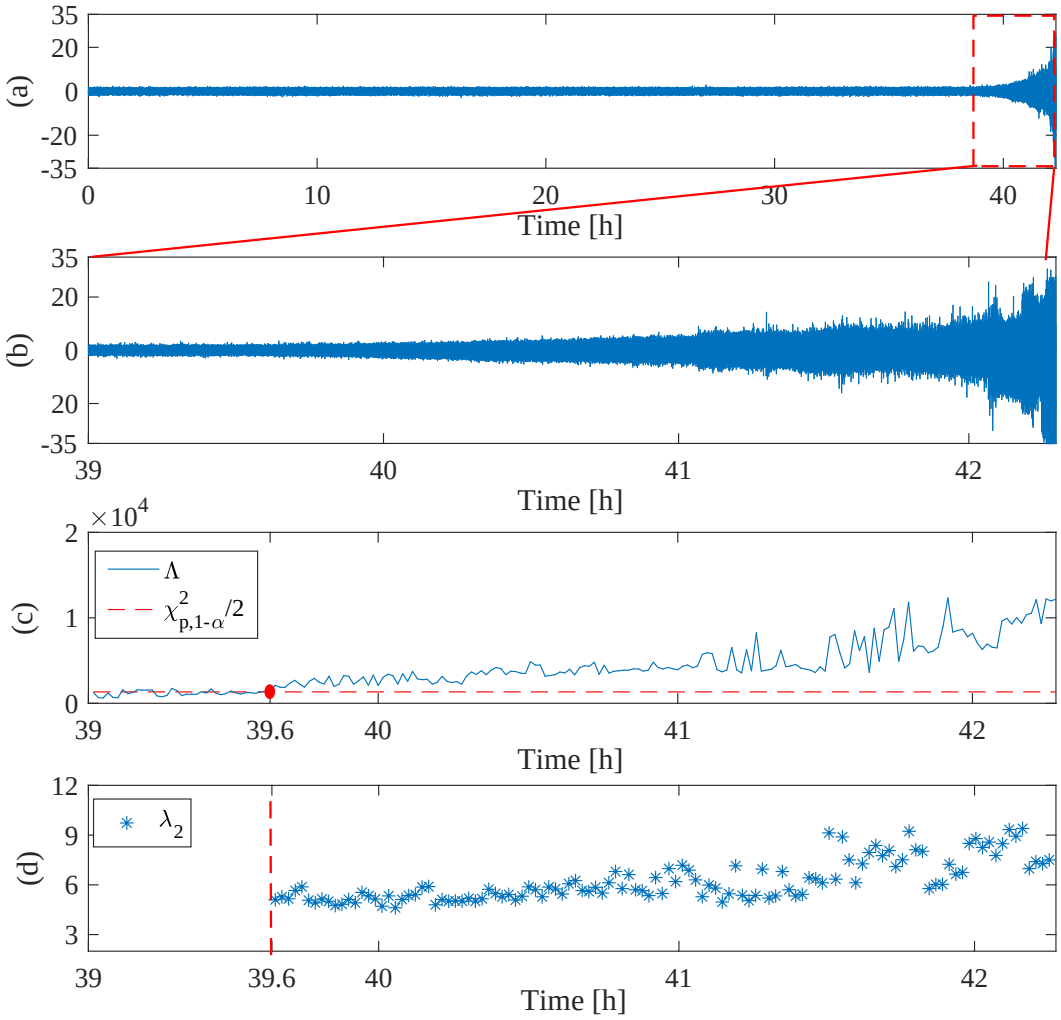

FIGURE 5.14: Illustration of fault detection and fault development tracking. (a) accelerated degradation vibration signal; (b) interval of fault occurrence; (c) generalized likelihood ratio Λ for detecting fault occurrence; (d) Poisson distribution parameter λ_2 for tracking the fault development.

fault type.

Figure 5.16 shows the histogram of the frequency data of the incipient fault signal, $f = 1/t_s, s \in \{1, 2, ..., S\}$. It is seen that most of the impulsive cycles fall within the frequency bin [112 135] Hz. This bin contains the outer race characteristic defect frequency f_{BPFO}, indicating that the transient pulses in the signal were generated to a certain extent by the outer race. The other posterior probabilities of different fault types are listed in Table 5.11. It is seen the probability of the outer race $p(f_{BPFO}|t_{1:S})$ is the highest, confirming the outer race fault. Such statistical analysis can be realized automatically without the visual examination of a spectrum.

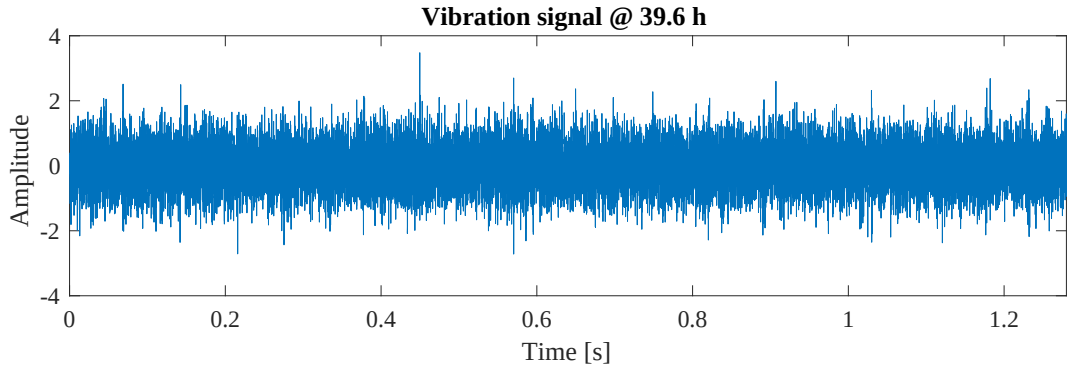

FIGURE 5.15: Signal with the incipient fault @ 39.6 h.

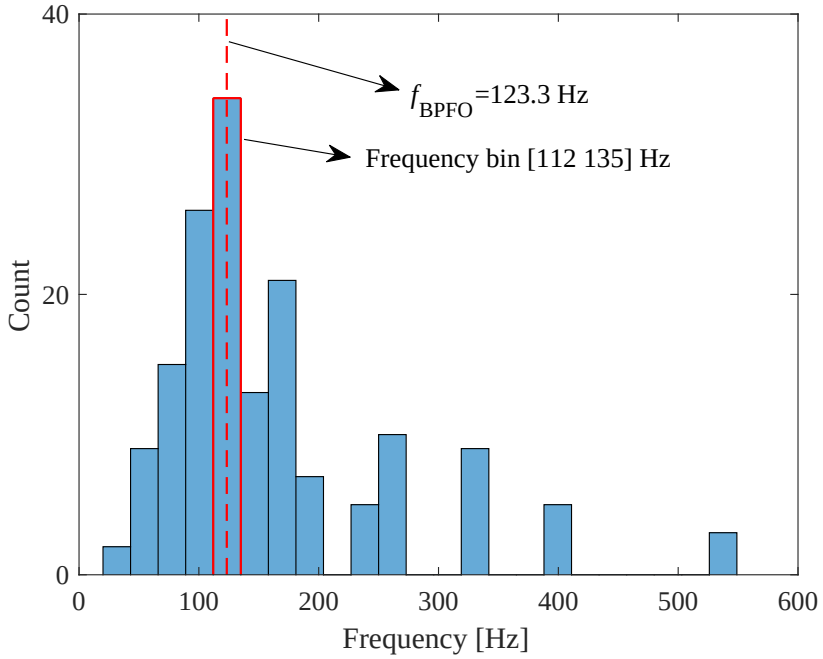

FIGURE 5.16: Histogram of the transient pulse cycle frequency, showing the red dash line relating to the fault characteristic frequency located in the highest bin.

TABLE 5.11: The posterior probabilities of fault type given $t_{1:S}$

Fault type	BPFI	BPFO	BSF	FTF
Posterior probability	$p(f_{BPFI}\|t_{1:S})$	$p(f_{BPFO}\|t_{1:S})$	$p(f_{BSF}\|t_{1:S})$	$p(f_{FTF}\|t_{1:S})$
	0.18	0.66	0.14	0.02

In order to confirm this statistical analysis, the posterior probability spectrum is shown in Figure 5.17(a). The rotating frequency f_r at 39.5 Hz and the outer race fault characteristic frequency f_{BPFO} at 123.3 Hz are very clearly visible in the posterior probability spectrum $S(k)$. For comparison, the state-of-the-art envelope spectrum of the

bandpass signal is computed and displayed in Figure 5.17(b), where the Kurtogram is used to select the most informative frequency band. The signal was band pass filtered between 3200 Hz to 4266 Hz as the most informative frequency region. In the envelope spectrum, although the characteristic frequency f_{BPFO} and its second harmonic can be seen, the spectrum contains many other unknown frequency components. The results of the signal at 39.6 h (bearing 3_1) shows that the stochastic model is capable of modeling the different states in the bearing fault signal to identify the characteristic defect frequencies.

