The University Of Sheffield.

Frequency Analysis of Linear and Nonlinear Systems for Applications in Fault Detection and Medical Diagnosis

By

Sikai Zhang

Supervised by:

Prof Zi-Qiang Lang

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

University of Sheffield

Faculty of Engineering

Department of Automatic Control and Systems Engineering

October 2019

Abstract

The frequency analysis is highly demanded to process the industrial or medical data for fault detection and diagnosis, especially when the investigated machine or human tissues are stimulated by the periodic signals. The PhD research work aims to develop the new methods for system frequency feature extraction and selection of features for machine learning oriented classification and to apply these methods in the fault detection and medical diagnosis. To analyse the system characteristics with input-output data, novel modelling and model frequency feature extraction method is proposed. The method is effective in revealing the physically meaningful characteristics of systems. To select the useful features for machine learning oriented classification, an orthogonal least squares based feature selection method is proposed. Compared to traditional methods, the proposed method has a faster computation speed and higher accuracy. These novel methods are then applied to two real-world problems, which are wind turbine fault detection and preterm birth prediction. In the wind turbine fault detection application, the results show that the modelling and model feature extraction method is powerful in the damage sensitive feature extraction, while traditional methods do not work well. In the preterm birth prediction, the proposed methods can extract and select features from the magnetic impedance spectroscopy data of the pregnant women's cervix tissue. The results demonstrate that the magnetic impedance spectroscopy data have the capability to predict the spontaneous preterm birth. These application studies demonstrate that the proposed methods have great potential to be used in many engineering system fault detection and medical diagnosis related applications.

Acknowledgements

I am extremely grateful to my supervisor Professor Zi-Qiang Lang for the support of the academic training. During the four years of my PhD, Prof Lang shows a great patient to me. I am benefited from his precious advice especially in academic writing. Prof Lang also provides me with part-time works to help me financially. I would like to thank my colleagues, Yunpeng Zhu, Rajintha Gunawardena, Ke Sun, and Jinny Robson, for their advice on my research. I would also like to give thanks to my brothers and sisters in Christ for their prayers. Finally, I would like to give thanks to my parents for their selfless love.

For the data used in Chapter 2, acknowledgement is made for the measurements used in this work provided through data-acoustics.com Database.

Contents

List of Figures

List of Figures

List of Tables

List of Algorithms

List of Acronyms

AC: Alternating Current

AE: Acoustic Emission

ARX: AutoRegressive eXogenous

CCA: Canonical Correlation Analysis

CIFE: Conditional Infomax Feature Extraction

CL: Cervical Length

CMIM:Conditional Mutual Information Maximisation

DFT: Discrete Fourier Transform

DISR: Double Input Symmetrical Relevance

EIS: Electrochemical Impedance Spectroscopy

ERR: Error Reduction Ratio

ESR: Error-to-Signal Ratio

fFN: fetal FibronectiN

FPGA: Field-Programmable Gate Array

FRF: Frequency Response Function

FROLS: Forward Regression with Orthogonal Least-Squares

GFRFs: Generalised Frequency Response Functions

ICAP: Interaction Capping

IDFT: Inverse Discrete Fourier Transform

iOFR: iterative Orthogonal Forward Regression

JMI: Joint Mutual Information

LDA: Linear Discriminant Analysis

MIM: Mutual Information Maximisation

MIS: Magnetic Impedance Spectroscopy

mRMR: minimal-Redundancy-Maximal-Relevance

NARX: Nonlinear AutoRegressive eXogenous

NOFRFs: Nonlinear Output Frequency Response Functions

O&M: Operation and Maintenance

OBCC: Orthogonal point-Biserial Correlation Coefficient

OLS: Orthogonal Least-Squares

PCA: Principal Component Analysis

SCADA: Supervisory Control And Data Acquisition

SERR: Sum of Error Reduction Ratio

sPTB: spontaneous Pre-Term Birth

SSR: Residual Sum of Squares

SST: Total Sum of Squares

TSA: Time Synchronous Averaging

TSOLS: Two-Stage Orthogonal Least-Squares

List of Notation

a: the crack depth

a_m and $A_m(j\omega)$: the input-output transmission path in mode m and its spectrum

c: the flexibility of the uncracked beam

δc: the local flexibility caused by the crack

C: the flexibility of the cracked beam

$C_p(\tau)$: the power cepstrum

d: the diameter of the rolling element

$d(t)$ and $D(j\omega)$: a unit impulse train and its spectrum

D: the pitch diameter

$D(\cdot, \cdot)$: the relevance between a feature and a target

$\mathbf{e}$: the error vector

$e(k)$: the k^{th} data point in the additive noise sequence, $e(k) \in \mathbb{R}$

f_c: the characteristic defect frequency

f_r: the resonance frequency

f_s: the sampling frequency in Hz

$\mathcal{F}$: the discrete Fourier transform

$\mathcal{F}^{-1}$: the inverse discrete Fourier transform

G: the likelihood-ratio test statistic

G_g: the thermal capacitance

$G_n(e^{j\omega T})$: the n^{th} order NOFRF at frequency ω

h: the height of the rectangular cross-section of the beam

$h_n(\tau_1, \ldots, \tau_n)$: the n^{th} order Volterra kernel, $h_n \in \mathbb{R}$

$H(e^{j\omega T})$: the frequency response function

$I(e^{j\omega T})$: the spectrum of the current

$I(\cdot;\cdot)$: the mutual information

$\mathcal{I}$: the information matrix

$\mathbf{J}$: the Jacobian matrix

k: the index of sequence, $k \in \mathbb{Z}$

ℓ: the order of NARX model, $\ell \in \mathbb{Z}^+$

$L(\cdot)$: the log-likelihood function

m: the mass of the beam

m_c: the current number of candidate features, $m_c \in \mathbb{Z}^+$

m_s: the current number of selected features, $m_s \in \mathbb{Z}^+$

m_{uc}: the number of the unchanged terms in TSOLS process , $m_{uc} \in \mathbb{Z}^+$

M: the number of total features or total terms of polynomial NARX model, $M \in \mathbb{Z}^+$

M_s: the number of features is needed to select, $M_s \in \mathbb{Z}^+$

n_u: is the maximum lag for u, $n_u \in \mathbb{Z}^+$

n_y: the maximum lag for y, $n_y \in \mathbb{Z}^+$

N: the order of system nonlinearity, $N \in \mathbb{Z}^+$

N_s: the number of observation, $N_s \in \mathbb{Z}^+$

$\mathcal{O}$: the asymptotic upper bound notation

$P(\cdot, \cdot)$: the joint probability distribution

q: the number of all possible distinct term of the NARX model

$r(\cdot, \cdot)$: the Pearson correlation coefficient

$r^b(\cdot, \cdot)$: the point-biserial correlation coefficient

$R(\cdot)$: the redundancy of the features

R_{ga}: the thermal resistance between the external environment and the generator

$R_{uy}(\tau)$: the cross-correlation function between u and y

$S_{uy}(\tau)$: the cross spectral density between u and y

t: the time in second

T: the sampling period in second

T_a: the ambient temperature

T_d: the reciprocal of the inner race element passing frequency

T_g: the wind turbine generator winding temperature

$u(k)$: the k^{th} data point in the input sequence, $u(k) \in \mathbb{R}$

$U(e^{j\omega T})$: the spectrum of $u(k)$

$V(e^{j\omega T})$: the spectrum of the voltage

V_w: the wind speed

$\mathbf{w}$: the orthogonalised vector of $\mathbf{x}$

$w(t)$ and $W(j\omega)$: a weighting function indicating the contact energy in bearing and its spectrum

W: the Wald test statistic

W_s: the set composed of the selected orthogonalised regressors or features

$\mathbf{x}$: a term in the NARX model or a feature

$\mathbf{X}$: the regressor matrix of the NARX model

$\mathbf{X_r}$: the matrix composed of rest candidate regressors or features

X_s: the set composed of the selected features

$\mathbf{X_s}$: the matrix composed of selected regressors or features

$\mathbf{y}$: the output vector

$y(k)$: the k^{th} data point in the output sequence, $y(k) \in \mathbb{R}$

$Y(e^{j\omega T})$: the spectrum of $y(k)$

Z: the number of rolling elements

$Z(e^{j\omega T})$: the impedance spectroscopy

α: a constant factor for NOFRFs computation, $\alpha \in \mathbb{R}^+$

α_m: the system characteristics in mode m

θ: the parameter of the NARX model

$\mathbf{\Theta}$: the parameter vector of the NARX model

ϕ: the contact angle

ω_{bl}: the ball spinning frequency

ω_{cg}: the cage frequency

ω_{id}: the inner race defect frequency

ω_m: the system characteristics in mode m

ω_n: the nature frequency of the normal beam

ω_{od}: the outer race defect frequency

ω_{rc}: the rolling element defect frequency

ω_s: the shaft rotation frequency in rad/s

Chapter 1

Introduction

1.1 Background and motivation

For economic, safety, and health considerations, maintenance is necessary in many systems. The traditional maintenance technique is the reactive maintenance, which takes place at breakdowns. A better maintenance technique is the preventative maintenance, in which the system is maintained based on the time or usage. However, the cost of the preventive maintenance is usually high [1]; in addition, the preventative maintenance cannot be carried out in the some areas, such as health care, without the insight of the system's health conditions. To this end, the condition-based maintenance has been proposed. The condition-based maintenance requires the technologies to quantify the health conditions of a system. Based on the health conditions, the specific and targeted maintenance can be applied. Therefore, the success of the condition-based maintenance highly relies on technologies those can accurately evaluate the health condition of systems.

The frequency analysis has been widely used in many fields in the fault detection and diagnosis, especially after the fast Fourier transform (FFT) was developed by Cooley and Tukey [2]. The vibration and acoustic emission signals are frequently used in frequency analysis for the fault detection of the mechanical elements or systems, such as bearings, gearboxes, and bridges. The the electronic signals, such as current and impedance spectroscopy, are popular in the frequency analysis for the fault detection of motors or medical applications such as cancer diagnosis. This PhD project is motivated by the need to develop the methods for

the wind turbines generator fault detection and the preterm birth prediction for the pregnant women using electronic impedance spectroscopy data.

For the wind turbine generator fault detection, the traditional frequency analysis usually requires expensive additional equipments. In fact, most MW-scale wind turbines have installed supervisory control and data acquisition (SCADA) systems, which collect important signals from the wind turbines for fault detection, such as the generator temperature, tower vibration, and power output. However, as the SCADA data are recorded every 10 minutes, whose sampling frequency is too low to capture the most of the dynamics of wind turbines, the researchers mainly focus on the static relationship between the SCADA parameters, and these relationships cannot be used to conduct frequency analysis. For example, the power curve which describes the static relationship between the wind speed and the power output is applied for the wind turbine fault detection [3]. The static relationships contain less information than the dynamic relationship. In our case, the static relationship cannot detect the generator fault at all. Fortunately, it is found that the generator temperature changes periodically at a lower frequency (i.e. one day period), so the 10 minutes sampling period is enough to capture the dynamic characteristics of the generator temperature. Therefore, in Chapter 6, the dynamic relationship between the generator temperature, wind speed, and the ambient temperature has been studied, and the frequency analysis has been carried out for the wind turbine generator fault detection.

For the preterm birth prediction, the traditional methods adopt two risk factors, i.e. cervix length and fetal fibronectin (fFN), to predict the risk of preterm birth, but the accuracy of the prediction is low [4]. The recent research has identified cumulative changes in hydration as a sign of the ripening process which precedes labour [5]. The changes in hydration can be assessed by the cervix electrical impedance, which is a type of frequency features. The challenge to use the impedance features is the number of features is large, and the features are highly redundant. In Chapter 5, the orthogonal least squares based feature selection method with efficient redundancy control has been applied to select useful impedance features, and it has been investigated whether the impedance features are feasible for the preterm birth prediction.

1.2 Key contributions and outline

The relationship between each chapter is shown in Figure 1.1. Two novel methods are developed in Chapter 4 and Chapter 6, respectively, which are followed by the two applications in Chapter 5 and Chapter 7.

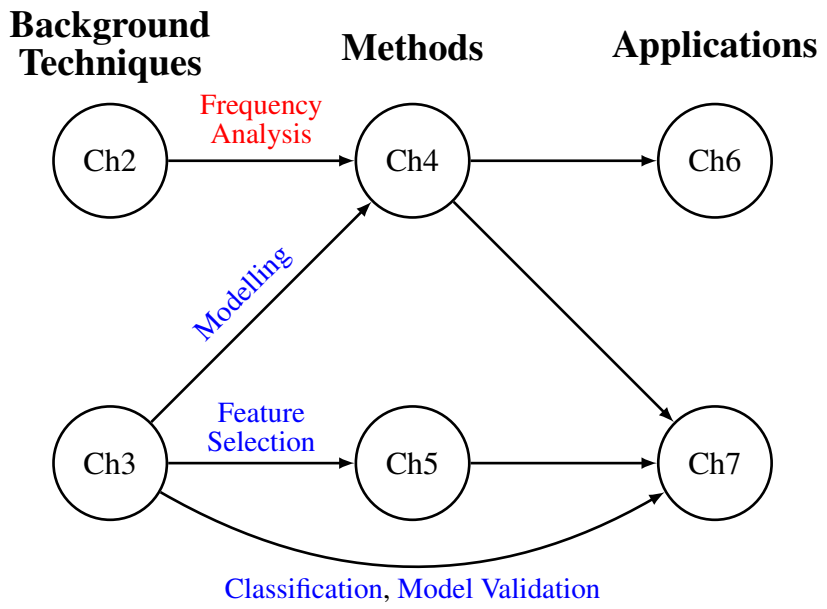


Figure 1.1: The relationship between different chapters in the thesis.

The overview of each chapter and the main contributions are as follows:

- **Chapter 2:**

  This chapter reviews the existing methods for the frequency analysis and their application in fault detection and diagnosis. The methods are divided into signal based and model based. In the model based frequency analysis, the review will focus on the frequency responses function (FRF) analysis and the nonlinear output frequency response functions (NOFRFs) analysis. The applications of frequency analysis in the bearing defects, gear defects, beam cracks, and cancer detection are reviewed. The most common methods in each field are presented.

- **Chapter 3:**

  This chapter introduces a range of topics covering modelling, feature selection, classification, and model validation. The preliminary knowledge is reviewed. As the model

structure adopted in the thesis is the nonlinear autoregressive with exogenous inputs (NARX) model, the modelling techniques for NARX models are reviewed. The idea of filter feature selection is introduced. Two feature selection methods with different redundancy control techniques are reviewed. For the classification, the review focuses on the logistic regression model. Three performance evaluation methods for logistic regression model are reviewed. As the data for the preterm birth medical diagnosis are imbalanced and small in size, the special cross-validation techniques are required. Four typical cross-validation techniques have been introduced.

- **Chapter 4:**

  In this chapter, the NARX modelling techniques are first applied to a linear system with a signal equilibrium and then to a nonlinear system with dual equilibria. The frequency features (i.e. FRF and NOFRFs features) of the system are then extracted from the system models. The simulation results demonstrate the effectiveness of modelling and model feature extraction method in monitoring the changes of the system characteristics.