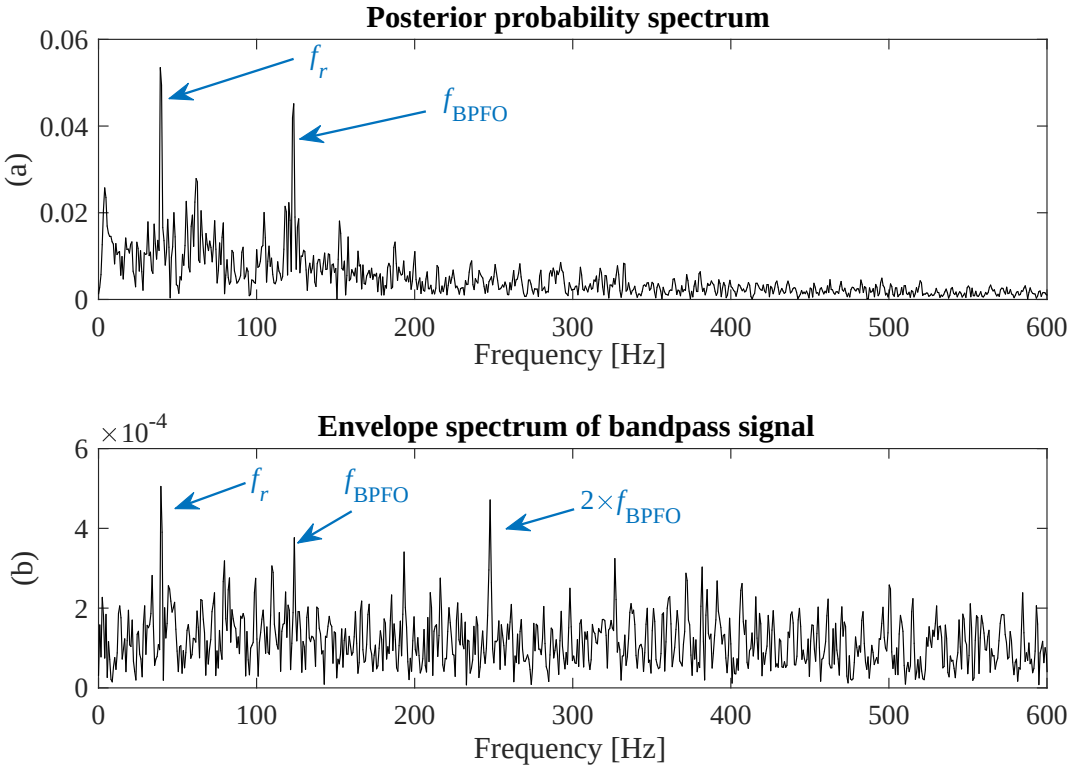

FIGURE 5.17: Spectral analysis of the signal at 39.6 h. (a) Posterior probability spectrum $S(k)$; (b) Envelope spectrum of bandpass signal with maximum kurtosis in band [3.2; 4.3] kHz selected from kurtogram.

At last, the capability of the proposed method to reconstruct the fault signal is validated on the signal at the fault occurrence point. Figure 5.18(a) displays the zoom of time domain in the interval [0.8, 1] s. It is seen that the raw vibration shows a rather stationary behavior, without the obvious fault manifestation. It is impossible to identify visually the spikes of an outer race fault in the time domain; most transient pulses with low energy are masked by background noise. After reconstructing the fault signal, Figure 5.18(b) shows that the majority of the transient pulses have been reconstructed; the time interval between the adjacent pulses is about 8.3 ms, very close to the outer race

characteristic frequency.

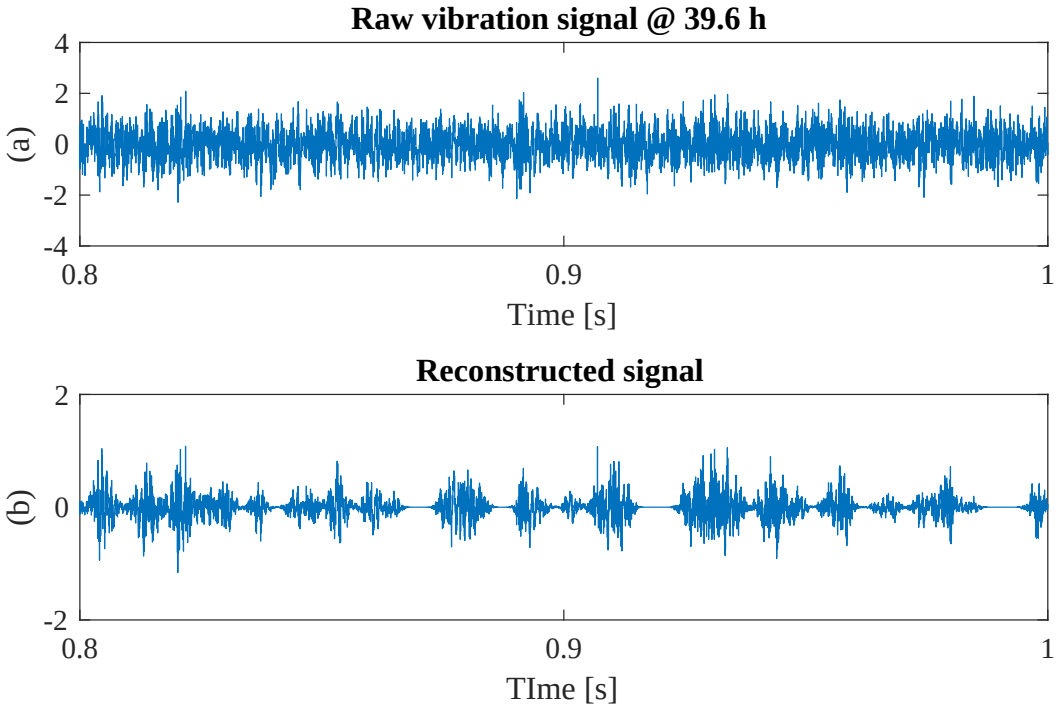

FIGURE 5.18: Fault signal reconstruction. (a) Raw vibration signal at 39.6 h; (b) Reconstructed fault signal.

5.4 Conclusion

This chapter has illustrated the application of the proposed integrated auto-diagnostic technique for detection, identification, quantification of a fault and fault signal reconstruction. Two kinds of experimental data have been considered in this chapter. The results on 36 sets of bearing data associated to three different fault cases has shown the efficiency of this framework. Almost all faults have been detected through the stochastic indicator GLR. Although, the bearing characteristic frequency was modulated by the shaft frequency thus producing a series of sidebands and harmonics, which has a big influence on the identification, the proposed identification technique overcame this problem well. For the signal reconstruction technique, the transients interval became clearer after fault signal reconstruction. In addition, from Figure 5.1 and 5.7, the fault size has been assessed close to the real size, and within the 90% confidence intervals, thus validating the technique.

In the accelerated degradation data, the failure is natural, not artificial like the first experimental validation. On this set of data, it is important to detect an incipient fault in

advance before it continues to develop. After validation, it is found that the technique is effective for detecting the fault occurrence time, and is also capable of providing information on its propagation and tracking quantitatively the evolution of the crack.

Conclusions and perspectives

Conclusion

The main objective of this thesis is the investigation of automatic signal processing methods for different diagnostic tasks in one integrated framework. The remarkable advantages are that it performs without the historic data, without few hyperparameters, and without any manual intervention in the process. The literature review shows that the most existing techniques are focused on one specific diagnostic task. But these techniques are almost independent, which requires users to have a high level of experience and knowledge when dealing with different diagnostic problems. Some methods attempt to fulfill such aspiration, yet not reaching the automated and integrated level as this thesis does.

The thesis introduced an automated diagnosis framework rooted on a stochastic model of the vibration signals produced by rolling element bearing faults. The proposed method can achieve the fault detection, fault type identification, fault signal reconstruction and fault size characterization at once. The distinguishing feature of this method is that it is performed automatically, which means that no human intervention or manual tuning of the algorithms is required for solving these subtasks. The whole process depends mainly on only one key hyperparameter, i.e. the window length N_w in the STFT.

Firstly, a technique based on hypothesis test is introduced to detect the healthy status of the vibration signal. The test statistics is the GLR constructed through the EDHMM parameter, i.e. the covariance matrix $\mathbb{C}$ in different states, and it asymptotically follows a chi-squared distribution according to Wilk's theorem. Unlike the supervised approach, this technique does not require prior knowledge or training data.

Secondly, the fault signal is extracted from the noisy vibration signal based on the

Bayesian theorem and EDHMM parameters. After modeling through EDHMM, the STFT coefficients related to the failure can be singled out and estimated as the active state, and then the inverse STFT is used to reconstruct the fault signal. Compared with the Wiener filter, this technique can be seen as a time-varying filter, which takes account of the posterior probability γ_n.

Subsequently, two different techniques are proposed based on the EDHMM parameters to identify the fault type. The PPS is the spectrum of the posterior probability, a robust alternative to the standard envelope spectrum, free of any pre-processing. Unlike the signal spectrum, the PPS is not sensitive to the amplitude of the transient pulses, so it can avoid the influence of modulation, thereby reducing the sidebands in the spectrum. However, in order to completely escape visual inspection, a simple statistical method is introduced for identifying automatically the fault type. It is based on the duration time of each transient period, to calculate the probabilities of different fault types. Then the fault identification is achieved by a probability comparison.

At last, the fault size quantification is addressed. This technique is based on the hypothesis that a proportional relationship exists between the transient pulse duration and fault size. So we can get the linear relationship between the duration time and fault size based on the geometry structure of the rolling element bearing. Therefore, after some mathematical manipulation, the estimated fault size $\hat{l}$ and the coda error Δe can be estimated as the slope and the intercept, respectively, of the fitted lines with respect to machine speed. The Poisson parameter λ_2, i.e. the expected duration time of transient pulse, can also track quantitatively the fault size evolution.

The above techniques for different diagnostic tasks are all based on the EDHMM parameters. A nonstationary bearing signal cannot be entirely modeled by a deterministic model, while EDHMM, as one stochastic model, has gifted flexibility and adaptability to cope with the complexity of non-stationary phenomena. Finally, it should be mentioned that the proposed integrated auto-diagnostic framework strongly depends on this stochastic model, which poses a risk to the diagnosis. This will be considered in future research. In the last chapter, the capability of the proposed automated diagnosis framework has been validated through two different experimental scenarios. The results of the experimental validation not only meets the expectation of diagnosis, but also reflect its unique advantages, i.e. the automation and integration.

Perspectives

So far, it is time to close this thesis. The previous chapters described how to model the signal, how to estimate the model parameters, how to use the parameters for integrated diagnosis and the performance in the experimental validation. Although everything seems perfect, there are several aspects that have not been considered in the thesis. In this part, some of the future work directions related to this research are suggested.

- **Simplicity and reliability**
  One of the most crucial aspects that should be kept in mind is that the ultimate goal of scientific research is to find practical application. In the practical CBM, the reliability of the technique rather than its sophistication is the primary pursuit. It is often required that part or even all of the proposed technique should be suitable for the slave computer or microcomputer, which means the method should be performed in a simple and automated way. Therefore, computational optimization of this integrated diagnosis framework is the next step to be done.

- **Ability for variable speed conditions**
  In addition, in most industrial environments, the speed is often an uncontrollable variable, e.g. as with wind turbines. However, the current framework is only valid at constant speed. Angle-time analysis [2][12] could potentially be jointed with the EDHMM in future work to achieve the integrated diagnosis for variable rotating speed signals.

- **Diversity of applicable object**
  This research attempted to establish an integrated diagnosis framework for rolling element bearings. So another future work is the application of integrated diagnosis in condition monitoring applications for other rotational machine components such as gears, pumps, blades etc.

- **Comprehensiveness of the framework**
  One last recommendation for future scientific work is related to the embedding of the RUL forecast. The reliable RUL estimation is a challenging task and also very useful in real CBM. In the current framework, the Poisson parameter λ_2 (i.e. the expected duration time of transient pulse) can quantitatively track the evolution of the fault size. Since this consideration could lead to increased prognostic performance, further

research effort in this field is suggested.

Appendix A

Parameter estimation

In this section, the detail mathematical process about the Eq.(3.23) and (3.25) is demonstrated. The logarithmic likelihood function of the observation is obtained by marginalizing over the hidden variables as,

$$L(\boldsymbol{\theta}) = \ln p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta}) = \ln\{\sum_{\boldsymbol{Z}} p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})\}, \tag{A.1}$$

where, $\boldsymbol{Z} = z_{1:N}$. The goal is to find the optimal parameter $\boldsymbol{\theta}$ that maximizes the likelihood function $L(\boldsymbol{\theta})$. But in this likelihood function, it is found that the summation over the hidden states appears inside the logarithm. As the hidden variables $\boldsymbol{Z}$ are also unknown, which brings difficulties to direct maximization. We therefore turn to the iterative algorithm to find an efficient framework for maximizing indirectly the likelihood function. Some manipulation on the logarithmic likelihood function $L(\boldsymbol{\theta})$ are performed as follow,

$$\begin{aligned}
L(\boldsymbol{\theta}) &= \ln\{\sum_{\boldsymbol{Z}} p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})\} \\
&= \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \frac{p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old})} \\
&\geq \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln \frac{p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old})} \\
&= \sum_{\boldsymbol{Z}} \{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})\} - \sum_{\boldsymbol{Z}} \{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old})\}.
\end{aligned}$$

The updated $\boldsymbol{\theta}^{new}$ can be found by maximizing the logarithmic likelihood function

$L(\boldsymbol{\theta})$ in each iteration. As maximizing $L(\boldsymbol{\theta})$ with respect to the parameters $\boldsymbol{\theta}$, then we remove the items that are not related to the parameters $\boldsymbol{\theta}$. And using the $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ as the new objective function, then it can be simplified as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\theta}). \tag{A.2}$$

We substitute the joint probability $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ into $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ and make use the definitions of γ and ξ, $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ can be written as,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{i} \gamma_1(i) \ln \pi_i + \sum_{n=2}^{N} \sum_{i} \sum_{j} \xi_n(i,j) \ln a_{ij} + \sum_{n=1}^{N} \sum_{i} \gamma_n(i) \ln p(\boldsymbol{Y}_n|\boldsymbol{\theta}_k). \tag{A.3}$$

Maximization with respect to $\boldsymbol{\pi}$ is achieved using appropriate Lagrange multipliers. First, setting the derivatives of the equation (A.3) with respect to π_i to zero, together with the constrain $\sum_i \pi_i = 1$, we can obtain $\frac{\gamma_1(i)}{\pi_i} + \rho = 0$. Joint with the the constrain $\sum_i \pi_i = 1$, the formula can be written as $\sum_i \gamma_1(i) = -\rho$. Then the parameter $\boldsymbol{\pi}$ is given by,

$$\pi_i = \frac{\gamma_1(i)}{\sum_i \gamma_1(i)}. \tag{A.4}$$

We take the derivative with respect to the parameter $\mathbb{C}_i$ as follow,

$$\begin{aligned}
\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) &= -\sum_{n=1}^{N} \gamma_n(i) \left[\frac{\partial [\ln |\mathbb{C}_i| + \boldsymbol{Y}_n^{\dagger} \mathbb{C}_i^{-1} \boldsymbol{Y}_n]}{\partial \mathbb{C}_i} \right] \\
&= -\sum_{n=1}^{N} \gamma_n(i) \left[tr(\mathbb{C}_i^{-1} \partial \mathbb{C}_i) - \boldsymbol{Y}_n^{\dagger} \mathbb{C}_i^{-1} \partial \mathbb{C}_i \mathbb{C}_i^{-1} \boldsymbol{Y}_n \right] \\
&= -\sum_{n=1}^{N} \gamma_n(i) \left[tr(\mathbb{C}_i^{-1} \partial \mathbb{C}_i) - tr(\boldsymbol{Y}_n^{\dagger} \mathbb{C}_i^{-1} \partial \mathbb{C}_i \mathbb{C}_i^{-1} \boldsymbol{Y}) \right] \\
&= -\sum_{n=1}^{N} \gamma_n(i) \left[tr(\mathbb{C}_i^{-1} \partial \mathbb{C}_i) - tr(\mathbb{C}_i^{-1} \boldsymbol{Y} \boldsymbol{Y}^{\dagger} \mathbb{C}_i^{-1} \partial \mathbb{C}_i) \right] \\
&= -\sum_{n=1}^{N} \gamma_n(i) \left[tr([\mathbb{C}_i^{-1} - C_k^{-1} \boldsymbol{Y} \boldsymbol{Y}^{\dagger} \mathbb{C}_i^{-1}] \partial \mathbb{C}_i) \right]
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})}{\partial \mathbb{C}_i} &= -\sum_{n=1}^{N} \gamma_n(i) [\mathbb{C}_i^{-1} - \mathbb{C}_i^{-1} \boldsymbol{Y} \boldsymbol{Y}^{\dagger} \mathbb{C}_i^{-1}] = 0 \\
\Longrightarrow \mathbb{C}_i &= \frac{\sum_{n=1}^{N} \gamma_n(i) \boldsymbol{Y} \boldsymbol{Y}^{\dagger}}{\sum_{n=1}^{N} \gamma_n(i)}.
\end{aligned} \tag{A.5}$$