- **Chapter 6:**

  In Chapter 6, the modelling and model feature extraction method proposed in Chapter 4 is applied in the wind turbine generator fault detection. As the model performs as a sensor, which keeps updating to provide real-time features of the generator, the model is called the model sensor of the wind turbine generator, and the fault detection method based on the model sensor is referred to as the model sensor method. The SCADA data of three wind turbines are used for this study. The NOFRFs features are used as the damage sensitive indicator and extracted using the model sensor method. The effectiveness of the model sensor method and the NOFRFs features is verified by real data analysis. The main contributions in this chapter are given as follows:

  - the structure of the model describing the dynamic relationship between the generator temperature, wind speed, and ambient temperature is determined from the first principles;
  - the model sensor method is applied to update the parameter of the dynamic model

monthly;

- the NOFRFs feature extraction under different operating point is proposed and applied;
- the strategy about how to deploy model sensors on the wind turbines for the early fault alarm is developed;
- the fault detection performance of the linear and nonlinear model sensor is compared.

- **Chapter 5:**

  In this chapter, an orthogonal least squares (OLS) based feature selection method is proposed for both binomial and multinomial classification. The squared orthogonal correlation coefficients are defined analysed as the feature ranking criterion. The equivalence between the canonical correlation coefficient, Fisher's criterion, and the sum of the squared orthogonal correlation coefficients are proved to demonstrate the statistical implication of the proposed method. It is also shown that the OLS based feature selection method has speed advantages in a greedy search. The simulation tests of continuous feature selection are designed to compare the proposed method with other popular feature selection methods, and the proposed method achieves the best results. The main contributions in this chapter are given as follows:

  - The new statistics based on *squared orthogonal correlation coefficients* are defined.
  - The equivalence between the squared orthogonal correlation coefficients, the canonical correlation coefficient, and Fisher's criterion are proved.
  - The speed advantage of OLS based method in a greedy search is analysed by evaluating computational complexity.
  - The OLS based feature selection algorithm for both binomial and multinomial classification cases are developed.

- **Chapter 7:**

  In this chapter, the modelling and model analysis method and the OLS based feature

selection methods are applied to preterm birth prediction. The data used for analysis is the magnetic impedance spectroscopy (MIS). The MIS is measured in a non-contact manner, and the time domain data are provided for this study. The traditional MIS analysis is based on the frequency domain measurement. The modelling and model analysis method is to extract the impedance spectra from the time series models which are trained with the time domain data. The features are selected from the impedance spectra by the OLS based feature selection method. The results demonstrated that the effectiveness of the MIS and the proposed methods in sPTB prediction. The main contributions in this chapter are given as follows:

- the new idea of extracting MIS from time series model is proposed;
- the traditional features are compared with the model based features;
- the calibrated features are compared with the non-calibrated features;
- the efficient MIS features are identified in the study that can potentially be applied in future studies.

1.3 List of publications

Journal

- S. Zhang, Z.Q. Lang, Y.P. Zhu, "SCADA-data-based wind turbine fault detection: a dynamic model sensor method," in preparation for submission to *Control Engineering Practice*.
- S. Zhang, Z.Q. Lang, "Orthogonal least-squares based fast feature selection for classification," submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Conference

- S. Zhang, Y.P. Zhu, L. Zhao, Z.Q. Lang, Victoria Stern, Jamie Healey, Brian H Brown, Andrew Humphries, Dilly O Anumba, "Nonlinear Logistic Regression Analysis of Electrical Impedance Spectroscopy Data Improves the Accuracy of Prediction of Preterm Birth," in Reproductive Sciences, San Diego, 2018.

Workshop

- S. Zhang, J. Lin, Q. Lin, Z.Q. Lang, "A data-driven method for the detection of fly-off faults of large rotating machine rotor parts," the 30th International Workshop on Principles of Diagnosis (DX-19), 2019.

Chapter 2

Frequency Features of Signals and Systems and Application in Fault Detection

2.1 Introduction

Frequency analysis can extract useful features from signals and systems. The fault detection and diagnosis can be realised using the features by simply setting a threshold or applying more complicated machine learning methods [6]. Most systems investigated by frequency analysis are the machines or the elements working in a periodic environment. For example, the gearboxes or the bearings in a rotating machine, and the circuit under the alternating current. The signals sampled from such systems have periods with certain frequencies. Naturally, the characteristics of the signals in the frequency domain are extracted to represent the system operating conditions. Actually, frequency analysis can also be used in the areas without periodic behaviours. In these cases, the frequency features are regarded as the synthesis characteristics of the systems. For example, the natural frequency of a beam is a common feature for crack detection, where the beams are not necessary to work in the periodic behaviour [7].

From a system's perspective, the signals can be divided into system inputs and system outputs. The input signals are the external environmental conditions, such as load and ambi-

ent temperature, which are independent from the system. The output signals are the physical variables affected by the system operation, such as power output of a generator and vibration of gearbox, which are dependent on the characteristics of the system. If the frequency features are extracted from outputs only without considering the effect of the system inputs, the frequency analysis is signal based. The signal based frequency analysis has spectrum analysis, power spectrum analysis, cepstrum analysis, etc. The advantage of the signal based frequency analysis is its simplicity. However, as the system outputs are influenced not only by the system characteristics but also system inputs, signal based methods are not sensitive to small change of the system characteristics [8]. To remove the effect of inputs, the relationship between the inputs and the outputs can be found first, and then the frequency analysis is carried out. The relationship is the model of the system, so such methods are called model based frequency analysis. Based on different model assumptions, the frequency analysis methods can be categorised. For a linear model, the frequency analysis can be carried out by studying the transfer function or the frequency response function (FRF) of the system. For a nonlinear model, the frequency analysis requires more advanced techniques, such as the generalised frequency response functions (GFRFs) [9], and the nonlinear output frequency response functions (NOFRFs) [10].

This review gives a general background of the frequency analysis for fault detection. As the foundation of the model based frequency analysis, the signal based frequency analysis methods are introduced first. Then, the model based frequency analysis methods for linear and nonlinear systems are reviewed, respectively. The fault detection application examples are given for both the signal and model based frequency analysis.

2.2 Signal based frequency analysis

2.2.1 Spectrum analysis

Consider a general sequence $\{y(k)\}_{-N_s+1}^{N_s-1}$ is sampled at sampling frequency f_s in Hz, and the corresponding sampling period is $T = 1/f_s$, where $N_s \in \mathbb{Z}^+$, $y \in \mathbb{R}$, and $f_s \in \mathbb{R}^+$. The sequence $\{y(k)\}_{-N_s+1}^{N_s-1}$ can be interpreted as the summation of $2N_s - 1$ cosine signals with

different frequency, amplitudes and phases, which is given by [11]

$$y(k) = \frac{1}{2N_s - 1} \sum_{i=-N_s+1}^{N_s-1} |Y(e^{j\omega_i T})| \cos\left(\omega_i T k + \angle Y(e^{j\omega_i T})\right) \tag{2.1}$$

where $|\cdot|$ is the operator to compute the amplitude of the spectrum, $\angle\cdot$ is the operator to compute the phase angle of the spectrum,

$$\omega_i = \frac{2\pi f_s}{2N_s - 1} i \qquad i = -N_s + 1, -N_s + 2, \ldots, N_s - 1 \tag{2.2}$$

and

$$Y(e^{j\omega_i T}) = \mathcal{F}\{y\}(e^{j\omega_i T}) = \sum_{k=-N_s+1}^{N_s-1} y(k) e^{-j\omega_i T k} \tag{2.3}$$

is the discrete Fourier transform (DFT) of the sequence $\{y(k)\}_{-N_s+1}^{N_s-1}$. The inverse process (2.1), which reconstructs $y(k)$ from $Y(e^{j\omega_i T})$, is referred to inverse discrete Fourier transform (IDFT). By Euler rule, the formula of IDFT can be rewritten as

$$\begin{aligned} y(k) = \mathcal{F}^{-1}\{Y\}(k) = \frac{1}{2N_s - 1} \sum_{i=-N_s+1}^{N_s-1} Y(e^{j\omega_i T}) e^{j\omega_i T k}, \\ k = -N_s + 1, -N_s, \ldots, N_s - 1. \end{aligned} \tag{2.4}$$

$Y(e^{j\omega_i T})$ is a function of the discrete frequency ω_i in rad/s. The frequency resolution, which is the interval between ω_{i-1} and ω_i, is $2\pi f_s/(2N_s - 1)$ in rad/s. We define the discrete variable $\omega \in \{\omega_{-N_s+1}, \ldots, \omega_{N_s-1}\}$. The frequency domain function $Y(e^{j\omega T})$ is called the spectrum of the sequence $\{y(k)\}_{-N_s+1}^{N_s-1}$. The analysis on $Y(e^{j\omega T})$ is the spectrum analysis or Fourier analysis.

2.2.2 Power spectrum analysis

Power spectrum analysis is also frequently used in frequency analysis. Firstly, the cross-correlation function between the sequence $\{y(k)\}_{-N_s+1}^{N_s-1}$ and $\{u(k)\}_{-N_s+1}^{N_s-1}$ is obtained by [12]

$$R_{uy}(\tau) = u(k) * y(k) = \frac{1}{N_s} \sum_{k=0}^{N_s-1} [u(k) - \bar{u}_0][y(k - \tau) - \bar{y}_\tau] \tag{2.5}$$

where $\tau = 0, 1, \ldots, N_s - 1$, $\bar{u}_0$ is the sample mean of the sequence $\{u(k)\}_0^{N_s-1}$, and $\bar{y}_\tau$ is the sample mean of the sequence $\{y(k)\}_{-\tau}^{N_s-1-\tau}$. When $u(k) = y(k)$, the cross-correlation function R_{yy} is referred as the auto-correlation function of the sequence $\{y(k)\}_{-N_s+1}^{N_s-1}$. Secondly,

the cross spectral density, which is the DFT of the cross-correlation, is given by

$$S_{uy}(e^{j\omega T}) = \mathcal{F}\left\{R_{uy}\right\}(e^{j\omega T}) = \sum_{\tau=0}^{N_s-1} R_{uy}(\tau)e^{-j\omega T\tau}. \tag{2.6}$$

When $u(k) = y(k)$, the cross spectral density S_{yy} is referred as the power spectral density (or power spectrum), which is the DFT of the auto-correlation. The analysis on $S_{yy}(e^{j\omega T})$ is the power spectrum analysis.

2.2.3 Cepstrum analysis

In spectrum analysis, the frequency components are composed of the positive integer multiples of a frequency called fundamental frequency. The frequency components constitute a harmonic family, in which any frequency component is a harmonic.

This kind of spectra are common in gearboxes [13]. The different harmonic families of a gearbox originate from the different elements, such as the shafts, the ball-bearings, and the gearmesh frequencies. Through analysing each harmonic family, the health condition of the individual elements can be monitored. However, the harmonic families in some cases are difficult to separate by the traditional spectrum analysis. For example, when the speeds of the high speed shaft and the low speed shaft are close, the sidebands of the harmonics of the gearmesh frequency are overlapped to make it difficult to distinguish in the spectrum [14].

To address this issue, the cepstrum analysis is introduced. The (power) cepstrum is defined as the IDFT of the logarithmic power spectrum, which is given by [14]

$$C_p(\tau) = \mathcal{F}^{-1}\left\{\log\left(S_{yy}\right)\right\}(\tau). \tag{2.7}$$

It is known that the power spectrum S_{yy} is the DFT of the auto-correlation function $R_{yy}(\tau)$. Thus, the only difference between the cepstrum $C_p(\tau)$ and the auto-correlation function $R_{yy}(\tau)$ is that the IDFT is performed on the logarithm of the power spectrum rather than the power spectrum itself. The logarithm emphasises the small value harmonics, which leads that the cepstrum is more suitable for harmonics family separation than the auto-correlation. The independent variable of the cepstrum τ, referred to as quefrency, is a measure of time as in the case of the auto-correlation. A peak at a certain quefrency corresponds to the period

of the fundamental frequency of a harmonic family, by which the harmonic families can be separated.

2.2.4 High-frequency resonance technique

The high-frequency resonance technique is consider as the benchmark method for the fault detection and diagnosis in rolling element bearings [15]. The feasibility of the high-frequency resonance technique is based on the fact that an impulse vibration is generated at each time a defect in a rolling element bearing makes contact with another surface in the bearing [16]. The impulse vibration has a wide range of frequencies. However, when the bearing under the load, the spectrum of the bearing vibration is strongly masked by the vibration of gears or other machine elements. Accordingly, the wide distributed energy from the impulse vibration is difficult to detect by traditional spectrum analysis. Fortunately, as the system usually has higher natural frequencies than the frequency generated by the other machine elements, the impulse vibration excites high frequency resonances f_r, which can be used for the fault detection and diagnosis. Figure 2.1 gives an example of the spectrum of an inner race fault bearing. The radial vibration data were measured on the bearing housing at 51200 Hz sampling frequency. The record takes 10 seconds. The rotation speed is 29 Hz. The high frequency resonances can be observed in the figure.

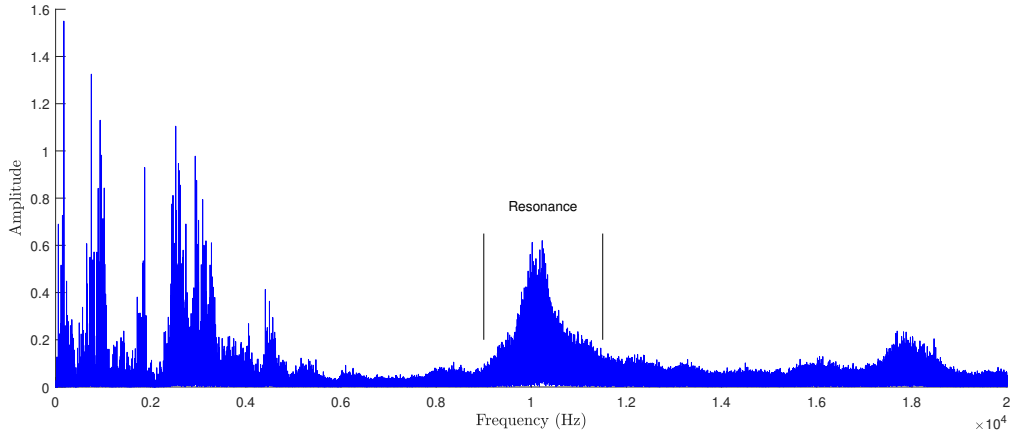


Figure 2.1: The resonance in the spectrum of an inner race fault bearing.

The impulse generated by the defect are periodic. The frequency of occurrence of the impulse is referred as to characteristic defect frequency, which is denoted as f_c. The char-

acteristic defect frequency is at relative low frequency range comparing with the resonance frequencies. The resonances as carrier signal is considered as being amplitude modulated at the characteristic defect frequency. Therefore, the spectrum of the defected bearing around the resonance has the frequency components at $f_r \pm f_c$. As the characteristic defect frequency is transformed to high frequency range, the transformed characteristic defect frequency is not influenced by the low frequency vibration of the other machine elements. Through demodulating the resonances, the characteristic defect frequency can be recovered for the fault detection.