Appendix B

Proof of the convergence

In this section, the detailed proof about the convergence of the iterative algorithm in Chapter 3.4 is given. Intuitively, as long as it can be proved that the new estimated parameter $\boldsymbol{\theta}^{new}$ makes the logarithmic likelihood function $L(\boldsymbol{\theta}^{new})$ greater than the previous iteration, then the iterative algorithm converges. The proof of the convergence is equivalent to prove $L(\boldsymbol{\theta}^{new}) > L(\boldsymbol{\theta}^{old})$. If proven, the $L(\boldsymbol{\theta})$ will be maximized after a sufficient number of iterations. Some manipulations about the logarithmic likelihood function $L(\boldsymbol{\theta})$ are made as follow,

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \ln p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta}) \\
&= \ln \frac{p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln \frac{p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{Z}} \{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})\} - \sum_{\boldsymbol{Z}} \{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta})\} \\
&= Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) - H(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}),
\end{aligned}
$$

where,

$$
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} \{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Y}_{1:N}, \boldsymbol{Z}|\boldsymbol{\theta})\} \tag{B.1}
$$

$$H(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} \{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta})\}. \tag{B.2}$$

As shown in Appendix A, the new parameter $\boldsymbol{\theta}^{new}$ is obtained by maximizing the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$, i.e. $\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$. Absolutely, the relationship is beyond doubt,

$$Q(\boldsymbol{\theta}^{new}, \boldsymbol{\theta}^{old}) > Q(\boldsymbol{\theta}^{old}, \boldsymbol{\theta}^{old}). \tag{B.3}$$

So if $H(\boldsymbol{\theta}^{new}, \boldsymbol{\theta}^{old}) < H(\boldsymbol{\theta}^{old}, \boldsymbol{\theta}^{old})$ is also proved, then it can explain the convergence, $L(\boldsymbol{\theta}^{new}) > L(\boldsymbol{\theta}^{old})$. The demonstration of $H(\boldsymbol{\theta}^{new}, \boldsymbol{\theta}^{old}) < H(\boldsymbol{\theta}^{old}, \boldsymbol{\theta}^{old})$ is given as follow,

$$\begin{aligned} H(\boldsymbol{\theta}^{new}, \boldsymbol{\theta}^{old}) - H(\boldsymbol{\theta}^{old}, \boldsymbol{\theta}^{old}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{new})}{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old})} \\ &\leq \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old}) \frac{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{new})}{p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{old})} \\ &= \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{Y}_{1:N}, \boldsymbol{\theta}^{new}) \\ &= 0. \end{aligned} \tag{B.4}$$

The convergence is proven,

$$L(\boldsymbol{\theta}^{new}) > L(\boldsymbol{\theta}^{old}). \tag{B.5}$$

Bibliography

[1] ISO 281:2007. "Rolling Bearings—Dynamic Load Ratings and Rating Life". In: *International Organization for Standardization: Geneva, Switzerland* (2007).

[2] Dany Abboud, Jerome Antoni, Mario Eltabach, and Sophie Sieg Zieba. "Angle time cyclostationarity for the analysis of rolling element bearing vibrations". In: *Measurement* (2015), pp. 29–39.

[3] Nur Ashar Aditiya, Muhammad Rizky Dharmawan, Zaqiatud Darojah, and Raden Sanggar. "Fault diagnosis system of rotating machines using Hidden Markov Model (HMM)". In: *2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*. IEEE. 2017, pp. 177–181.

[4] Alireza Moazen Ahmadi, Carl Howard, and Dick Petersen. "The path of rolling elements in defective bearings: observations, analysis and methods to estimate spall size". In: *Journal of Sound and Vibration* (2016), pp. 277–292.

[5] R.J. Alfredson and Jhansi Mathew. "Time domain methods for monitoring the condition of rolling element bearings". In: *Nasa Sti/recon Technical Report A* (1985), pp. 102–107.

[6] Jerome Antoni. "A Critical Overview of the "Filterbank-Feature-Decision" Methodology in Machine Condition Monitoring". In: *Acoustics Australia* (2021), pp. 1–8.

[7] Jerome Antoni. "Cyclic spectral analysis of rolling-element bearing signals: Facts and fictions". In: *Journal of Sound and vibration* (2007), pp. 497–529.

[8] Jerome Antoni. "Cyclostationarity by examples". In: *Mechanical Systems and Signal Processing* (2009), pp. 987–1036.

[9] Jerome Antoni. "Fast computation of the kurtogram for the detection of transient faults". In: *Mechanical Systems and Signal Processing* (2007), pp. 108–124.

[10] Jerome Antoni. "The infogram: Entropic evidence of the signature of repetitive transients". In: *Mechanical Systems and Signal Processing* (2016), pp. 73–94.

[11] Jerome Antoni. "The spectral kurtosis: a useful tool for characterising non-stationary signals". In: *Mechanical systems and signal processing* (2006), pp. 282–307.

[12] Jerome Antoni, Dany Abboud, and Sophie Baudin. "Time-angle periodically correlated processes". In: *Cyclostationarity: theory and methods*. Springer, 2014, pp. 3–14.

[13] jerome Antoni, Frederic Bonnardot, Amani Raad, and Mohamed El Badaoui. "Cyclostationary modelling of rotating machine vibration signals". In: *Mechanical systems and signal processing* (2004), pp. 1285–1314.

[14] Jerome Antoni and Pietro Borghesani. "A statistical methodology for the design of condition indicators". In: *Mechanical Systems and Signal Processing* (2019), pp. 290–327.

[15] Jerome Antoni and Robert Bond Randall. "The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines". In: *Mechanical systems and signal processing* (2006), pp. 308–331.

[16] Bulent Ayhan, Chiman Kwan, Stephen Chu, and Kennth Loparo. "Integrated Bearing Diagnostics and Fault Severity Estimation Using HMM". In:

[17] Mehran Azimi, Panos Nasiopoulos, and Rabab Kreidieh Ward. "Offline and online identification of hidden semi-Markov models". In: *IEEE Transactions on Signal Processing* (2005), pp. 2658–2663.

[18] Piero Baraldi, Luca Podofillini, Lusine Mkrtchyan, Enrico Zio, and Vinh Dang. "Comparing the treatment of uncertainty in Bayesian networks and fuzzy expert systems used for a human reliability analysis application". In: *Reliability Engineering & System Safety* (2015), pp. 176–193.

[19] Ross Barrett and David Holdsworth. "Frequency tracking using hidden Markov models with amplitude and phase information". In: *IEEE Transactions on Signal Processing* (1993), pp. 2965–2976.

[20] Leonard E Baum and John Alonzo Eagon. "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology". In: *Bulletin of the American Mathematical Society* (1967), pp. 360–363.

[21] Leonard E Baum and Ted Petrie. "Statistical inference for probabilistic functions of finite state Markov chains". In: *The annals of mathematical statistics* (1966), pp. 1554–1563.

[22] Eric Bechhoefer, Andreas Bernhard, David He, and Pat Banerjee. "Use of hidden semimarkov models in the prognostics of shaft failure". In: *HIP* (2006), p. 3.

[23] Jaouher Benali, Mounir Sayadi, Farhat Fnaiech, Brigitte Morello, and Noureddine Zerhouni. "Importance of the fourth and fifth intrinsic mode functions for bearing fault diagnosis". In: *14th International Conference on Sciences and Techniques of Automatic Control & Computer Engineering-STA'2013*. IEEE. 2013, pp. 259–264.

[24] Michael Berouti, Richard Schwartz, and John Makhoul. "Enhancement of speech corrupted by acoustic noise". In: *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE. 1979, pp. 208–211.

[25] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[26] Steven Boll. "Suppression of acoustic noise in speech using spectral subtraction". In: *IEEE Transactions on acoustics, speech, and signal processing* (1979), pp. 113–120.

[27] Frederic Bonnardot, Robert Randall, and Francois Guillet. "Extraction of second-order cyclostationary sources—application to vibration analysis". In: *Mechanical Systems and Signal Processing* (2005), pp. 1230–1244.

[28] Pietro Borghesani and Jerome Antoni. "CS2 analysis in presence of non-Gaussian background noise – Effect on traditional estimators and resilience of log-envelope indicators". In: *Mechanical Systems and Signal Processing* (2017), pp. 378–398.

[29] Pietro Borghesani, Paolo Pennacchi, and Steven Chatterton. "The relationship between kurtosis-and envelope-based indexes for the diagnostic of rolling element bearings". In: *Mechanical Systems and Signal Processing* (2014), pp. 25–43.

[30] Pietro Borghesani, Paolo Pennacchi, Robert Randall, Nader Sawalhi, and Roberto Ricci. "Application of cepstrum pre-whitening for the diagnosis of bearing faults under variable speed conditions". In: *Mechanical Systems and Signal Processing* (2013), pp. 370–384.

[31] Tony Boutros. "Fault detection and diagnosis in machining processes and rotating machinery using fuzzy approach and hidden Markov model". PhD thesis. University of Ottawa (Canada), 2006.

[32] Tony Boutros and Ming Liang. "Detection and diagnosis of bearing and cutting tool faults using hidden Markov models". In: *Mechanical Systems and Signal Processing* (2011), pp. 2102–2124.

[33] David Brillinger. "Fourier analysis of stationary processes". In: *Proceedings of the IEEE* (1974), pp. 1628–1643.

[34] Marco Buzzoni, Jerome Antoni, and Gianluca d'Elia. "Blind deconvolution based on cyclostationarity maximization and its application to fault identification". In: *Journal of Sound and Vibration* (2018), pp. 569–601.

[35] Marco Buzzoni, Elia Soave, Gianluca D'Elia, Emiliano Mucchi, and Giorgio Dalpiaz. "Development of an indicator for the assessment of damage level in rolling element bearings based on blind deconvolution methods". In: *Shock and Vibration* (2018).

[36] Francesco Cartella, Jan Lemeire, Luca Dimiccoli, and Hichem Sahli. "Hidden semi-Markov models for predictive maintenance". In: *Mathematical Problems in Engineering* (2015).

[37] *Case Western Reserve University Bearing Data Center Website*. http://csegroups.case.edu/bearingdatacenter/home.

[38] Biao Chen and Peter Willett. "Detection of hidden Markov model transient signals". In: *IEEE Transactions on Aerospace and Electronic systems* (2000), pp. 1253–1268.

[39] Yeunung Chen. "Cepstral domain talker stress compensation for robust speech recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1988), pp. 433–439.

[40] Zhongsheng Chen and Yongmin Yang. "Fault diagnostics of helicopter gearboxes based on multi-sensor mixtured hidden Markov models". In: *Journal of vibration and acoustics* (2012).

[41] Zhongsheng Chen, Yongmin Yang, Zheng Hu, and Qinghu Zeng. "Fault prognosis of complex mechanical systems based on multi-sensor mixtured hidden semi-Markov models". In: *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* (2013), pp. 1853–1863.

[42] Yao Cheng, Ning Zhou, Weihua Zhang, and Zhiwei Wang. "Application of an improved minimum entropy deconvolution method for railway rolling element bearing fault diagnosis". In: *Journal of sound and vibration* (2018), pp. 53–69.

[43] Li Deng, Michael Aksmanovic, Xiaodong Sun, and Jeff Wu. "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states". In: *IEEE transactions on speech and audio processing* (1994), pp. 507–520.

[44] Guangming Dong and Jin Chen. "Noise resistant time frequency analysis and application in fault diagnosis of rolling element bearings". In: *Mechanical Systems and Signal Processing* (2012), pp. 212–236.

[45] Ming Dong. "A novel approach to equipment health management based on auto-regressive hidden semi-Markov model (AR-HSMM)". In: *Science in china series F: information sciences* (2008), pp. 1291–1304.

[46] Ming Dong and David He. "A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology". In: *Mechanical systems and signal processing* (2007), pp. 2248–2266.

[47] Ming Dong and David He. "Hidden semi-Markov model-based methodology for multi-sensor equipment health diagnosis and prognosis". In: *European Journal of Operational Research* (2007), pp. 858–878.

[48] Ming Dong, David He, Prashant Banerjee, and Jonathan Keller. "Equipment health diagnosis and prognosis using hidden semi-Markov models". In: *The International Journal of Advanced Manufacturing Technology* (2006), pp. 738–749.

[49] Ming Dong and Ying Peng. "Equipment PHM using non-stationary segmental hidden semi-Markov model". In: *Robotics and Computer-Integrated Manufacturing* (2011), pp. 581–590.

[50] Konstantin Dragomiretskiy and Dominique Zosso. "Variational mode decomposition". In: *IEEE transactions on signal processing* (2013), pp. 531–544.

[51] Jean-Paul Dron, Fabrice Bolaers, and Lanto Rasolofondraibe. "Improvement of the sensitivity of the scalar indicators (crest factor, kurtosis) using a de-noising method by spectral subtraction: application to the detection of defects in ball bearings". In: *Journal of Sound and Vibration* (2004), pp. 61–73.

[52] Chaoqun Duan, Viliam Makis, and Chao Deng. "Optimal Bayesian early fault detection for CNC equipment using hidden semi-Markov process". In: *Mechanical Systems and Signal Processing* (2019), pp. 290–306.

[53] Van Dyer and R.M. Stewart. "Detection of Rolling Element Bearing Damage by Statistical Vibration Analysis". In: *Journal of Mechanical Design* (1978), pp. 229–235.

[54] Yariv Ephraim and David Malah. "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator". In: *IEEE transactions on acoustics, speech, and signal processing* (1985), pp. 443–445.

[55] Yariv Ephraim and David Malah. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". In: *IEEE Transactions on acoustics, speech, and signal processing* (1984), pp. 1109–1121.

[56] Iain Kenton Epps. "An investigation into vibrations excited by discrete faults in rolling element bearings". In: (1991).

[57] Lei Fan, Shaoping Wang, Haibin Duan, and Hongliang Ran. "Fatigue crack fault diagnosis and prognosis based on hidden semi-Markov model". In: *The Journal of Engineering* (2019), pp. 406–410.

[58] Marcin Firla, Zhong-Yang Li, Nadine Martin, Christian Pachaud, and Tomasz Barszcz. "Automatic characteristic frequency association and all-sideband demodulation for the detection of a bearing fault". In: *Mechanical Systems and Signal Processing* (2016), pp. 335–348.

[59] Nolazco Flores and Steve Young. "Continuous speech recognition in noise using spectral subtraction and HMM adaptation". In: *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 1994, p. 409.

[60] John Freguson. "Variable duration models for speech". In: *Proc. Symposium on the Application of Hidden Markov Models to Text and Speech, 1980*. 1980.

[61] Omid Geramifard, Jianxin Xu, and W-Y Chen. "An HMM-based semi-nonparametric approach for fault diagnostics in rotary electric motors". In: *2012 IEEE International Symposium on Industrial Electronics*. IEEE. 2012, pp. 1218–1223.

[62] John Goutsias and Jerry Mendel. "Optimal simultaneous detection and estimation of filtered discrete semi-Markov chains". In: *IEEE transactions on information theory* (1988), pp. 551–568.

[63] Wei Guo, Tse Peter, and Alexandar Djordjevich. "Faulty bearing signal recovery from large noise using a hybrid method based on spectral kurtosis and ensemble empirical mode decomposition". In: *Measurement* (2012), pp. 1308–1322.

[64] Beniamino Hadj-Amar, Barbel Finkenstadt, Mark Fiecas, and Robert Huckstepp. "A Spectral Hidden Markov Model for Nonstationary Oscillatory Processes". In: (2020).

[65] FangZhou Hao, Zili Li, Haiyan Wen, and Hongshan Zhao. "Remaining lifetime prediction of distribution transformer based on improved hidden semi-markov model". In: *Journal of Physics: Conference Series*. IOP Publishing. 2019, p. 012052.

[66] David He, Shenliang Wu, Pat Banerjee, and Eric Bechhoefer. "Probabilistic model based algorithms for prognostics". In: *2006 IEEE Aerospace Conference*. IEEE. 2006, 10–pp.

[67] Justyna Hebda-Sobkowicz, Radoslaw Zimroz, and Agnieszka Wylomanska. "Selection of the Informative Frequency Band in a Bearing Fault Diagnosis in the Presence of Non-Gaussian Noise—Comparison of Recently Developed Methods". In: *Applied Sciences* (2020), p. 2657.

[68] Ronald Howard. *Dynamic probabilistic systems: Markov models*. Courier Corporation, 2012.

[69] Norden Huang, Zheng Shen, Steven Long, Manli Wu, Hsing Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry Liu. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis". In: *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences* (1998), pp. 903–995.