The characteristic defect frequency is dependent on the location of the defect, which makes it possible to diagnose in which elements of the bearing the defect appears. For a bearing with a stationary outer race, the characteristic defect frequencies in different locations are given by [17]

- cage frequency:

$$\omega_c g = \frac{\omega_s}{2}\left(1 - \frac{d}{D}\cos\phi\right) \tag{2.8}$$

- ball spinning frequency:

$$\omega_b l = \frac{D\omega_s}{2d}\left(1 - \frac{d^2}{D^2}\cos^2\phi\right) \tag{2.9}$$

- outer race defect frequency:

$$\omega_{od} = Z\omega_c g = \frac{Z\omega_s}{2}\left(1 - \frac{d}{D}\cos\phi\right) \tag{2.10}$$

- inner race defect frequency:

$$\omega_{id} = Z(\omega_s - \omega_c g) = \frac{Z\omega_s}{2}\left(1 + \frac{d}{D}\cos\phi\right) \tag{2.11}$$

- rolling element defect frequency

$$\omega_{re} = 2\omega_b l = \frac{D\omega_s}{d}\left(1 - \frac{d^2}{D^2}\cos^2\phi\right) \tag{2.12}$$

where ω_s is the shaft rotation frequency in rad/s, d is the diameter of the rolling element, D is the pitch diameter, Z is the number of rolling elements and ϕ is the contact angle.

The whole procedural can be illustrated by Figure 2.2. First, the sampled signal $y(k)$ is filtered by a bandpass filter. The centre frequency of the bandpass filter should be the resonance to be studied. The bandwidth of the bandpass filter is at least double the highest characteristic defect frequency, so that the filter can pass the carrier frequency (i.e. resonance) with one pair of modulation sidebands [16]. Second, the filtered signal is demodulated by an envelope detector. The envelope detector can be constructed by a half-wave and a peak holder. Figure 2.2 gives an example that the half-wave rectifier only keeps the values of the bandpass filtered signal which are higher than the threshold, and then the signal is smoothed by the peak holder. Finally, the signal is transformed into frequency domain by DFT.

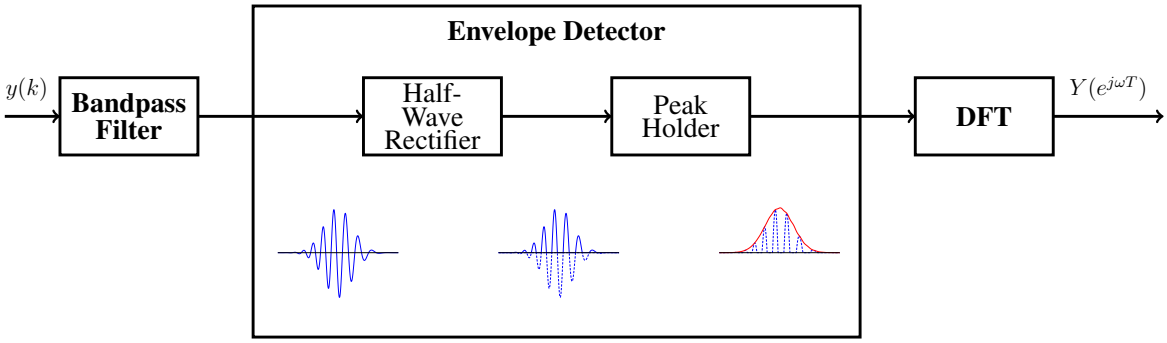


Figure 2.2: The procedure of high-frequency resonance technique.

2.3 Model based frequency analysis

The signal based frequency analysis, which does not take the system inputs into consideration, cannot reflect the system characteristics when the system inputs affect the output significantly. For example, when the system works under the varied loads, such as long-term health monitoring of large structures like bridges and wind turbines, the system outputs such as the structural vibrations are strongly depended on the ambient loads [18]. To take away the influence of the input, the mathematical model, which describes the relationship between the input and the output, is derived first. The model based frequency analysis is used to extract frequency features from the model.

In this section, both the linear and the nonlinear model based frequency analysis methods are reviewed. The system input and output are sampled in time domain. The sampling

frequency is f_s in Hz, and the corresponding sampling period is $T = 1/f_s$. The sampled input is $\{u(k)\}_{-N_s+1}^{N_s-1}$ and the sampled output is $\{y(k)\}_{-N_s+1}^{N_s-1}$. The following methods are applied on the two sequences.

2.3.1 Frequency response function

When the system is linear, the input and the output of the system satisfy the properties of superposition and homogeneity [19]. Under a sinusoidal input, e.g. $u(k) = \sin(\omega T k)$, the steady state output of a linear system is a sinusoidal signal at the same frequency as the input. However, the amplitude and the phase of the output is different from the input. The difference of the amplitude and the phase can be described by two functions of the input frequency, which are $|H(e^{j\omega T})|$ and $\angle H(e^{j\omega T})$, respectively. Then the output of the linear system under sinusoidal input is given by

$$y(k) = |H(e^{j\omega T})| \sin\left(\omega T k + \angle H(e^{j\omega T})\right). \tag{2.13}$$

The function $H(e^{j\omega T})$ is referred as frequency response function (FRF) of the linear system. To apply DFT on both side of (2.13), the relationship between the spectrum of the input $U(e^{j\omega T})$ and the spectrum of the output $Y(e^{j\omega T})$ can be obtained. It is observed that (2.13) is time shifted by $\frac{1}{\omega T}\angle H(e^{j\omega T})$ from the input. To use the time-shift property [11], the relationship of the input and output spectra can be written as

$$\begin{aligned} Y(e^{j\omega T}) =& |H(e^{j\omega T})| e^{j\omega T \cdot \frac{1}{\omega T}\angle H(e^{j\omega T})} U(e^{j\omega T}) \\ =& |H(e^{j\omega T})| e^{j\angle H(e^{j\omega T})} U(e^{j\omega T}) \\ =& H(e^{j\omega T}) U(e^{j\omega T}). \end{aligned} \tag{2.14}$$

In other word, FRF is the ratio of the input and the output spectra. Through FRF, the ratio of the input and output under different frequencies can be analysed. FRF as a frequency feature can be used to reflect the system characteristics for fault detection.

FRF can be obtained from the time domain model. First, the time domain model can be derived either by the first principle or the black-box modelling. Then, the linear time domain model is converted into the transfer function $H(z)$ by z-transform in discrete time. Finally, the FRF of the system is attained by replacing z by $e^{j\omega T}$ in the transfer function $H(z)$ [20].

It is also uncomplicated to obtain the FRF from the frequency domain directly. The test sinusoidal signals are reliably generated and commonly used in experimental instruments [19]. One direct way is to compute FRF through the spectrum of the test input and the measured output by (2.14). However, the output usually contains the immeasurable noise. To remove the noise, the cross-correlation between the input $\{u(k)\}_{-N_s+1}^{N_s-1}$ and the noisy output $\{y(k)+e(k)\}_{-N_s+1}^{N_s-1}$, where $e(k)$ represents noise, are computed first by

$$\begin{aligned} R_{uy_n}(\tau) &= u(k) * (y(k)+e(k)) \\ &= u(k) * y(k) + u(k) * e(k) \\ &= R_{uy}(\tau) + R_{ue}(\tau) \qquad \tau = 0, 1, \ldots, N_s - 1. \end{aligned} \tag{2.15}$$

When the noise $e(k)$ is uncorrelated with the input $u(k)$, i.e. $R_{ue}(\tau) = 0$, we have $R_{uy_n}(\tau) = R_{uy}(\tau)$. Then, the cross-spectrum S_{uy} and power spectrum S_{uu} can be computed by (2.6). Finally, through the convolution property of DFT, the noise free FRF of the linear system can be obtained by [21, 11]

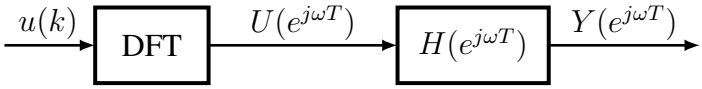
$$\begin{aligned} H(e^{j\omega T}) =& \frac{S_{uy}(e^{j\omega T})}{S_{uu}(e^{j\omega T})} \\ =& \frac{U(e^{j\omega T})Y(e^{j\omega T})}{U(e^{j\omega T})U(e^{j\omega T})} \\ =& \frac{Y(e^{j\omega T})}{U(e^{j\omega T})}. \end{aligned} \tag{2.16}$$

2.3.2 Nonlinear output frequency response functions

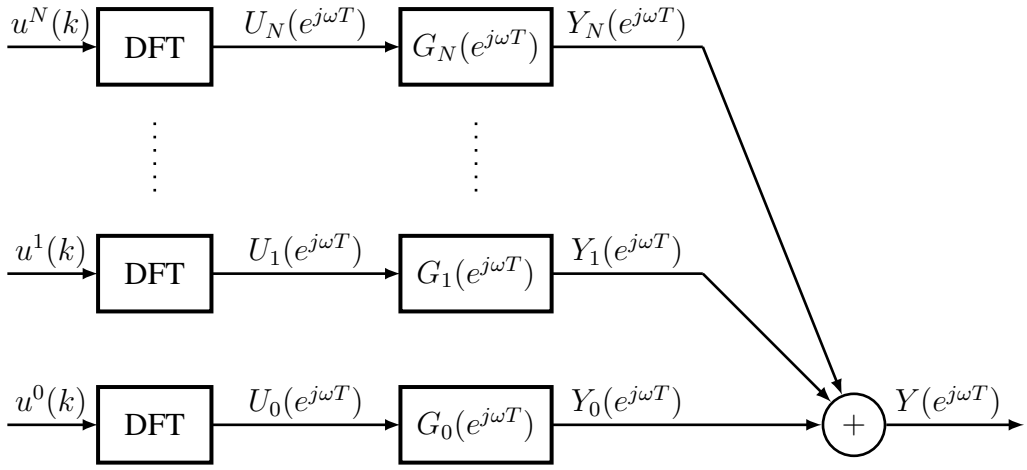
The concept of the nonlinear output frequency response functions (NOFRFs) [10] is an extension of linear FRF concept to nonlinear systems. Based on NOFRFs, the spectrum of the system output can generally be considered as the summation of the $N+1$ spectra shown in Figure 2.3, where N is the order of the system nonlinearity. Each spectrum member $Y_i(e^{j\omega T})$ is generated by

$$Y_i(e^{j\omega T}) = G_i(e^{j\omega T})U_i(e^{j\omega T}) \tag{2.17}$$

where $U_i(e^{j\omega T})$ is the spectrum of $u^i(k)$ and $G_i(e^{j\omega T})$ is i^{th} order NOFRF.



(a) The frequency response of linear system constructed by FRF.



(b) The frequency response of nonlinear system constructed by NOFRFs.

Figure 2.3: Comparison between FRF and NOFRFs.

Definition of NOFRFs

The NOFRFs are proposed based on the Volterra series representation of a nonlinear system as follows

$$y(k) = h_0 + \sum_{n=1}^{N} \sum_{\tau_1=0}^{N_s-1} \cdots \sum_{\tau_n=0}^{N_s-1} h_n(\tau_1, \ldots, \tau_n) \prod_{i=1}^{n} u(k-\tau_i) \qquad k = 0, \cdots, N_s - 1. \tag{2.18}$$

where $h_n(\tau_1, \ldots, \tau_n) \in \mathbb{R}$ is the n^{th} order Volterra kernel, and N is the order of the system nonlinearity. The nonlinear system (2.18) is stable, and the equilibrium is h_0. The spectra of the system input $\{u(k)\}_{-N_s+1}^{N_s-1}$ and output $\{y(k)\}_0^{N_s-1}$ are denoted as $U(e^{j\omega T})$ and $Y(e^{j\omega T})$, respectively. The output frequency response of the system (2.18) can be described as [10]

$$Y(e^{j\omega T}) = \sum_{n=0}^{N} Y_n(e^{j\omega T}), \tag{2.19}$$

where

$$\begin{aligned} Y_0(e^{j\omega T}) &= \begin{cases} h_0 & \text{for} \quad \omega = 0 \\ 0 & \text{for} \quad \omega \neq 0 \end{cases} \\ Y_n(e^{j\omega T}) &= \frac{1}{(2N_s-1)^{n-1}} \sum_{\omega_1+\cdots+\omega_n=\omega} H_n(j\omega_1,\ldots,j\omega_n) \prod_{i=1}^{n} U(e^{j\omega_i T}). \end{aligned} \tag{2.20}$$

$$H_n(j\omega_1,\ldots,j\omega_n) = \sum_{\tau_1=-N_s+1}^{N_s-1} \cdots \sum_{\tau_n=-N_s+1}^{N_s-1} h_n(\tau_1,\ldots,\tau_n) e^{-jT(\omega_1\tau_1+,\cdots,+\omega_n\tau_n)} \tag{2.21}$$

is known as the n^{th} order Generalised Frequency Response Function (GFRF), which is a description of the characteristics of nonlinear systems in the frequency domain, and

$$\sum_{\omega_1+\cdots+\omega_n=\omega} H_n(j\omega_1,\ldots,j\omega_n) \prod_{i=1}^{n} U(e^{j\omega_i T}) \tag{2.22}$$

denotes the summation of $H_n(j\omega_1,\ldots,j\omega_n) \prod_{i=1}^{n} U(e^{j\omega_i T})$ over the n-dimensional hyperplane $\omega_1 + \cdots + \omega_n = \omega$.

For $n = 0, 1, \ldots, N$, the spectrum of the $u^n(k)$ at the frequency ω is given by

$$\begin{aligned} U_0(e^{j\omega T}) &= \begin{cases} 1 & \text{for} \quad \omega = 0 \\ 0 & \text{for} \quad \omega \neq 0 \end{cases} \\ U_n(e^{j\omega T}) &= \frac{1}{(2N_s-1)^{n-1}} \sum_{\omega_1+\cdots+\omega_n=\omega} \prod_{i=1}^{n} U(e^{j\omega_i T}). \end{aligned} \tag{2.23}$$

Thus, we can define the n^{th} order NOFRF at frequency ω as

$$G_n(e^{j\omega T}) = \frac{Y_n(e^{j\omega T})}{U_n(e^{j\omega T})} \tag{2.24}$$

under the condition

$$U_n(e^{j\omega T}) \neq 0 \quad \text{for} \quad n = 0, 1, \ldots, N. \tag{2.25}$$

Therefore, $Y_n(e^{j\omega T})$ in (2.19) can be expressed as

$$Y_n(e^{j\omega T}) = G_n(e^{j\omega T}) U_n(e^{j\omega T}). \tag{2.26}$$

Consequently, the output frequency response of system (2.18) can be represensted using the NOFRFs as

$$Y(e^{j\omega T}) = \sum_{n=0}^{N} G_n(e^{j\omega T}) U_n(e^{j\omega T}). \tag{2.27}$$

The condition (2.25) implies that the NOFRF $G_n(e^{j\omega T})$ only exists when $U_n(e^{j\omega T}) \neq 0$. As $U_0(e^{j\omega T})$ is non-zero only at 0 frequency, we can use G_0 to stand for $G_0(e^{j\omega T})$. In addition, based on the definition (2.24), G_0 is identical to the system's stable equilibrium h_0.

Evaluation of the NOFRFs under Harmonic Inputs

In this research, harmonic signals with frequency ω_c ($\omega_c \neq 0$) will be used as the system inputs for NOFRFs evaluation. In this case, the possible nonnegative frequency components in the system output are $\{p\omega_c | p = 0, 1, \ldots, N\}$, and according to [22], the output frequency response of system (2.27) can now be described as

$$Y(jp\omega_c) = \sum_{i=0}^{q} G_{p+2i}(jp\omega_c) U_{p+2i}(jp\omega_c) \tag{2.28}$$

where $q = \lfloor (N-p)/2 \rfloor$. When a nonlinear system is subject to a harmonic input, the existence of the NOFRFs over different frequencies is shown in Table 2.1.

Table 2.1: The existence of NOFRFs over different frequencies when a nonlinear system is subject to harmonic input with 1 representing existence and 0 nonexistence.

Ω \ G	$G_0(e^{j\Omega T})$	$G_1(e^{j\Omega T})$	$G_2(e^{j\Omega T})$	$G_3(e^{j\Omega T})$	$G_4(e^{j\Omega T})$	$G_5(e^{j\Omega T})$	$\cdots$
0	1	0	1	0	1	0	$\cdots$
ω_c	0	1	0	1	0	1	$\cdots$
$2\omega_c$	0	0	1	0	1	0	$\cdots$
$3\omega_c$	0	0	0	1	0	1	$\cdots$
$\vdots$	0	0	0	0	$\vdots$	$\vdots$	$\vdots$

The NOFRF $G_n(e^{j\omega T})$ is insensitive to the change of the input by a constant factor α ($\alpha \neq 0$) [10], that is,

$$G_n(e^{j\omega T})\big|_{u(t)=u^*(t)} = G_n(e^{j\omega T})\big|_{u(t)=\alpha u^*(t)}. \tag{2.29}$$

Thus, if the model is excited by the harmonic input $u^*(t)$ scaled by $\bar{N}$ different constants $\alpha_1, \alpha_2, \cdots, \alpha_{\bar{N}}$, respectively, to produce $\bar{N}$ different system frequency responses $Y^i(e^{j\omega T}),\ i =$

$1, \ldots, \bar{N}$, the following equation can be obtained.

$$\mathbf{Y}(e^{jp\omega_c T}) = \mathbf{AU}(e^{jp\omega_c T})\mathbf{G}(e^{jp\omega_c T}) \tag{2.30}$$

where

$$\begin{aligned}
\mathbf{Y}(e^{jp\omega_c T}) &= \left[Y^1(e^{jp\omega_c T}), Y^1(e^{jp\omega_c T}), \ldots, Y^{\bar{N}}(e^{jp\omega_c T})\right]^{\mathrm{T}} \\
\mathbf{AU}(e^{jp\omega_c T}) &= \begin{bmatrix} U_p^1(e^{jp\omega_c T}) & U_{p+2}^1(e^{jp\omega_c T}) & \cdots & U_{p+2q}^1(e^{jp\omega_c T}) \\ U_p^2(e^{jp\omega_c T}) & U_{p+2}^2(e^{jp\omega_c T}) & \cdots & U_{p+2q}^2(e^{jp\omega_c T}) \\ \vdots & \vdots & \ddots & \vdots \\ U_p^{\overline{N}}(e^{jp\omega_c T}) & U_{p+2}^{\overline{N}}(e^{jp\omega_c T}) & \cdots & U_{p+2q}^{\overline{N}}(e^{jp\omega_c T}) \end{bmatrix} \\
\mathbf{G}(e^{jp\omega_c T}) &= \left[G_p(e^{jp\omega_c T}), G_{p+2}(e^{jp\omega_c T}), \ldots, G_{p+2q}(e^{jp\omega_c T})\right]^{\mathrm{T}}
\end{aligned} \tag{2.31}$$

and $U_n^i(e^{jp\omega_c T}),\ i = 1, \ldots, \bar{N},\ n = 0, \ldots, N$ is the spectrum of the input $(\alpha_i u^*(k))^n$ at the frequency $p\omega_c T$. To avoid Equation (2.30) to be underdetermined, it is required that $\bar{N} \geq N$. Consequently, the NOFRFs of nonlinear systems subject to a harmonic input can be determined by the least squares method, that is

$$\begin{aligned}
\mathbf{G}(e^{jp\omega_c T}) = \left[\mathbf{AU}(e^{jp\omega_c T})^{\mathrm{T}}\mathbf{AU}(e^{jp\omega_c T})\right]^{-1}\mathbf{AU}(e^{jp\omega_c T})^{\mathrm{T}}\mathbf{Y}(e^{jp\omega_c T}) \\
p = 0, 1, \ldots N.
\end{aligned} \tag{2.32}$$

2.4 Applications

2.4.1 Frequency feature measurements

For the fault detection by the frequency analysis, the different types of the time domain data can be measured. In mechanical systems, the vibration and acoustic emission (AE) are the most frequently used measurements [17, 18]. The vibration is measured by accelerometers, in which piezoelectric transducers are the most common [23]. The acoustic emission is measured by AE transducers, which are designed to detect the very high frequency stress wave [24]. The high frequency wave can be caused by the generation and propagation of cracks. This makes the acoustic emission be able to detect the growth of the cracks before they appear on the surface, while the vibration can only detect the cracks after they appear [17]. In

addition, the noise from neighbouring components, e.g. the gearbox or shaft around a bearing which is under study, are usually lower than 50 kHz [17]. These noises do not effect the acoustic emission which can reach to 2 MHz [24].

Impedance spectroscopy (IS) is another powerful frequency data used for determining the coatings on metals [25, 26], detecting corrosion cracking [27, 28, 29], analysing the biomechanical change of human tissues [5, 30], etc. The measurements, for generating impedance spectroscopy, are current and potential for electrochemical or electronic systems [31]. The general measurement approach is to apply an alternating voltage or current stimulus with certain frequencies ω to the electrodes and observe the response, i.e. the resulting current or voltage. Let $V(e^{j\omega T})$ denote the spectrum of the voltage and $I(e^{j\omega T})$ denote the spectrum of the current, and the impedance spectroscopy is given by

$$Z(e^{j\omega T}) = \frac{V(e^{j\omega T})}{I(e^{j\omega T})} \tag{2.33}$$

Compared to the FRF given in (2.14), the impedance spectroscopy $Z(e^{j\omega T})$ is the FRF of a linear system whose input is current and output is voltage. The impedance spectroscopy of the system is effected by the physical properties, such as diffusivity, rate constants, viscosity, and moisture, which can be used for the system fault detection [32].

2.4.2 Rolling element bearing

The rolling element bearing is a critical components in the rotating machines. The early alarm for the bearing defects helps the system maintenance, so that the more severe consequent failure can be prevented. The typical defect mode of rolling element bearing is the scratched crack, which is a result of surface fatigue caused by the repeated loading of the shaft [33]. Therefore, the main issue of rolling element bearing fault detection is to examine whether there is surface defect on the bearing.

The high-frequency resonance technique (or envelope analysis) is the benchmark method for the rolling element bearing fault detection [15]. It is known that bearings have some characteristic defect frequencies which are sensitive to the bearing fault. However, the characteristic defect frequencies are at the low frequency range, which is overlapped with the frequency range of the noise signals generated by the gearbox or other machine elements. To

solve this issue, the high-frequency resonance technique extracts the characteristic defect frequencies by demodulating high frequency resonances, which are not influenced by the noise. The high frequency resonances is excited by the impulse signal, which is generated whenever a defect of the bearing makes contact with another surface under the load [34]. These high frequency resonances are amplitude modulated at the characteristic defect frequencies.

The bearing fault can also be detected by the model based frequency analysis. The frequency response of the defected bearing vibration can be modelled by the first principles. Consider the inner-race-defect-induced impulse train as the input $u(t) = d(t)w(t)$, where $w(t)$ is a weighting function indicating the contact energy, and $d(t)$ is a unit impulse train defined by

$$d(t) = \sum_{k=-\infty}^{\infty} \delta(t - kT_d) \tag{2.34}$$

where T_d is the reciprocal of the inner race element passing frequency. The bearing vibration at the outer race or bearing housing as the output $y(t)$, the white-box model is given by [33]

$$y(t) = \sum_{m=1}^{n} \int_{-\infty}^{t} d(\tau)w(\tau)a_m(\tau)e^{-\alpha_m(t-\tau)} \cos\left(\omega_m(t-\tau)\right) d\tau. \tag{2.35}$$

For the mode m, the system characteristics α_m and ω_m are independent of the location of the input and the output, while the characteristics of the input-output transmission path a_m are effected by the locations of both the input and the output. The frequency response for the mode m is given by [33]

$$Y_m(j\omega) = \left[D(j\omega) * W(j\omega) * A_m(j\omega)\right] H_m(j\omega) \tag{2.36}$$

where $Y_m(j\omega)$, $D(j\omega)$, $W(j\omega)$ and $A_m(j\omega)$ are the spectrum of $y(t)$, $d(t)$, $w(t)$ and $a_m(t)$, respectively, and H_m is the spectrum of $e^{-\alpha_m t}\cos\left(\omega_m t\right)$.

Acoustic emission are also applied in the fault detection of bearings as complementary diagnostic tool for the vibration based methods [35]. Similar frequency analysis methods for the vibration measurement can be used for acoustic emission which is at high frequency level. It has been reported that acoustic emission offers earlier fault detection than the vibration [17, 35].

2.4.3 Gear

Gearboxes, which are widely used in heavy rotating machine such as wind turbines and helicopters, generally operate under tough environmental conditions. Gears as the key component of the gearbox are subject to pitting and fatigue cracks [36]. The accurate gear fault detection is critical to both the safety and the economic aspects.

Similar to the fault detection of bearing, the common techniques of gear fault detection is based on the vibration signals. The typical spectrum of the defected gear vibration is composed of the tooth meshing frequency and its harmonics, along with sidebands due to modulation. The modulation is caused by the impulse signals generated by the gear fault. The impulsive modulation leads to large numbers of sidebands spaced at the speed of the defected gear [14]. Therefore, the increment of the number or amplitude of the sidebands can be used for indicating the gear fault. The group of the sidebands with the equal spacing in the spectrum is a sideband family. The sidebands in the same family are generated from the same source [37]. However, the sideband families are mixed in the spectrum of the gear vibration, which are difficult to be separated. To address this issue, the cepstrum is introduced to detect and quantify the periodically spaced spectral components [14], which is illustrated in Section 2.2.3.

Time synchronous averaging (TSA) signal based frequency analysis techniques are also used in the gear fault detection. TSA is a noise removal technique, which is suit to analyse the the signal measured from the periodic rotating machine. The algorithm to compute TSA of the signal is provided by [38], and discussed in [39, 40].

Based on TSA signal $y(k)$, FM_0 index is used for detect major tooth faults, such as breakage, which results in an increase in peak-to-peak level but no significant change in the amplitude at the meshing frequency. FM_0 is defined as [41]

$$\text{FM}_0 = \frac{y_{pp}}{\sum_{i=1}^{n} |Y(j\omega_i)|} \tag{2.37}$$

where y_{pp} is the peak-to-peak amplitude of the TSA, $Y(j\omega_i)$ is the spectrum of $y(k)$ at the i^{th} harmonic of the gear mesh frequency.

2.4.4 Beam crack

The crack of beams can also be detected by analysing their vibration characteristics in frequency domain. For small cracks, the natural frequency of the beam is a sensitive index to indicate the crack [7]. It is found that the natural frequency of the cracked beam is lower than the normal beam. The reduction in natural frequency of the cracked beam can be explained through the linear white-box model introduced by [7]. First, the stiffness of the cracked beam is given by

$$K = \frac{1}{C} = \frac{1}{c + \Delta c} \tag{2.38}$$

where C is the flexibility of the cracked beam, c is the flexibility of the uncracked beam, and Δc is the local flexibility caused by the crack. For small crack depth the local flexibility Δc is proportional to $(a/h)^2$, where a is the crack depth and h is the height of the rectangular cross-section of the beam. Thus, the local flexibility can be expressed by

$$\Delta c = \lambda \left(\frac{a}{h}\right)^2 \tag{2.39}$$

where λ is a constant. The natural frequency square of the beam with small crack can be obtained by

$$(\omega_n + \Delta\omega_n)^2 = \frac{K}{m} = \frac{1}{cm}\left(1 + \frac{\Delta c}{c}\right) \tag{2.40}$$

where m is the mass of the beam and ω_n is the nature frequency of the normal beam. Due to $\omega_n = 1/cm$, the change of the nature frequency of the cracked beam is given by

$$\Delta\omega_n \approx -\frac{\lambda}{2\omega_n c}\left(\frac{a}{h}\right)^2. \tag{2.41}$$

This explains that, for a small crack depth, the reduction in natural frequency is proportional to the square of the crack depth ratio a/h.

However, the numerical and experimental results show the nonlinear effects are more sensitive to the crack than the feature of the linear model such as natural frequency and mode shapes [42]. For nonlinear frequency analysis of the beam, NOFRFs have been reported to be sensitive to the crack. In [22], frequency domain modelling was applied to obtain the NOFRFs directly from the input-output data of the beams. On the experimental test rig, two different harmonic inputs with the same waveform but different intensities were generated

by a shaker to excite three beams; one is crack-free, one is with a slight crack defect, and one is with a deep crack. The response vibration data were measured from an accelerometer at 8 kHz sampling frequency. The results showed that the NOFRFs successfully indicate the crack size. The larger NOFRFs values normally indicate larger crack sizes. In [43], the NOFRFs analysis was implemented through nonlinear white-box modelling of the beam. A finite element model for the relationship between the beam bending vibration and the external force was built by the first principles. According to the requirement of NOFRFs computation [10], two groups of harmonic inputs with same frequencies but different amplitudes were used. In each group, the frequencies of the harmonic inputs are from 1 to 200 Hz in step of 5Hz. The numerical responses of the model were computed by the Runge-Kutta method. The NOFRFs analysis shows that the high order NOFRFs are extremely sensitive to the crack on the beam. In [44], the NOFRFs of computed through the nonlinear black-box modelling for the structural damage detection. The method was introduced for the general structural damage including the beam crack, while the example given by the paper is the plates with a hole damage. The input-output data were obtained from the experimental set-up with three plates; one is undamaged, one is with a small hole, and the other one is with a large hole. However, only one pair of the input-output data was measured for each plate, which is not enough for computing NOFRFs directly through frequency domain modelling like the method given in [22]. This situation is common in practice. To solve this problem, the NARX models were built through the input-output data first. Then, more input-output data pairs were generated from the prediction of the NARX model. Finally, the NOFRFs were computed from the data pairs. The results showed that the NARX modelling and NOFRFs analysis are effective to detect the structural damages and distinguish the damage sizes.