[70] Qiang Huang, Jun Li, and Hai Wen Le. "Application of the CEMS on Fault Diagnosis for Rotary Machine". In: *Applied Mechanics and Materials*. Trans Tech Publ. 2013, pp. 947–950.

[71] Fei Jiang, Kang Ding, Guolin He, and Canyi Du. "Sparse dictionary design based on edited cepstrum and its application in rolling bearing fault diagnosis". In: *Journal of Sound and Vibration* (2021), p. 115704.

[72] Michael Johnson. "Capacity and complexity of HMM duration modeling techniques". In: *IEEE signal processing letters* (2005), pp. 407–410.

[73] Biing-Hwang Juang. "On the hidden Markov model and dynamic time warping for speech recognition—A unified view". In: *AT&T Bell Laboratories Technical Journal* (1984), pp. 1213–1243.

[74] Biing Hwang Juang and Laurence R Rabiner. "Hidden Markov models for speech recognition". In: *Technometrics* (1991), pp. 251–272.

[75] Souhayb Kass, Amani Raad, and Jerome Antoni. "Self-running bearing diagnosis based on scalar indicator using fast order frequency spectral coherence". In: *Measurement* (2019), pp. 467–484.

[76] Akram Khaleghei and Viliam Makis. "Parameter and residual life estimation for a hidden semi-markov model of a deteriorating system". In: *J Multidiscip Eng Sci Technol (JMEST)* (2015), pp. 2532–2541.

[77] K Knill and S Young. "Hidden Markov models in speech and language processing". In: *Corpus-based methods in language and speech processing*. Springer, 1997, pp. 27–68.

[78] Amlan Kundu, George Chen, and Charles Persons. "Transient sonar signal classification using hidden Markov models and neural nets". In: *IEEE Journal of Oceanic Engineering* (1994), pp. 87–99.

[79] Chiman Kwan, Xiaodong Zhang, Roger Xu, and Leonard Haynes. "A novel approach to fault diagnostics and prognostics". In: *2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422)*. IEEE. 2003, pp. 604–609.

[80] Francesco Larizza, Carl Howard, and Steven Grainger. "Defect size estimation in rolling element bearings with angled leading and trailing edges". In: *Structural Health Monitoring* (2021), pp. 1102–1116.

[81] Francesco Larizza, Alireza Moazen-Ahmadi, Carl Howard, and Steven Grainger. "The importance of bearing stiffness and load when estimating the size of a defect in a rolling element bearing". In: *Structural Health Monitoring* (2019), pp. 1527–1542.

[82] Thanh Trung Le, Christophe Berenguer, and Florent Chatelain. "Multi-branch Hidden semi-Markov modeling for RUL prognosis". In: *2015 Annual Reliability and Maintainability Symposium (RAMS)*. IEEE. 2015, pp. 1–6.

[83] Thanh Trung Le, Christophe Berenguer, and Florent Chatelain. "Prognosis based on Multi-branch Hidden semi-Markov Models: A case study". In: *IFAC-PapersOnLine* (2015), pp. 91–96.

[84] Yaguo Lei, Zhengjia He, and Yanyang Zi. "Application of an intelligent classification method to mechanical fault diagnosis". In: *Expert Systems with Applications* (2009), pp. 9941–9948.

[85] Yaguo Lei, Jing Lin, Zhengjia He, and Yanyang Zi. "Application of an improved kurtogram method for fault diagnosis of rolling element bearings". In: *Mechanical Systems and Signal Processing* (2011), pp. 1738–1749.

[86] Yaguo Lei, Bin Yang, Xinwei Jiang, Feng Jia, Naipeng Li, and Asoke Nandi. "Applications of machine learning to machine fault diagnosis: A review and roadmap". In: *Mechanical Systems and Signal Processing* (2020), p. 106587.

[87] Norman Levinson. "The Wiener (root mean square) error criterion in filter design and prediction". In: *Journal of Mathematics and Physics* (1946), pp. 261–278.

[88] Bing Li, Pei-lin Zhang, Dong-sheng Liu, Shuang-shan Mi, Guo-quan Ren, and Hao Tian. "Feature extraction for rolling element bearing fault diagnosis utilizing generalized S transform and two-dimensional non-negative matrix factorization". In: *Journal of Sound and Vibration* (2011), pp. 2388–2399.

[89] Bing Li, Pei-lin Zhang, Hao Tian, Shuang-shan Mi, Dong-sheng Liu, and Guo-quan Ren. "A new feature extraction and selection scheme for hybrid fault diagnosis of gearbox". In: *Expert Systems with Applications* (2011), pp. 10000–10009.

[90] Xin Li, Jing Cai, Hongfu Zuo, and Huaiyuan Li. "Optimal cost-effective maintenance policy for a helicopter gearbox early fault detection under varying load". In: *Mathematical Problems in Engineering* (2017).

[91] Xin Li, Viliam Makis, Hongfu Zuo, and Jing Cai. "Optimal Bayesian control policy for gear shaft fault detection using hidden semi-Markov model". In: *Computers & Industrial Engineering* (2018), pp. 21–35.

[92] Lin Liang, Fei Liu, Maolin Li, Kangkang He, and Guanghua Xu. "Feature selection for machine fault diagnosis using clustering of non-negation matrix factorization". In: *Measurement* (2016), pp. 295–305.

[93] Jae Soo Lim and Alan V Oppenheim. "Enhancement and bandwidth compression of noisy speech". In: *Proceedings of the IEEE* (1979), pp. 1586–1604.

[94] Huo Lin, Fei Simiao, Lv Chuan, and Wang Zili. "A novel methodology based on hidden semi-Markov model for equipment health assessment". In: *Vibroengineering PROCEDIA* (2014), pp. 271–276.

[95] Qinming Liu and Ming Dong. "Online health management for complex nonlinear systems based on hidden semi-markov model using sequential monte carlo methods". In: *Mathematical Problems in Engineering* (2012).

[96] Qinming Liu, Ming Dong, Wenyuan Lv, Xiuli Geng, and Yupeng Li. "A novel method using adaptive hidden semi-Markov model for multi-sensor monitoring equipment health prognosis". In: *Mechanical Systems and Signal Processing* (2015), pp. 217–232.

[97] Qinming Liu, Ming Dong, and Ying Peng. "A novel method for online health prognosis of equipment based on hidden semi-Markov model using sequential Monte Carlo methods". In: *Mechanical Systems and Signal Processing* (2012), pp. 331–348.

[98] Tongshun Liu, Kunpeng Zhu, and Liangcai Zeng. "Diagnosis and prognosis of degradation process via hidden semi-Markov model". In: *IEEE/ASME Transactions on Mechatronics* (2018), pp. 1456–1466.

[99] Yi Liu, Zhansi Jiang, Huang Haizhou, and Jiawei Xiang. "Asymmetric penalty sparse model based cepstrum analysis for bearing fault detections". In: *Applied Acoustics* (2020), p. 107288.

[100] Zhiliang Liu, Yaqiang Jin, Ming J Zuo, and Zhipeng Feng. "Time-frequency representation based on robust local mean decomposition for multicomponent AM-FM signal analysis". In: *Mechanical systems and signal processing* (2017), pp. 468–487.

[101] Zhiliang Liu, Ming Zuo, Yaqiang Jin, Deng Pan, and Yong Qin. "Improved local mean decomposition for modulation information mining and its application to machinery fault diagnosis". In: *Journal of Sound and Vibration* (2017), pp. 266–281.

[102] Nadine Martin, Corinne Mailhes, and Xavier Laval. "Automated machine health monitoring at an expert level". In: *Acoustics Australia* (2021), pp. 1–13.

[103] Robert McAulay and Marilyn Malpass. "Speech enhancement using a soft-decision noise suppression filter". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1980), pp. 137–145.

[104] Andrew McCormick and Asoke Nandi. "Cyclostationarity in rotating machine vibrations". In: *Mechanical systems and signal processing* (1998), pp. 225–242.

[105] Geoff McDonald, Qing Zhao, and Ming Zuo. "Maximum correlated Kurtosis deconvolution and application on gear tooth chip fault detection". In: *Mechanical Systems and Signal Processing* (2012), pp. 237–255.

[106] P.D. McFadden and J.D. Smith. "Model for the vibration produced by a single point defect in a rolling element bearing". In: *Journal of sound and vibration* (1984), pp. 69–82.

[107] Ramin Moghaddass and Ming Zuo. "An integrated framework for online diagnostic and prognostic health monitoring using a multistate deterioration process". In: *Reliability Engineering & System Safety* (2014), pp. 92–104.