2.4.5 Cancer

The same signal based and model based frequency analysis can be applied in the medical diagnosis such as cancer detection, and preterm birth prediction. The early diagnosis of cancer at a curable stage is crucial for the successful treatment of the disease. Electrochemical impedance spectroscopy (EIS), which is the frequency characteristics of the human tissue, has been considered as a label-free, mediator-free strategy for extracting precancerous fea-

tures in a fast, simple, and low cost fashion [45]. Some researchers have reported that the EIS is sensitive to the cancer cells [46, 47, 48]. However, the traditional EIS measurement devices can only record frequency domain data, which are measured when the output is in the steady state. Sufficient exploitation of EIS data over both steady and transient states have the potential to better reveal the electrical properties of tissues under investigation.

2.5 Summaries

This chapter reviews the frequency analysis techniques for fault detection. The damage sensitive features can be extracted through signal or model based frequency analysis. For the signal based frequency analysis, the features are extracted from the system output only. For the model based frequency analysis, the features are extracted through analysing the relationship between the input and output data. In the end, some application examples about machine fault detection and medical diagnosis using frequency analysis are provided.

Chapter 3

Techniques for Modelling, Feature Selection, Classification, and Model Validation

3.1 Introduction

This chapter is concerned with the introduction of the preliminaries for the following chapters and is divided into 4 parts. First the NARX modelling techniques are reviewed, which are used for the model based frequency feature extraction in Chapter 4 and 7. The filter feature selection techniques are reviewed as the background of our proposed feature selection method in Chapter 5. Finally, as the logistic regression model is used for the preterm birth prediction in Chapter 7, the related model evaluation and validation techniques are reviewed.

3.2 NARX modelling

When the real system which is unknown and only the input-output data of the system is known, the data can be used to build a NARX model through system identification methods, e.g. Forward Regression with Orthogonal Least-Squares (FROLS), to represent the real system. Here, the polynomial type NARX model is applied. Consider two general sequences $\{u(k)\}_{-N_s+1}^{N_s-1}$ and $\{y(k)\}_{-N_s+1}^{N_s-1}$ are sampled from the input and the output signals of a system

in the time domain, and the sampling period is T. The following methods can be applied to determine a NARX model of the underlying system.

3.2.1 Orthogonal least-squares (OLS) for polynomial NARX model

The polynomial type of the NARX model given by

$$\begin{aligned} y(k) = \theta_0 &+ \sum_{i_1=1}^{q} \theta_{i_1} s_{i_1}(k) + \sum_{i_1=1}^{q} \sum_{i_2=i_1}^{q} \theta_{i_1 i_2} s_{i_1}(k) s_{i_2}(k) + \cdots \\ &+ \sum_{i_1=1}^{q} \cdots \sum_{i_\ell=i_{\ell-1}}^{q} \theta_{i_1 i_2 \cdots i_\ell} s_{i_1}(k) s_{i_2}(k) \cdots s_{i_\ell}(k) + e(k) \end{aligned} \tag{3.1}$$

where ℓ is the degree of polynomial nonlinearity, $\theta_{i_1 i_2 \cdots i_\ell}$ are model parameters, and

$$s_n(k) = \begin{cases} y(k-n), & 1 \le n \le n_y \\ u(k-n+n_y+1), & n_y+1 \le n \le q = n_y + n_u + 1 \end{cases} \tag{3.2}$$

where $n_y < N_s$ and $n_u < N_s$ are the maximum lags for the system output and input, and q is the number of all possible terms for s_n given k.

Consider (3.1) as a generic linear-in-the-parameters representation

$$y(k) = \sum_{i=1}^{M} \theta_i x_i(k) + e(k) \tag{3.3}$$

where $\{y(k)\}$ with $k = 0, 1, \ldots, N_s - 1$ is the system output sequence, $x_i(k)$ with $i = 1, 2, \ldots, M$ is the regressor formed by the product of some $s_n(k)$. In this case, the number of the data is N_s, and the number of the term of the polynomial NARX model is M.

Rewrite (3.3) into a matrix representation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{e} \tag{3.4}$$

where

$$\begin{aligned}
\mathbf{y} &= [y(0), y(1), \ldots, y(N_s-1)]^\top \\
\mathbf{\Theta} &= [\theta_1, \theta_2, \ldots, \theta_M]^\top \\
\mathbf{e} &= [e(0), e(1), \ldots, e(N_s-1)]^\top \\
\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M] &= \begin{bmatrix} x_1(0) & x_2(0) & \cdots & x_M(0) \\ x_1(1) & x_2(1) & \cdots & x_M(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(N_s-1) & x_2(N_s-1) & \cdots & x_M(N_s-1) \end{bmatrix}
\end{aligned} \tag{3.5}$$

In this linear model, $\mathbf{X}$ is called the regressor matrix and $\mathbf{y}$ is called the target vector. If the regressor matrix $\mathbf{X}$ is full rank in columns, the matrix can decomposed as

$$\mathbf{X} = \mathbf{WA} \tag{3.6}$$

where $\mathbf{A}$ is an $M \times M$ unit upper triangular matrix and $\mathbf{W}$ is an $N_s \times M$ matrix with orthogonal columns $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M$. There are several ways to obtain $\mathbf{A}$ and $\mathbf{W}$ in (3.6), such as Gram-Schmidt, modified Gram-Schmidt, and Householder transformation [49]. Then (3.4) can be rewritten as

$$\mathbf{y} = \mathbf{WA\Theta} + \mathbf{e} = \mathbf{Wg} + \mathbf{e} \tag{3.7}$$

where $\mathbf{g} = [g_1, g_2, \ldots, g_M]^\top$. Thanks to the orthogonal property of $\mathbf{W}$, the inner product of the target vector $\mathbf{y}$ can be written into

$$\begin{aligned}
\mathbf{y}^\top\mathbf{y} &= \mathbf{g}^\top\mathbf{W}^\top\mathbf{Wg} + \mathbf{e}^\top\mathbf{e} \\
&= \sum_{i=1}^{M} g_i^2 \mathbf{w}_i^\top \mathbf{w}_i + \mathbf{e}^\top\mathbf{e}
\end{aligned} \tag{3.8}$$

where g_i is computed by orthogonal least-squares (OLS) [49]

$$g_i = \frac{\mathbf{y}^\top\mathbf{w}_i}{\mathbf{w}_i^\top\mathbf{w}_i}, \quad \text{for all } i = 1, 2, \ldots, M. \tag{3.9}$$

Divide both sides by $\mathbf{y}^\top\mathbf{y}$, Equation (3.8) can be written as

$$\begin{aligned}
1 &= \frac{\sum_{i=1}^{M} g_i^2 \mathbf{w}_i^\top \mathbf{w}_i}{\mathbf{y}^\top\mathbf{y}} + \frac{\mathbf{e}^\top\mathbf{e}}{\mathbf{y}^\top\mathbf{y}} \\
&= \sum_{i=1}^{M} \text{ERR}_i + \text{ESR}
\end{aligned} \tag{3.10}$$

where ERR_i refers to the error reduction ratio and ESR refers to the error-to-signal ratio. It can be found that the sum of the error reduction ratio ($\text{SERR} = \sum_{i=1}^{M} \text{ERR}_i$) is higher, the fitting error reflected by ESR is lower. Therefore, the SERR provides a effective means of determining the goodness of data fitting.

3.2.2 Forward Regression with OLS (FROLS)

When the q and ℓ are known, the number of the terms of the polynomail NARX (3.1) is

$$M = \begin{pmatrix} q+\ell \\ \ell \end{pmatrix} = \frac{(q+\ell)!}{q!\ell!}. \tag{3.11}$$

To use excessive terms in the model may lead to the overfitting issue [21]. Therefore, only the most significant terms should be selected from $\mathbf{X}$ into the regressor matrix $\mathbf{X_s}$ through feature selection methods. The feature selection method used in this paper is the forward regression, which means the regressor matrix $\mathbf{X_s}$ is empty before the feature selection, and the most significant terms are add into $\mathbf{X_s}$ [50]. On the contrary, the backward regression has the regressor matrix $\mathbf{X_s}$ including all the terms of $\mathbf{X}$ in the beginning, and then the less significant terms are removed from $\mathbf{X_s}$ [50].

In the forward regression, a criteria is usually required to compare the different candidate terms. The criteria value is computed for each candidate term to determine which term is the most significant. When the criteria is SERR which is computed by OLS, the feature selection algorithm is called FROLS. SERR can be used as the criteria based on the assumption that the better fitting performance implies the better regressor matrix $\mathbf{X_s}$. FROLS algorithm is given in Algorithm 1. The set W_s stores the orthogonal terms $\mathbf{w}_i$ which are selected by FROLS. The orthogonal terms $\mathbf{w}_i$ can be obtained by different methods, e.g. Gram-Schmidt orthogonalisation which is given by

$$\begin{aligned} \mathbf{u}_i &= \mathbf{x}_i - \sum_{r=1}^{m_s} \frac{\mathbf{w}_r^\top \mathbf{x}_i}{\mathbf{w}_r^\top \mathbf{w}_r} \mathbf{w}_r, \quad \text{for all } i = 1, 2, \ldots, m_c \\ \mathbf{w}_i &= \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|} \end{aligned} \tag{3.12}$$

where $\|\cdot\|$ is the Euclidean length, m_c is the number of candidates in the term library $\mathbf{X}$, and m_s is the number of the terms in W_s. The orthogonal term $\mathbf{w}_i$ is normalised into unit length.

When W_s is empty, $\mathbf{w}_i$ simply normalises $\mathbf{x}_i$ to unit length, i.e. $\mathbf{w}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$. After the orthogonalised term $\mathbf{w}_i$ is obtained, the corresponding ERR_i computed. The FROLS algorithm tries to find a term which maximises the SERR, but in practice, we try to find maximum ERR rather than SERR. As the orthogonalisation makes each term contribute to SERR independently, the new selected term does not change the SERR of the previous selected terms. Thus, the maximum of SERR is reached when the new selected term has the maximum of ERR. The stop criteria here is the number of the selected terms reach to M_s, which leads the polynomial NARX models always contain M_s terms. Another often used stop criteria is "ESR $= 1 -$ SERR $\leq \rho$", where ρ is a small number (say $\rho = 10^{-2}$). Under this stop criteria, the number of the terms of the polynomial NARX models is flexible.

Algorithm 1 Forward regression with OLS algorithm

function FROLS($\mathbf{X}, \mathbf{y}$) ▷ Select the most significance terms from $\mathbf{X}$
  $\mathbf{X_s} \leftarrow$ empty array ▷ Store selected terms
  $W_s \leftarrow \emptyset$ ▷ Orthogonalise terms in $\mathbf{X_s}$, and store in W_s
  SERR $\leftarrow 0$
  $m_c \leftarrow M$
  while do not meet *stop criteria* **do**
    for $\mathbf{x}_i \leftarrow$ each column of $\mathbf{X}$ **do**
      $\mathbf{w}_i \leftarrow$ make $\mathbf{x}_i$ to be orthogonal to the each feature in W_s
      $g_i \leftarrow \dfrac{\mathbf{y}^\top \mathbf{w}_i}{\mathbf{w}_i^\top \mathbf{w}_i}$
      $\text{ERR}_i \leftarrow \dfrac{g_i^2 \mathbf{w}_i^\top \mathbf{w}_i}{\mathbf{y}^\top \mathbf{y}}$
    end for
    $\text{ERR}_{max} \leftarrow$ maximum among $\{\text{ERR}_1, \text{ERR}_2, \ldots, \text{ERR}_{m_c}\}$
    SERR $\leftarrow$ SERR $+ \text{ERR}_{max}$ ▷ Maximise ERR is equivalent to maximise SERR
    $\mathbf{X_s} \leftarrow$ add $\mathbf{x}_{max}$ into $\mathbf{X_s}$
    $W_s \leftarrow$ add $\mathbf{w}_{max}$ into W_s
    $\mathbf{X} \leftarrow$ remove $\mathbf{x}_{max}$ from $\mathbf{X}$
    $m_c \leftarrow m_c - 1$
  end while
  return $\mathbf{X_s}$ ▷ The selected terms are in $\mathbf{X_s}$
end function

3.2.3 Term refinement

When the criteria of feature selection is SERR, the optimal subset of terms should have the maximum SERR in the all possible subsets, whose total number is M_s-combination of M elements. It is normally computational expensive to exhaustively compares SERR of all possible subsets. However, the algorithm like FROLS, which is called non-exhaustive search or greedy search, cannot guarantee to find the optimal subset (the detailed discussion is given in Section 3.3.1). Therefore, based on the terms selected by FROLS, term refinement can be applied to replace certain terms for increasing the SERR.

Iterative orthogonal forward regression

Guo et al. [51] find FROLS selecting suboptimal terms (or wrong term) often happens at the first term selection. The reason is when FROLS selects the first term, the candidate terms can not be orthogonalised to any selected terms. Therefore, a wrong term which carries the information from linear combination of more than one correct terms is likely be selected as the first term. The iterative orthogonal forward regression (iOFR) is developed to find a better first term than FROLS, which can eventually give larger SERR. First, FROLS selects M_s terms to construct a candidate library for the first term. Using FROLS to construct the candidate library is due to the expectation that FROLS selects a majority of correct terms and a few incorrect terms. Then, each term of the candidate library is adopted as the first term. Finally, for each first term, the rest of $M_s - 1$ terms are searched by FROLS. Thus, iOFR gives M_s subsets, one of which is the previous selection result of FROLS. The subset which has maximum SERR is selected, so iOFR can guarantee the SERR of the selected subset is no less than the SERR of the subset selected by FROLS.

Two-stage orthogonal least-squares

Zhang et al. in [52] introduce a two-stage OLS (TSOLS) method to refine the terms selected by FROLS. In the first stage, the terms are selected by the ordinary FROLS method. In the second stage (called the refinement stage), each selected term is reviewed to check whether a better term exists to replace that term.

The TSOLS method is illustrated in Algorithm 2. After the first stage, the term library $\mathbf{X}$ is divided into the selected term library $\mathbf{X_s}$ and the rest term library $\mathbf{X_r}$. In each time, the first selected term of $\mathbf{X_s}$ is removed from $\mathbf{X_s}$, and put into $\mathbf{X_r}$. Then, a new term is selected from the rest term library $\mathbf{X_r}$ by the FROLS method in the line 9. The new term is added behind the last place in $\mathbf{X_s}$. The term in $\mathbf{X_s}$ move in cycle to find whether the term in the first position of $\mathbf{X_s}$ can be replaced. The removed term is possible to be selected again, which is equivalent to switch the positions of the first and the last terms in $\mathbf{X_s}$. The variable m_{uc} counts when the term does not change. If term-unchanged happens consecutively M_s times, no term can be changed. Then the refinement process stops.