[108] Subhasis Nandi, Hamid A Toliyat, and Xiaodong Li. "Condition monitoring and fault diagnosis of electrical motors—A review". In: *IEEE transactions on energy conversion* (2005), pp. 719–729.

[109] Willem Niehaus, Stephan Schmidt, and Stephan Heyns. "NIC Methodology: A probabilistic methodology for improved informative frequency band identification

by utilizing the available healthy historical data under time-varying operating conditions". In: *Journal of Sound and Vibration* (2020), p. 115642.

[110] Hasan Ocak. *Fault detection, diagnosis and prognosis of rolling element bearings: frequency domain methods and hidden markov modeling*. Case Western Reserve University, 2004.

[111] Hasan Ocak and Kenneth Loparo. "A new bearing fault detection and diagnosis scheme based on hidden Markov modeling of vibration signals". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. IEEE. 2001, pp. 3141–3144.

[112] Hasan Ocak and Kenneth Loparo. "HMM-based fault detection and diagnosis scheme for rolling element bearings". In: (2005).

[113] Henry Ogbemudia Omoregbee and Stephan Heyns. "Fault detection in roller bearing operating at low speed and varying loads using Bayesian robust new hidden Markov model". In: *Journal of Mechanical Science and Technology* (2018), pp. 4025–4036.

[114] Aziz Kubilay Ovacikli, Patrik Paajarvi, James LeBlanc, and Johan Carlson. "Uncovering harmonic content via skewness maximization-a Fourier analysis". In: *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE. 2014, pp. 481–485.

[115] Choon-Su Park, Young-Chul Choi, and Yang-Hann Kim. "Early fault detection in automotive ball bearings using the minimum variance cepstrum". In: *Mechanical Systems and Signal Processing* (2013), pp. 534–548.

[116] William Pearlman and Robert Gray. "Source coding of the discrete Fourier transform". In: *IEEE Transactions on information theory* (1978), pp. 683–692.

[117] Ying Peng and Ming Dong. "A prognosis method using age-dependent hidden semi-Markov model for equipment health prediction". In: *Mechanical Systems and Signal Processing* (2011), pp. 237–252.

[118] J. Porter and S. Boll. "Optimal estimators for spectral restoration of noisy speech". In: *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE. 1984, pp. 53–56.

[119] Barry Quinn, Ross Barrett, and Stephen Searle. "The estimation and HMM tracking of weak narrowband signals". In: *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 1994, pp. IV–341.

[120] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* (1989), pp. 257–286.

[121] Lawrence R Rabiner, Stephen E Levinson, and M Mohan Sondhi. "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition". In: *Bell System Technical Journal* (1983), pp. 1075–1105.

[122] Robert Randall. "A history of cepstrum analysis and its application to mechanical problems". In: *Mechanical Systems and Signal Processing* (2017), pp. 3–19.

[123] Robert Randall and Nader Sawalhi. "Signal processing tools for tracking the size of a spall in a rolling element bearing". In: *IUTAM symposium on emerging trends in rotor dynamics*. Springer. 2011, pp. 429–440.

[124] Robert B. Randall and Jerome Antoni. "Rolling element bearing diagnostics—A tutorial". In: *Mechanical systems and signal processing* (2011), pp. 485–520.

[125] Andrew P Reilly and Boualem Boashash. "Comparison of time-frequency signal analysis techniques with application to speech recognition". In: *Advanced Signal Processing Algorithms, Architectures, and Implementations III*. International Society for Optics and Photonics. 1992, pp. 339–350.

[126] Martin Russell. "A segmental HMM for speech pattern modelling". In: *1993 IEEE international conference on acoustics, speech, and signal processing*. IEEE. 1993, pp. 499–502.

[127] Nader Sawalhi and Robert Randall. "Simulating gear and bearing interactions in the presence of faults: Part I. The combined gear bearing dynamic model and the simulation of localised bearing faults". In: *Mechanical Systems and Signal Processing* (2008), pp. 1924–1951.

[128] Nader Sawalhi and Robert Randall. "Simulating gear and bearing interactions in the presence of faults: Part II: Simulation of the vibrations produced by extended bearing faults". In: *Mechanical Systems and Signal Processing* (2008), pp. 1952–1966.

[129] Nader Sawalhi and Robert Randall. "Vibration response of spalled rolling element bearings: Observations, simulations and signal processing techniques to track the spall size". In: *Mechanical Systems and Signal Processing* (2011), pp. 846–870.

[130] Nader Sawalhi, Robert Randall, and Hiroaki Endo. "The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis". In: *Mechanical Systems and Signal Processing* (2007), pp. 2616–2633.

[131] Nader Sawalhi, Wenyi Wang, and Andrew Becker. "Vibration signal processing for spall size estimation in rolling element bearings using autoregressive inverse filtration combined with bearing signal synchronous averaging". In: *Advances in Mechanical Engineering* (2017), p. 1687814017703007.

[132] Miloud Sedira and Ahmed Felkaoui. "Rotating machinery Diagnostic Using Hidden Markov Models (HMMs)". In: ().

[133] Amit Shrivastava and Sulochana Wadhwani. "Development of fault detection system for ball bearing of induction motor using vibration signal". In: *International Journal Of Scientific Research* (2013).

[134] Fatima Sloukia, Mohamed Aroussi, Hicham Medromi, and Mohamed Wahbi. "Bearings prognostic using mixture of gaussians hidden markov model and support vector machine". In: *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2013, pp. 1–4.

[135] Jonathan Smith. "The local mean decomposition and its application to EEG perception data". In: *Journal of the Royal Society Interface* (2005), pp. 443–454.

[136] Wade Smith, Zhiqi Fan, Zhongxiao Peng, Huaizhong Li, and Robert Randall. "Optimised Spectral Kurtosis for bearing diagnostics under electromagnetic interference". In: *Mechanical Systems and Signal Processing* (2016), pp. 371–394.

[137] Wade Smith and Robert Randall. "Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study". In: *Mechanical Systems and Signal Processing* (2015), pp. 100–131.

[138] Panu Somervuo, Aki Harma, and Seppo Fagerlund. "Parametric representations of bird sounds for automatic species recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* (2006), pp. 2252–2263.

[139] Cesar Ricardo Soto Ocampo, Jose Manuel Mera, Juan David Cano-Moreno, and Jose Luis Garcia-Bernardo. "Low-Cost, High-Frequency, Data Acquisition System for Condition Monitoring of Rotating Machinery through Vibration Analysis-Case Study". In: *Sensors* (2020), p. 3493.

[140] César Ricardo Soto-Ocampo, José Manuel Mera, Juan David Cano-Moreno, and José Luis Garcia-Bernardo. *Bearing Database*. Version V1. https://doi.org/10.5281/zenodo.3898942. June 2020.

[141] Roy Streit and Ross Barrett. "Frequency line tracking using hidden Markov models". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1990), pp. 586–598.

[142] Chun Su and Jinyun Shen. "A novel multi-hidden semi-Markov model for degradation state identification and remaining useful life estimation". In: *Quality and Reliability Engineering International* (2013), pp. 1181–1192.

[143] Kevin Tang, Li Fei-Fei, and Daphne Koller. "Learning latent temporal structure for complex event detection". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 1250–1257.

[144] Hongzhi Teng, Jianmin Zhao, Xisheng Jia, Yunxian Jia, Xinghui Zhang, and Liying Cai. "Experimental study on gearbox prognosis using total life vibration analysis". In: *2011 Prognostics and System Health Managment Confernece*. IEEE. 2011, pp. 1–6.

[145] Diego Tobon-Mejia, Kamal Medjaher, Noureddine Zerhouni, and Gerard Tripot. "A mixture of gaussians hidden markov model for failure diagnostic and prognostic". In: *2010 IEEE International Conference on Automation Science and Engineering*. IEEE. 2010, pp. 338–343.

[146] Diego Tobon-Mejia, Kamal Medjaher, Noureddine Zerhouni, and Gerard Tripot. "Hidden Markov models for failure diagnostic and prognostic". In: *2011 Prognostics and System Health Managment Confernece*. IEEE. 2011, pp. 1–8.

[147] Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*. IEEE. 1999, pp. 229–232.

[148] Shiu Kit Tso and King Pui Liu. "Demonstrated trajectory selection by hidden Markov model". In: *Proceedings of International Conference on Robotics and Automation*. IEEE. 1997, pp. 2713–2718.

[149] Manuel Vieira, Paulo Fonseca, Clara Amorim, and Carlos Teixeira. "Call recognition and individual identification of fish vocalizations based on automatic speech recognition: An example with the Lusitanian toadfish". In: *The Journal of the Acoustical Society of America* (2015), pp. 3941–3950.

[150] Biao Wang, Yaguo Lei, Naipeng Li, and Ningbo Li. "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings". In: *IEEE Transactions on Reliability* (2018), pp. 401–412.

[151] Dong Wang. "K-nearest neighbors based methods for identification of different gear crack levels under different motor speeds and loads: Revisited". In: *Mechanical Systems and Signal Processing* (2016), pp. 201–208.

[152] Huaqing Wang, Mengyang Wang, Junlin Li, Liuyang Song, and Yansong Hao. "A Novel Signal Separation Method Based on Improved Sparse Non-Negative Matrix Factorization". In: *Entropy* (2019), p. 445.

[153] Ning Wang, Shu-dong Sun, Zhi-qiang Cai, Shuai Zhang, and Can Saygin. "A hidden semi-markov model with duration-dependent state transition probabilities for prognostics". In: *Mathematical Problems in Engineering* (2014).

[154] Wenyi Wang, Nader Sawalhi, and Andrew Becker. "Size estimation for naturally occurring bearing faults using synchronous averaging of vibration signals". In: *Journal of Vibration and Acoustics* (2016).

[155] Jacek Wodecki, Anna Michalak, Radoslaw Zimroz, Tomasz Barszcz, and Agnieszka Wylomanska. "Impulsive source separation using combination of Nonnegative Matrix Factorization of bi-frequency map, spatial denoising and Monte Carlo simulation". In: *Mechanical Systems and Signal Processing* (2019), pp. 89–101.

[156] Haixi Wu, Zhonghua Yu, and Yan Wang. "Real-time FDM machine condition monitoring and diagnosis based on acoustic emission and hidden semi-Markov model". In: *The International Journal of Advanced Manufacturing Technology* (2017), pp. 2027–2036.

[157] Meng Wu, Xiaoxiao Dai, Yimin Zhang, Bradley Davidson, Moeness Amin, and Jun Zhang. "Fall detection based on sequential modeling of radar signal time-frequency features". In: *2013 IEEE International Conference on Healthcare Informatics*. IEEE. 2013, pp. 169–174.

[158] Xin Wu, Yaoyu Li, and Wei Teng. "Modified hidden semi-Markov models for motor wear prognosis". In: *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology* (2012), pp. 174–179.

[159] Qili Xiao, Yilin Fang, Quan Liu, and Shujuan Zhou. "Online machine health prognostics based on modified duration-dependent hidden semi-Markov model and

high-order particle filtering". In: *The International Journal of Advanced Manufacturing Technology* (2018), pp. 1283–1297.

[160] Wenbo Xiao, Jun Chen, Guangming Dong, Yu Zhou, and Zhiyang Wang. "A multichannel fusion approach based on coupled hidden Markov models for rolling element bearing fault diagnosis". In: *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* (2012), pp. 202–216.

[161] Ge Xin, Jerome Antoni, and Nacer Hamzaoui. "An exploring study of hidden Markov model in rolling element bearing diagnostis". In: *Surveillance 8*. 2015.

[162] Ge Xin, Nacer Hamzaoui, and Jerome Antoni. "Semi-automated diagnosis of bearing faults based on a hidden Markov model of the vibration signals". In: *Measurement* (2018), pp. 141–166.

[163] Jie Yang, Yangsheng Xu, and Chiou S Chen. "Human action learning via hidden Markov model". In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* (1997), pp. 34–44.

[164] Yong-sheng Yang, An-bo Ming, You-yun Zhang, and Yong-sheng Zhu. "Discriminative non-negative matrix factorization (DNMF) and its application to the fault diagnosis of diesel engine". In: *Mechanical Systems and Signal Processing* (2017), pp. 158–171.

[165] Yongsheng Yang, Youyun Zhang, and Yongsheng Zhu. "Application of the multi-kernel non-negative matrix factorization on the mechanical fault diagnosis". In: *Advances in Mechanical Engineering* (2015), p. 1687814015584494.

[166] Shun-Zheng Yu. "Hidden semi-Markov models". In: *Artificial intelligence* (2010), pp. 215–243.

[167] Bin Zhang, Chris Sconyers, Carl Byington, Romano Patrick, Marcos Orchard, and George Vachtsevanos. "Anomaly detection: A robust approach to detection of unanticipated faults". In: *2008 International Conference on Prognostics and Health Management*. IEEE. 2008, pp. 1–8.

[168] Bin Zhang, Chris Sconyers, Carl Byington, Romano Patrick, Marcos E Orchard, and George Vachtsevanos. "A probabilistic fault detection approach: Application to bearing fault detection". In: *IEEE Transactions on Industrial Electronics* (2010), pp. 2011–2018.

[169] Xiaodong Zhang, Roger Xu, Chiman Kwan, Qiulin Liang Stevenand Xie, and Leonard Haynes. "An integrated approach to bearing fault diagnostics and prog-

nostics". In: *Proceedings of the 2005, American Control Conference, 2005.* IEEE. 2005, pp. 2750–2755.

[170] XiaoLi Zhang, BaoJian Wang, and XueFeng Chen. "Intelligent fault diagnosis of roller bearings with multivariable ensemble-based incremental support vector machine". In: *Knowledge-Based Systems* (2015), pp. 56–85.

FOLIO ADMINISTRATIF

THESE DE L'UNIVERSITE DE LYON OPEREE AU SEIN DE L'INSA LYON

NOM : JIN DATE de SOUTENANCE : 25/05/2022

Prénoms : Yaqiang

TITRE : Integrated auto-diagnosis based on stochastic model for rolling element bearings

NATURE : Doctorat Numéro d'ordre : 2022LYSEI045

Ecole doctorale : Mécanique, énergétique, Génie Civil, Acoustique (MEGA)

Spécialité : Génie Mécanique

RESUME : Aujourd'hui, le problème le plus fondamental de la surveillance d'état dans la plupart des installations industrielles est le diagnostic et le pronostic des défauts. L'une des approches les plus efficaces pour étudier ce problème est la surveillance de l'état basée sur l'analyse des signaux de vibration. Avec le développement de l'industrie, la maintenance multithread et l'acquisition multicanal se généralisent, ce qui met en avant des exigences de maintenance plus élevées. Sur la base de cette observation, il est proposé dans cette thèse un cadre de diagnostic automatisé pour le roulement qui intègre les étapes successives de détection de défaut, d'identification du type de défaut, de reconstruction du signal de défaut et de caractérisation de la taille du défaut. L'avantage est que le processus de diagnostic complet est réalisé en une seule fois, tout en impliquant un seul hyperparamètre clé, ce qui améliore le degré d'automatisation de la maintenance basée sur les conditions (CBM) actuelle et libère la participation humaine.

En présence de défaut naissant, les vibrations des roulements présentent des signatures symptomatiques sous forme d'impulsions répétitives. Cela peut être vu comme un signal non stationnaire dont les propriétés statistiques basculent entre deux états. La stratégie de maintenance proposée modélise ces caractéristiques avec un modèle de Markov caché à durée explicite (EDHMM) et utilise les paramètres estimés du modèle pour effectuer un diagnostic intégré sans nécessiter l'expertise de l'utilisateur. La détection d'un défaut est d'abord réalisée au moyen d'un test de rapport de vraisemblance construit sur les paramètres de l'EDHMM. Une approche de comptage statistique et une probabilité sont ensuite utilisées pour identifier automatiquement le type de défaut. Afin d'obtenir le signal de défaut dans certains cas, un filtre bayésien basé sur les paramètres EDHMM est construit. Enfin, la taille du défaut est estimée à partir des durées renvoyés par EDHMM.

Par la suite, la capacité du cadre d'autodiagnostic intégré est illustrée sur différents ensembles de données expérimentales. La première validation est réalisée sur les données de vibration pour des conditions spécifiques. Les résultats indiquent une diagnostic robuste et précise du roulement à éléments roulants. De plus, le résultat sur des données de dégradation accélérée montre également l'efficacité de la méthode, en particulier la capacité à détecter l'occurrence d'une défaillance et de suivre quantitativement son développement. Cette technique a un potentiel d'utilisation en CBM.

MOTS-CLÉS : maintenance du renseignement; Roulement à élément roulant; Auto-framework intégré; Détection de fautes; Identification du type de défaut; Reconstruction du signal de défaut; Caractérisation de la taille des défauts.

Laboratoire (s) de recherche : Laboratoire de vibration et acoustique (LVA)

Directeur de thèse: Jérôme Antoni

Président de jury :

Composition du jury : GRYLLIAS, Konstantinos (rapporteur)
WYLOMANSKA, Agnieszka (rapporteur)
EL BADAOUI, Mohamed (examinateur)
ANTONI, Jérôme (directeur de thèse)