Algorithm 2 TSOLS

1: **function** TSOLS($\mathbf{X}, \mathbf{y}, M_s$) ▷ Select the most significance M_s terms from $\mathbf{X}$
2: $\mathbf{X_s} \leftarrow$ FROLS($\mathbf{X}, \mathbf{y}, M_s$) ▷ Select M_s terms by the FROLS method
3: $\mathbf{X_r} \leftarrow$ The rest terms of $\mathbf{X}$ except for $\mathbf{X_s}$
4: $m_{uc} \leftarrow 0$ ▷ Number of the unchanged terms
5: **while** $m_{uc} = M_s$ **do** ▷ The process stops when no term can be changed
6: $\mathbf{X_s} = [\mathbf{x_{s1}}, \ldots, \mathbf{x_{s}}_{M_s}]$ ▷ Define the column index
7: $\mathbf{X_r} \leftarrow [\mathbf{X_r}, \mathbf{x_{s1}}]$ ▷ The first term in $\mathbf{X_s}$ returns to the rest term library $\mathbf{X_r}$
8: $\mathbf{X_s} \leftarrow [\mathbf{x_{s2}}, \mathbf{x_{s3}}, \ldots, \mathbf{x_{s}}_{M_s}]$
9: $\mathbf{x_s}'_{M_s} \leftarrow$ FROLS($\mathbf{X_r}, \mathbf{y}, 1$) ▷ Select a new term from $\mathbf{X_r}$
10: $\mathbf{X_r} \leftarrow$ Remove $\mathbf{x_s}'_{M_s}$ from $\mathbf{X_r}$
11: $\mathbf{X_s} \leftarrow \left[\mathbf{X_s}, \mathbf{x_s}'_{M_s}\right]$ ▷ Add the new term at the end $\mathbf{X_s}$
12: **if** $\mathbf{x_s}'_{M_s} = \mathbf{X_{s1}}$ **then**
13: $m_{uc} \leftarrow m_{uc} + 1$ ▷ The term $\mathbf{X_{s1}}$ is unchanged
14: **else**
15: $m_{uc} \leftarrow 0$ ▷ The term $\mathbf{X_{s1}}$ change to the new term
16: **end if**
17: **end while**
18: **return** $\mathbf{X_s}$ ▷ The selected terms are in $\mathbf{X_s}$
19: **end function**

3.2.4 Sampling frequency issues in system modelling

In practice, the sampling frequency has a significant effect on the system modelling. A rough rule of thumb is choose the sampling frequency so that the data are sampled 10 times during the settling time of a step response [53, 12]. The variance of the model parameters will increase rapidly with the increase of the sampling frequency, when the parameter estimation is sensitive to the noise. The high sampling frequency will also lead to the nearly linearly dependent columns in the regressor matrix $\mathbf{X}$, and introduce high frequent noise in the data. On the other hand, the slow sampling makes the essential dynamics of the system can not be captured.

Aliasing will occur when the half of sampling frequency is lower than the highest frequency component of the signal according to the Nyquist–Shannon sampling theorem. To solve this issue, after the sampling frequency has been determined, the data should be considered for prefiltering by a low pass filter. The bandwidth of the filter should be smaller than the half of the sampling frequency. To avoid the signal distortion, the filter should have a constant gain and zero phase for its passband. In addition, as the high frequency noise in the data can be filtered out by the low pass filter, the signal-to-noise ratio is increased.

3.3 Filter feature selection

Feature selection techniques are widely used in machine learning to select a subset of features which are useful to classification models. Given a feature library $X = \{\mathbf{x}_i | 1 \leq i \leq M\}$, the objective of the feature selection is to select a subset $X_s = \{\mathbf{x}_i | 1 \leq i \leq M_s\}$ where $M_s \leq M$. For N_s samples, the corresponding feature matrix is $\mathbf{X} := (x_{i,j})_{N_s \times M}$, and the target vector is $\mathbf{y} = [y_1, \ldots, y_{N_s}]^\top$. The filter feature selection is to rank each feature with a criteria, such as correlation coefficient and mutual information, and then to select the features according to the rank. The core problem of the filter feature selection is to maximise the relevance between the selected features and the target, while to minimise the redundancy between the selected features. The following subsections provide two ideas to solve the problem.

3.3.1 Maximal relevance and minimal redundancy (mRMR)

The dependency between the features and the target is referred as relevance, while the dependency between the features themselves is redundancy. Peng et al. [54] propose a straight forward algorithm to maximise relevance while minimising redundancy. Both relevance and redundancy are evaluated by mutual information. The mutual information between feature $\mathbf{x}_k$ and the target $\mathbf{y}$ is given by

$$I(\mathbf{x}_k;\mathbf{y}) = \sum_{i=1}^{N_s}\sum_{j=1}^{N_s} P_{\mathbf{x}_k y}(x_{i,k}, y_j) \log \frac{P_{\mathbf{x}_k y}(x_{i,k}, y_j)}{P_{\mathbf{x}_k}(x_{i,k}) P_{\mathbf{y}}(y_j)} \tag{3.13}$$

where $P_{\mathbf{x}_k\mathbf{y}}$ is joint probability distribution, and $P_{\mathbf{x}_k}$, $P_{\mathbf{y}}$ are marginal distributions. The relevance is summation of mutual information between individual feature and target, which is given by

$$D(X_s, \mathbf{y}) = \frac{1}{M_s} \sum_{\mathbf{x}_i \in X_s} I(\mathbf{x}_i;\mathbf{y}). \tag{3.14}$$

The redundancy is summation of mutual information between two features, i.e.

$$R(X_s) = \frac{1}{M_s^2} \sum_{\mathbf{x}_i, \mathbf{x}_j \in X_s} I(\mathbf{x}_i;\mathbf{x}_j). \tag{3.15}$$

The maximal relevance and minimal redundancy (mRMR) can be realised by maximising

$$\Phi(D, R) = D - R. \tag{3.16}$$

To find the optimal subset maximising Φ, there are $\binom{M}{M_s}$ possible subsets to be compared, which is called *exhaustive search*. A realistic approach is to select only one feature in one step. In each step, the previously selected features will not be changed. When m_s features have been selected into X_s, the next feature $\mathbf{x}_i$ will be selected if it maximises the criterion function

$$I(\mathbf{x}_i;\mathbf{y}) - \frac{1}{m_s} \sum_{\mathbf{x}_j \in X_s} I(\mathbf{x}_i;\mathbf{x}_j). \tag{3.17}$$

Therefore, the selected feature $\mathbf{x}_i$ maximises the relevance with the target, while minimising the redundancy to the previously selected features in X_s. As each step selects the feature maximising the criterion function without reconsidering previously selected features, the approach is called *greedy search*. However, greedy search can not guarantee to find the optimal subset. Actually, no non-exhaustive search can guarantee to find the optimal subset [55].

3.3.2 Orthogonal point-biserial correlation coefficient

Another approach to control redundancy is based on feature orthogonalisation. Before the feature selection, the features in X are orthogonalised to each other. Then, the feature selection is applied on the orthogonalised features.

According to this idea, Solares et al. [56] introduce a feature selection method for binary classification which is called orthogonal point-biserial correlation coefficient (OBCC) method. The relevance between the feature $\mathbf{x}_i$ and the target $\mathbf{y}$ is evaluated by Pearson correlation, which is given by

$$r(\mathbf{x}_i, \mathbf{y}) = \frac{\sum_{k=1}^{N_s} (x_{k,i} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{N_s} (x_{k,i} - \bar{x}_i)^2 \sum_{k=1}^{N_s} (y_k - \bar{y})^2}} \tag{3.18}$$

where $\bar{x}_i$ and $\bar{y}$ are sample means. When the target $\mathbf{y}$ is dichotomy (say group A and group B), the estimate of Pearson correlation coefficient (3.18) is simplified to point-biserial correlation coefficient (BCC) [57] given by

$$r^b(\mathbf{x}_i, \mathbf{y}) = \frac{\bar{x}_i^A - \bar{x}_i^B}{\sqrt{\sum_{k=1}^{N_s} (x_{k,i} - \bar{x}_i)^2}} \sqrt{\frac{N_s^A N_s^B}{N_s}} \tag{3.19}$$

where $\bar{x}_i^A$ and $\bar{x}_i^B$ are the mean values on the vector $\mathbf{x}_i$ in the two groups, while N_s^A and N_s^B are the number of the samples in the two groups.

Similar to mRMR, the OBCC method also adopts greedy search, where only one feature is selected from X at one time. In each time, a candidate feature $\mathbf{x}_i$ is orthogonalised with the previously selected features in X_s to obtained the orthogonalised feature $\mathbf{w}_i$. Then, the BCC $r^b(\mathbf{w}_i, \mathbf{y})$, which is defined as OBCC between feature $\mathbf{x}_i$ and the target $\mathbf{y}$, is computed. The feature which has the highest OBCC with the dichotomous target $\mathbf{y}$ will be selected.

The pseudo-code of the OBCC method is described in Algorithm 3. The set W_s stores the orthogonal features $\mathbf{w}_i$ which are selected by orthogonal BCC. The orthogonal features $\mathbf{w}_i$ can be obtained by different methods, e.g. when the elements of W_s and $\mathbf{x}_i$ are linearly independent, Gram-Schmidt orthogonalisation which is given by

$$\begin{aligned} \mathbf{u}_i &= \mathbf{x}_i - \sum_{j=1}^{m_s} \frac{\mathbf{w}_j^\top \mathbf{x}_i}{\mathbf{w}_j^\top \mathbf{w}_j} \mathbf{w}_j, \quad \text{for all } i = 1, 2, \ldots, m_c \\ \mathbf{w}_i &= \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|} \end{aligned} \tag{3.20}$$

where m_c is the number of candidates in the feature library X, and m_s is the number of the features in W_s. The orthogonal term $\mathbf{w}_i$ normalises into unit length. When W_s is empty, $\mathbf{w}_i$ is simply normalising $\mathbf{x}_i$ to unit length, i.e. $\mathbf{w}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$. The OBCC between $\mathbf{x}_i$ and $\mathbf{y}$ computed in the line 8. The main process in the lines 5-15 is repeated until the number of the selected features reaches M_s.

Algorithm 3 Orthogonal BCC

1: **function** OBCC($X, \mathbf{y}, M_s$) ▷ Select the most significance M_s features from X
2: $X_s \leftarrow \emptyset$ ▷ Store selected features
3: $W_s \leftarrow \emptyset$ ▷ Store orthogonalised features
4: $m_c \leftarrow M$ ▷ Number of the candidate features
5: **for** $m_s = 1$ to M_s **do** ▷ M_s is the maximum number of features can be selected
6: **for** $\mathbf{x}_i \leftarrow$ each element of X **do**
7: $\mathbf{w}_i \leftarrow$ make $\mathbf{x}_i$ to be orthogonal to all elements in W_s
8: $r^b(\mathbf{w}_i, \mathbf{y}) \leftarrow$ Equation (3.19)
9: **end for**
10: $r^b(\mathbf{w}_{max}, \mathbf{y}) \leftarrow$ maximum among $\left\{r^b(\mathbf{w}_1, \mathbf{y}), r^b(\mathbf{w}_2, \mathbf{y}), \ldots, r^b(\mathbf{w}_{m_c}, \mathbf{y})\right\}$
11: $X_s \leftarrow$ add $\mathbf{x}_{max}$ into X_s
12: $W_s \leftarrow$ add $\mathbf{w}_{max}$ into W_s
13: $X \leftarrow$ remove $\mathbf{x}_{max}$ from X
14: $m_c \leftarrow m_c - 1$
15: **end for**
16: **return** X_s ▷ The selected features are in X_s
17: **end function**

3.4 Performance evaluation for logistic regression

The logistic regression model is adopted as the classifier in this thesis. After M_s features are selected, the features are used as the regressor in a logistic regression model. The regressor

is given by

$$\mathbf{X}_s = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M_s} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M_s} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_s,1} & x_{N_s,2} & \cdots & x_{N_s,M_s} \end{bmatrix} \tag{3.21}$$

where N_s is the number of the instances. The probability is predicted by the logistic regression model

$$\boldsymbol{\pi} = \frac{1}{1 + e^{-(\beta_0 + \mathbf{X}_s\boldsymbol{\beta})}}. \tag{3.22}$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_{N_s}]^\top$ and the parameter vector is $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_{M_s}]^\top$. The target is given by $\mathbf{y} = [y_1, y_2, \ldots, y_{N_s}]^\top$. The target is a binary vector, which only takes values of 0 and 1. Assume the target follows the binomial distribution. Thus, the likelihood with respect to the parameters β_0 and $\boldsymbol{\beta}$ is given by

$$l(\beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{N_s} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} = \prod_{i=1}^{N_s} \frac{e^{y_i(\beta_0 + \mathbf{X}_s\boldsymbol{\beta})}}{1 + e^{\beta_0 + \mathbf{X}_s\boldsymbol{\beta}}}. \tag{3.23}$$

The corresponding log-likelihood is defined as

$$\begin{aligned} L(\beta_0, \boldsymbol{\beta}) &= \ln\left[l(\beta_0, \boldsymbol{\beta})\right] \\ &= \sum_{i=1}^{N_s} \left[y_i \ln\left(\pi_i\right) + (1 - y_i) \ln\left(1 - \pi_i\right)\right] \\ &= \sum_{i=1}^{N_s} y_i \left(\beta_0 + \sum_{j=1}^{M_s} x_{i,j}\beta_j\right) - \ln\left(1 + e^{\beta_0 + \mathbf{X}_s\boldsymbol{\beta}}\right). \end{aligned} \tag{3.24}$$

The optimal parameters of β_0 and $\boldsymbol{\beta}$ are obtained by the maximum-likelihood estimate (MLE). The MLE of β_0 and $\boldsymbol{\beta}$ are $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ that maximise $L(\beta_0, \boldsymbol{\beta})$. As no closed-form solution exists, $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ are obtained by iterative algorithms such as Newton-Raphson method and Fisher scoring method [58].

This section briefly introduces three commonly used statistic indexes to evaluate the performance of logistic regression models.

3.4.1 Receiver operating characteristic

The predicted probability from the logistic regression model is given by

$$\hat{\boldsymbol{\pi}} = \frac{1}{1 + e^{-\left(\hat{\beta}_0 + \mathbf{X}\hat{\boldsymbol{\beta}}\right)}} \tag{3.25}$$

where $\hat{\boldsymbol{\pi}} = [\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{N_s}]^\top$. The predicted probability range from 0 to 1. A threshold value π^* in the range for determining whether a instance belongs to class 1 or class 0. The threshold value π^* gives four probabilities:

- $P(\hat{\pi}_i > \pi^* | y_i = 1)$: true positive rate (TPR) or sensitivity. The predicted probability higher than the threshold, and the target agrees with the prediction.
- $P(\hat{\pi}_i > \pi^* | y_i = 0)$: false positive rate (FPR). The predicted probability higher than the threshold, but the target disagrees with the prediction.
- $P(\hat{\pi}_i < \pi^* | y_i = 1)$: false negative rate (FNR). The predicted probability lower than the threshold, but the target disagrees with the prediction.
- $P(\hat{\pi}_i < \pi^* | y_i = 0)$: true negative rate (TNR) or specificity. The predicted probability lower than the threshold, but the target agrees with the prediction.

The point (FPR, TPR) with the threshold π^* can be plotted. When the π^* changes from 0 to 1, the point (FPR, TPR) will move along a curve which is called receiver operating characteristic (ROC) curve (Figure 3.1). The aggregated classification performance can be measured by the area under the curve (AUC) [59].

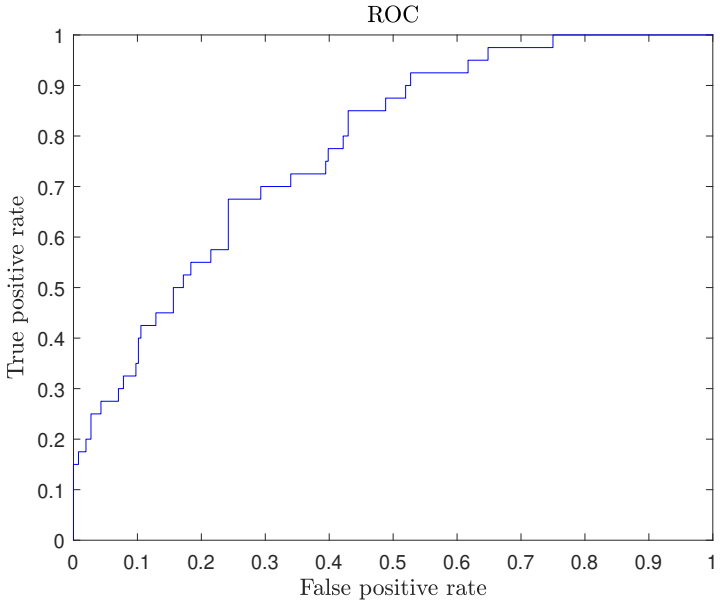


Figure 3.1: An example of ROC curve whose AUC $= 0.7771$.

3.4.2 Likelihood-ratio test

Different from ROC curve which used for evaluating classification performance, likelihood-ratio test is used for determining the significance of the features in the classification model. The fitted model (or unrestricted model) (3.22) is compared with the null model (or restricted model). The null model should be nested within the fitted model, which means the fitted model can be transformed into the null model by imposing constraints on its parameters. The null model normally used is given by

$$\boldsymbol{\pi}_0 = \frac{1}{1 + e^{-\beta_0}}, \tag{3.26}$$

which is the unrestricted model with the constrain $\boldsymbol{\beta} = \mathbf{0}$. The significance test is $H_0 : \boldsymbol{\beta} = \mathbf{0}$, the hypothesis is the independence between $\mathbf{X}_s$ and $\mathbf{y}$. The null hypothesis is true means the given model is not better than the null model which only has the intercept term. The likelihood-ratio test statistic is given by [60]

$$G^2 = -2\frac{L(\beta_0, \mathbf{0})}{L(\beta_0, \boldsymbol{\beta})}. \tag{3.27}$$

The statistic G^2 has an asymptotic χ^2 distribution with M_s degree of freedom. For a significance level α (e.g. 0.05), the null hypothesis is rejected when p-value is less than α.

3.4.3 Wald test

The Wald test is another common method to determine whether the regressor $\mathbf{X}_s$ is significant in the logistic regression model. An advantage of the Wald test over the likelihood-ratio test is that the Wald test does not require a null model. For the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$, the Wald test statistic is given by

$$W^2 = \hat{\boldsymbol{\beta}}^\top \Sigma_{\hat{\boldsymbol{\beta}}}^{-1} \hat{\boldsymbol{\beta}} \tag{3.28}$$

where $\Sigma_{\hat{\boldsymbol{\beta}}}$ is the covariance matrix of the parameter $\hat{\boldsymbol{\beta}}$. The matrix $\Sigma_{\hat{\boldsymbol{\beta}}}$ is the inverse of the information matrix $\mathcal{I}(\hat{\boldsymbol{\beta}})$ which has elements [61]

$$-E\left(\frac{\partial^2 L(\hat{\boldsymbol{\beta}})}{\partial\hat{\beta}_i \partial\hat{\beta}_j}\right) = \sum_{k=1}^{N_s} x_{k,i} x_{k,j} \hat{\pi}_k (1 - \hat{\pi}_k). \tag{3.29}$$

Thus, for the log-likelihood of the logistic regression, the information matrix is computed by

$$\mathcal{I}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \sum_{k=1}^{N_s} x_{k,1}^2 \hat{\pi}_k(1-\hat{\pi}_k) & \dots & \sum_{k=1}^{N_s} x_{k,1} x_{k,M_s} \hat{\pi}_k(1-\hat{\pi}_k) \\ \sum_{k=1}^{N_s} x_{k,2} x_{k,1} \hat{\pi}_k(1-\hat{\pi}_k) & \dots & \sum_{k=1}^{N_s} x_{k,2} x_{k,M_s} \hat{\pi}_k(1-\hat{\pi}_k) \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{N_s} x_{k,M_s} x_{k,1} \hat{\pi}_k(1-\hat{\pi}_k) & \dots & \sum_{k=1}^{N_s} x_{k,M_s}^2 \hat{\pi}_k(1-\hat{\pi}_k) \end{bmatrix}. \tag{3.30}$$

The Wald statistic W^2 is asymptotically distributed as χ^2 with M_s degree of freedom. Similar to the likelihood-ratio test, the null hypothesis is rejected when p-value less than the significance level α. The rejection of null hypothesis means the selected features $\mathbf{X}_s$ are statistically significant for the logistic regression model.

3.5 Cross-validation

The sampled data can be divided into two disjoint subsets, one is for the model training and another for the testing. The cross-validation is a statistical method to assess the model generalisation ability, which is the prediction performance of the model in the testing datasets. The model with high generalisation ability is insusceptible to the over-fitting, and has low variance in the bias-variance trade-off [62]. Therefore, the model is often evaluated and compared in cross-validation, where the prediction performance under the testing data rather than the training data is checked. The prediction performance under the testing dataset can be evaluated by the statistics, such as AUC and mean square error (MSE), which is called cross-validation criterion [63].

The cross-validation is best known in the model selection. The machine learning model normally is defined by a set of parameters which are derived in training and a set of hyper-parameters which are set before training [64]. In the logistic regression example, β_0 and $\boldsymbol{\beta}$ are parameters and M_s is the hyper-parameter. A set of models can be get by setting hyper-parameter as different values. The models can be evaluated by the cross-validation criterion, and the model selection is to choose a model which has the best criterion value.

Based on the different ways to divide the data into training and test datasets, several cross-validation methods are distinguished. Four cross-validation techniques are reviewed and compared.

3.5.1 k-fold cross-validation

The whole dataset is divided into k roughly identical size subsets. The k-fold cross-validation picks one subset as the testing data, and the rest $k-1$ subsets as the training data. The process is repeated k times, and each subset will be as the testing data once. The trained model can be assessed in the testing data by the statistic like MSE, so there are k statistic values are obtained. The cross-validation criterion is the averaged statistic values.

Leave one out cross-validation (LOOCV) is a special case when $k = N_s$, so the subset only contain 1 sample data. Since only one sample is held-out at a time, the cross-validation criterion is calculated from the averaged k individual held-out predictions. For the k-fold cross-validation, there are actually $C_k^{N_s}$ possible ways to split the data into k subsets, but only one of them is used. In LOOCV, as $C_{N_s}^{N_s} = 1$, all possible splitting ways have been used. Therefore, LOOCV is one of the exhaustive cross-validation methods [65].

3.5.2 Monte Carlo cross-validation

Monte Carlo cross-validation repeatedly splits the original dataset into two subsets, i.e. training dataset and testing dataset. The rule of thumb for the percentage of the training subset in the original data is about $75\% - 80\%$ [66]. For N_t data used in the training subset, there are $C_{N_t}^{N_s}$ possible repetitions to splits the data for Monte Carlo cross-validation. If N_r repetitions are used for the validation, the cross-validation criterion is computed by the average of N_r test data prediction performance. To choose a proper number of the repetitions for the model validation, the general rule can be followed [66]:

- The more proportion of data is in the training subset, the less bias is introduced in the validation.
- The more repetitions are used, the less uncertainty is introduced in the validation.
- The more proportion of data is in the training subset, the more repetitions are required to reduce the uncertainty.

Here, the bias in the cross-validation is the difference between the average of the testing prediction performance (i.e. cross-validation criterion) and the true value. The uncertainty is

the variance of the testing prediction performance.

The difference between Monte Carlo cross-validation and k-fold cross-validation is the testing datasets for the k-fold cross-validation are disjoint. For the small sample size, Monte Carlo cross-validation has the advantage to have larger size but overlapped testing datasets.

3.5.3 Stratified-holdout

It is likely that the splitting of the training and testing datasets is not representative, which means the proportions of each class in training and testing datasets are largely different. For the two classes example, an extreme splitting is that all the data of class 0 assign to the training dataset, and all the data of class 1 assign to the testing dataset. To solve this problem, the proportions of each class are preset for the training and the testing subsets. This procedure is called stratified-holdout [67, p. 149]. The stratification is especially important when the data are imbalanced [68].

3.5.4 Bootstrap

Bootstrap (or bagging) constructs the training dataset by sampling the original data with replacement. This means after a sample is selected into the training dataset, that sample is still available in the next selection. Thus, a training dataset may contain the duplicated data points. The unselected data are assign into the testing dataset, where the data is unduplicated. Similar to Monte Carlo cross-validation, the process repeated several times. In each repetition, the size of the testing dataset is varied.

A variant is called 0.632 bootstrap [69], when the size of the training dataset is same as the original dataset. In other word, the original dataset is sampled N_s times with replacement. The name is due to the fact that the opportunity of the data point is picked in the training data is a figure of

$$1 - \left(1 - \frac{1}{N_s}\right)^{N_s} = 1 - e^{-1} \approx 0.632. \tag{3.31}$$

Therefore, in 0.632 bootstrap, there are about 63.2% of the original data in the training dataset and 36.8% in the testing dataset.

3.6 Summaries

This chapter is mainly about modelling, feature selection, classification, and model validation techniques. For modelling, the FROLS algorithm is introduced for the NARX modelling. Some topics related to FROLS are also covered in this chapter, such as term refinement, and sampling frequency issues. For feature selection, the two ideas to develop the filter feature selection are reviewed. For classification, the logistic regression and its evaluation methods are introduced. For model validation, the techniques, especially those used for the small size and imbalanced data, are reviewed.

In Chapter 2, compared to FRF, NOFRFs are more difficult to be applied in the industrial areas. The reason is the computation of NOFRFs in (2.32) requires the system under investigation is stimulated by a group of the harmonic inputs which have the same frequency but different amplitudes. The condition for the system inputs is too strict to realisation in many industries, especially when the stimulation of the machines is from the natural environment, e.g. wind and sea wave. In addition, directly extracting the NOFRFs feature through the definition in Section 2.3.2 only utilises the steady-state of the system output, while the transient-state is wasted. To solve the two issues, this thesis will develop a novel method, which can utilise both the steady-state and transient-state data, to extract the NOFRFs under arbitrary inputs.

Inspired by the mRMR and the OBCC methods in Section 3.3, a novel filter feature selection method is developed. The proposed method gives better statistical meaning than OBCC, and achieves the better results than mRMR in linear classification using continuous features.

Chapter 4

Modelling and Model Feature Extraction for Nonlinear Systems with Multiple Stable Equilibria

4.1 Introduction

In the fault detection and condition monitoring, the critical parameters of a machine or its components are normally required to be monitored by sensors. However, it is not always feasible to measure the system parameters directly. To solve this issue, a model can be built from the system input-output data to represent the real system. Then, the features can be extracted from the data driven model to reveal the properties of the original machine system indirectly. The method is referred to as the modelling and model feature extraction method. The model structure used can be the NARX model and the modelling technique can be the FROLS, which has been introduced in Chapter 3. The model features can be extracted through frequency analysis as described in Chapter 2.

For a nonlinear system which has multiple stable equilibria, the model features around one equilibrium is different from another equilibrium. The aim of this study is to demonstrate the capability of the modelling and model feature extraction method to reflect the features of nonlinear systems around different stable equilibria. The demonstration starts with a simple linear system, which has a stable equilibrium A. An ARX model is built to represent the sys-

tem. Then, an example of the nonlinear system with two stable equilibria A and B are given. The NARX models are built in three scenarios to monitor the nonlinear system. First, the data for model training are collected when the system working around equilibrium A. Second, the data for model training are collected when the system working around equilibrium B. Third, the data for model training are collected when the system status switching between equilibrium A and B. The performance of the models is checked by examining whether the frequency features obtained from the models match the frequency features of the real system. Tests are also set for both the linear and nonlinear examples to examine, when a parameter of the real system changes gradually, whether the proposed method can track this change. It is found that in both examples the modelling and model feature extraction method effectively reflects the real system features and tracks the parameter change. The results imply the great application potential of the proposed method in multiple stable equilibria system fault detection and condition monitoring.

4.2 Methodology

4.2.1 Stability analysis of nonlinear systems with multiple equilibria

Suppose a general n-dimensional system is

$$\begin{aligned} \dot{y}_1 =& f_1(y_1, \ldots, y_n) \\ &\vdots \\ \dot{y}_n =& f_n(y_1, \ldots, y_n). \end{aligned} \tag{4.1}$$

An equilibrium is give by $(y_1^*, \ldots, y_n^*)$, so

$$\begin{aligned} 0 =& f_1(y_1^*, \ldots, y_n^*) \\ &\vdots \\ 0 =& f_n(y_1^*, \ldots, y_n^*). \end{aligned} \tag{4.2}$$

If we denote

$$\begin{aligned} v_1 =& y_1 - y_1^* \\ &\vdots \\ v_n =& y_n - y_n^*, \end{aligned} \tag{4.3}$$

the original system can be linearised as

$$\dot{\mathbf{v}} \approx \mathbf{J}_{(y_1^*,\ldots,y_n^*)}\mathbf{v}, \tag{4.4}$$

where $\mathbf{v} = (v_1, \ldots, v_n)^\top$ and $\mathbf{J}_{(y_1^*,\ldots,y_n^*)}$ is the Jacobian matrix given by

$$J_{i,j} = \frac{\partial f_i}{\partial y_j}(y_1^*, \ldots, y_n^*). \tag{4.5}$$

When all eigenvalues of $\mathbf{J}_{(y_1^*,\ldots,y_n^*)}$ have strictly negative real parts, the system is asymptotically stable at the equilibrium $(y_1^*, \ldots, y_n^*)$ [70].

The example of a two-dimensional nonautonomous nonlinear system is given by

$$\ddot{y} + d\dot{y} + cy + by^2 + ay^3 = u, \tag{4.6}$$

where y is the system output and u is the system input. When $b = 0$, the nonlinear system becomes Duffing's equation, which has two equilibria symmetrical about zero [71]. To make the location of equilibria be more general, the nonlinear system (4.6) is adopted to demonstrate how to describe the dynamics around the equilibria.

Let $y_1 = y$, $y_2 = \dot{y}$ and equation (4.6) can be described as

$$\begin{aligned} \dot{y}_1 &= f_1(y_1, y_2) = y_2 \\ \dot{y}_2 &= f_2(y_1, y_2) = -ay_1^3 - by_1^2 - cy_1 - dy_2, \end{aligned} \tag{4.7}$$

when $u = 0$. The equilibria can be obtained when $\dot{y}_1 = 0$ and $\dot{y}_2 = 0$. If $b^2 - 4ac \geq 0$ the equilibria are $(0, 0)$ and $\left(\frac{-b \pm \sqrt{b^2-4ac}}{2a}, 0\right)$.

The stability of the equilibria can be analysed by linearisation. Suppose that (y_1^*, y_2^*) is an equilibrium. Let

$$\begin{aligned} v_1 &= y_1 - y_1^* \\ v_2 &= y_2 - y_2^*. \end{aligned} \tag{4.8}$$

When v_1 and v_2 are small, the system can be linearised into

$$\begin{pmatrix} \dot{v}_1 \\ \dot{v}_2 \end{pmatrix} \approx \begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \tag{4.9}$$

The matrix

$$\mathbf{J}_{(y_1^*, y_2^*)} = \begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \frac{\partial f_1}{\partial y_2} \\ \frac{\partial f_2}{\partial y_1} & \frac{\partial f_2}{\partial y_2} \end{pmatrix}_{(y_1^*, y_2^*)} = \begin{pmatrix} 0 & 1 \\ -3ay_1^{*2} - 2by_1^* - c & -d \end{pmatrix} \tag{4.10}$$

is Jacobian matrix at equilibrium (y_1^*, y_2^*). When the equilibrium is $(0, 0)$, the Jacobian is

$$\mathbf{J}_{(0,0)} = \begin{pmatrix} 0 & 1 \\ -c & -d \end{pmatrix}. \tag{4.11}$$

When the equilibrium is $\left(\left(-b \pm \sqrt{b^2 - 4ac}\right)/2a, 0\right)$ denoted as $(y_1', 0)$, the Jacobian is

$$\mathbf{J}_{(y_1', 0)} = \begin{pmatrix} 0 & 1 \\ by_1' + 2c & -d \end{pmatrix}. \tag{4.12}$$

The determinant Δ, the trace τ and the discriminant $D = \tau^2 - 4\Delta$ of the Jacobian matrix is shown in the Table 4.1, where $y_1^+ = \frac{-b+\sqrt{b^2-4ac}}{2a}$ and $y_1^- = \frac{-b-\sqrt{b^2-4ac}}{2a}$. The type and stability of the equilibrium can be determined according to Figure 4.1 [72]. When $\Delta = 0$, $\tau = 0$ or $D = 0$, the determination of the type and stability of the equilibrium through the linearisation may be incorrect [72, p. 152]. It can be found that the equilibrium $(y_1^+, 0)$ is always saddle point when $y_1^+ < 0$, and the equilibrium $(y_1^-, 0)$ is always saddle point when $y_1^- > 0$.

Table 4.1: Characteristics of the Jacobian matrix for the equilibrium of the nonlinear system.

	$(0, 0)$	$(y_1^+, 0)$	$(y_1^-, 0)$
Δ	c	$y_1^+\sqrt{b^2 - 4ac}$	$-y_1^-\sqrt{b^2 - 4ac}$
τ	$-d$	$-d$	$-d$
D	$d^2 - 4c$	$d^2 + 4by_1^+ + 8c$	$d^2 + 4by_1^- + 8c$

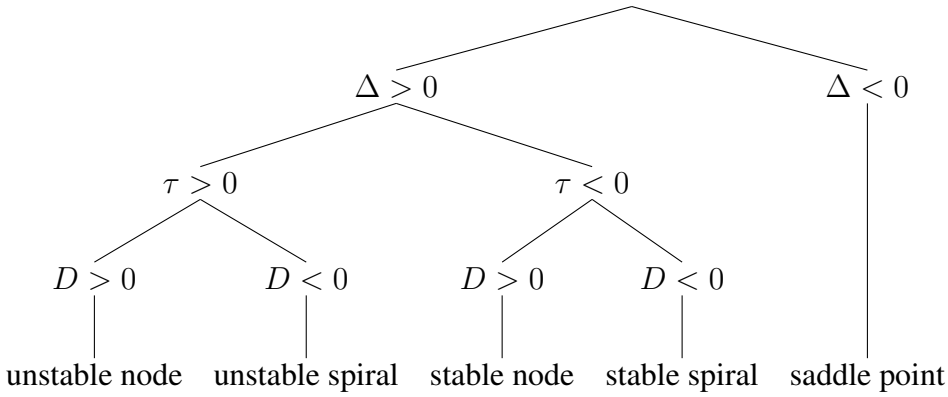


Figure 4.1: Classification of the equilibrium of the nonlinear system.

4.2.2 NARX model and NOFRFs model features

If the real system is unknown and only the input sequence $\{u(k)\}_{-N_s+1}^{N_s-1}$ output sequence $\{y(k)\}_{-N_s+1}^{N_s-1}$ are known, which are sampled at the sampling period T, the data can be used to build a NARX model (3.1) to represent the real system. The NARX model can be built by FROLS Algorithm 1. After the polynomial NARX models are obtained, the model features can be extracted for the system analysis. However, since the terms selected by FROLS may be different in the different polynomial models, the NARX model representation may not be unique. To solve this issue, the NOFRFs of the identified NARX model will be used to represent the model features and perform system analysis. In the following, two case studies will be used to demonstrated the ideas.

4.3 Case study 1: a linear system

Consider a special case when b and a of system (4.6) are 0. Thus, the system becomes linear, and an example is given by

$$\ddot{y} + \dot{y} + y = u. \tag{4.13}$$

Equation (4.13) can be written into the second order system

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{4.14}$$

where $y_1 = y$ and $y_2 = \dot{y}$. Obviously, this system only has one equilibrium $A(0, 0)$. The eigenvalues of the linear system are $-0.5 \pm 0.8660i$. As the real parts are less than 0, the system is stable. The phase portrait of the system is shown in Figure 4.2, where all trajectories are attracted into the equilibrium A.

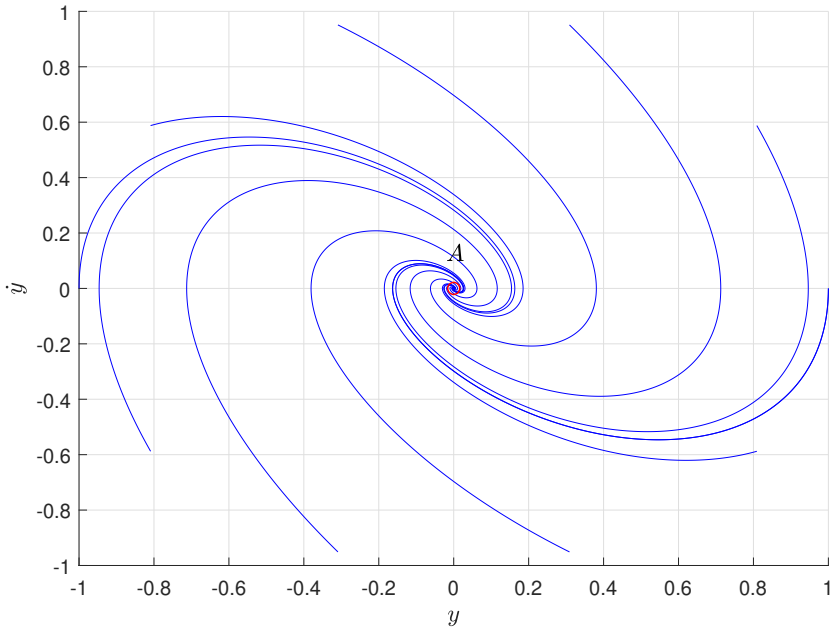


Figure 4.2: The phase portrait of the linear system (4.13).

The training data for modelling are collected under the harmonic input $u(k) = 0.1\cos(\omega_c kT)$ where $\omega_c = 0.2\pi$ and $T = 0.01$s. The corresponding output $y(k)$ is shown in Figure 4.3. The input-output data set $\mathcal{A}$ are used to train model M_A.

Table 4.3: Terms of ARX models through FROLS

Model term	Parameter	ERR
$y(k-1)$	1.9900	9.9995×10^{-1}
$y(k-2)$	-9.9005×10^{-1}	5.4615×10^{-5}
$u(k-1)$	9.9500×10^{-5}	1.5819×10^{-9}

The term library for modelling is constructed by linear terms with maximum 3 time delays for both input and output. Through FROLS, the 3 model terms can be found and shown in Table 4.3. Thus, an autoregressive with exogenous input (ARX) model [53] is obtained,

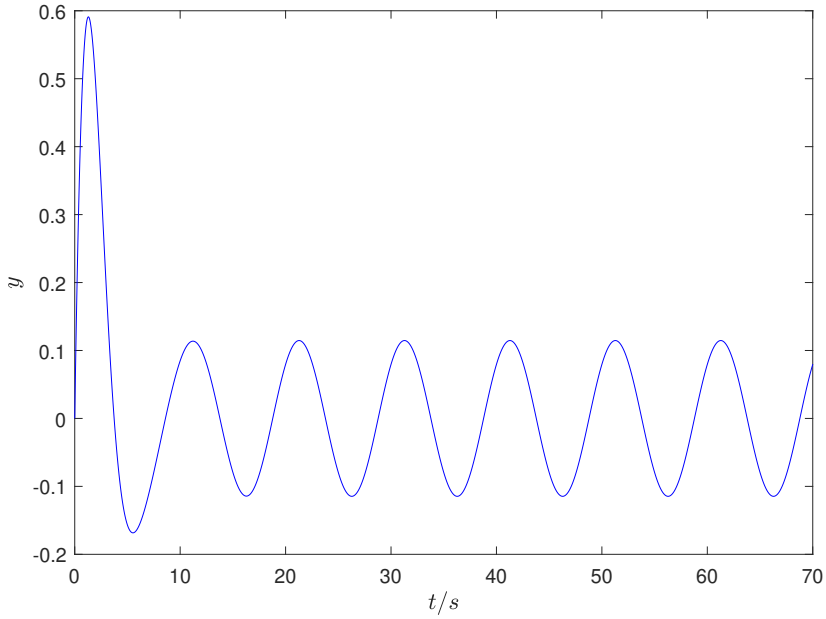


Figure 4.3: The harmonic responses in $\mathcal{A}$ for model training.

which is written as

$$\hat{y}(k) = 1.99\hat{y}(k-1) - 0.99\hat{y}(k-2) + 9.95 \times 10^{-5}u(k-1) \tag{4.15}$$

where $\hat{y}$ is the model simulated output. Therefore, the simulated output at arbitrary time $k = 1, \ldots, N_s$ can be computed when $u(k-1)$ and the initial conditions $y(-1)$, $y(0)$ are known. In Figure 4.4, it can be seen that the simulated output matches the measured output. The free oscillation trajectories can be observed in Figure 4.5. All trajectories are attracted to 0, which correctly reflect the location of the equilibrium of the real system. The frequency features of the real system and the ARX model are compared using FRF, which is shown in Figure 4.4. As FRF can be regarded as NOFRFs when the nonlinearity order is 1 ($N = 1$), the FRF under the frequency $\omega_c = 0.2\pi$ is computed by the approach for NOFRFs evaluation in Section 2.3.2. G_0 is also provided to compare with G_0 of the original system. Through the definition in Section 2.3.2, G_0 represent the equilibrium of the system. In Table 4.4, it is found that the model M_A correctly reflects the frequency features of the real system.

To show the frequency analysis can present the change in the system characteristics, the stiffness of the system is changed from 1 to 2. The data from the system are measured to update the the structure and parameters of the model M_A by the FROLS. It can be observed

from 4.6 that M_A can well track the change of the FRF of the original system. As the stiffness does not change the location of the equilibrium, G_0 is always 0.

Table 4.4: The NOFRFs around the equilibrium $A(0, 0)$

(a) The NOFRFs of the real system (4.13)

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$
0	1.8556×10^{-18}	
ω_c		1.1463

(b) The NOFRFs of M_A

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$
0	9.2738×10^{-17}	
ω_c		1.1463

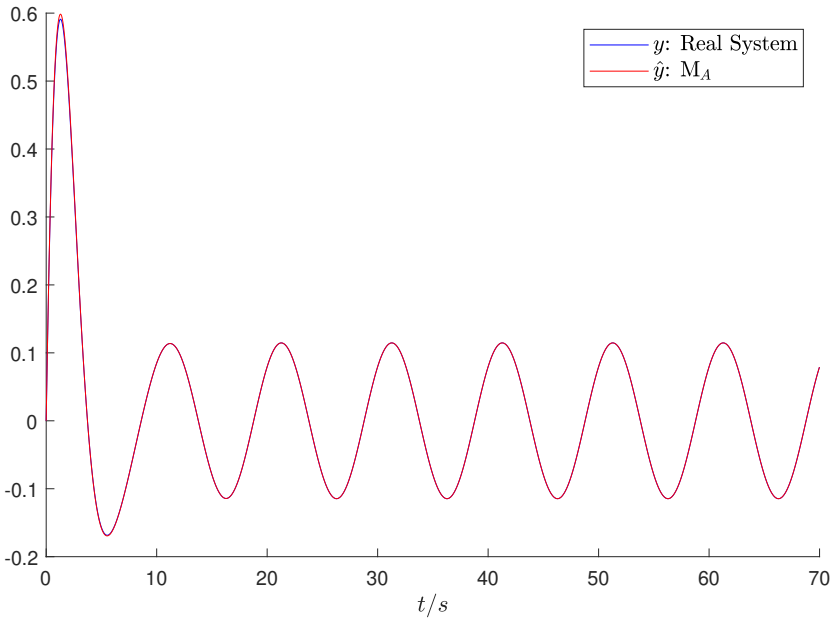


Figure 4.4: The harmonic responses of M_A.

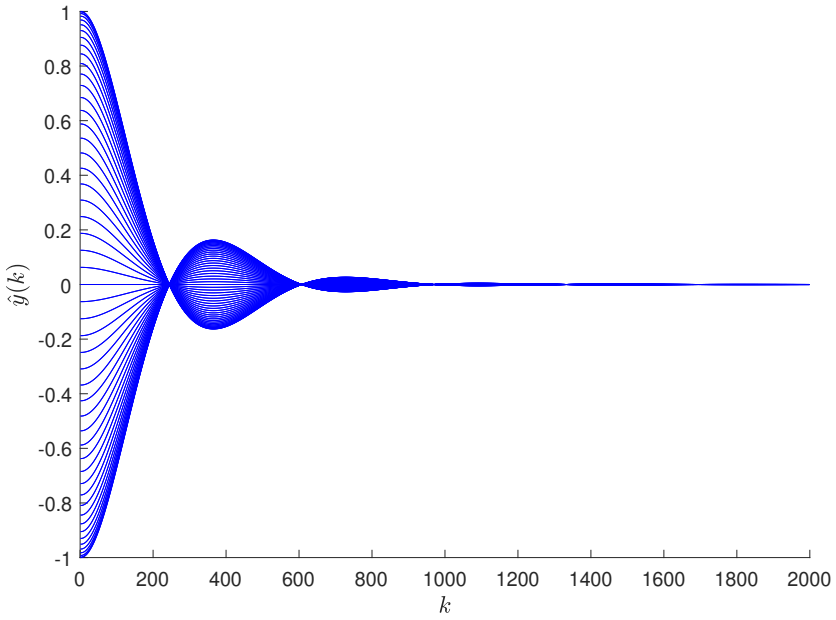


Figure 4.5: The free oscillation trajectories of M_A.

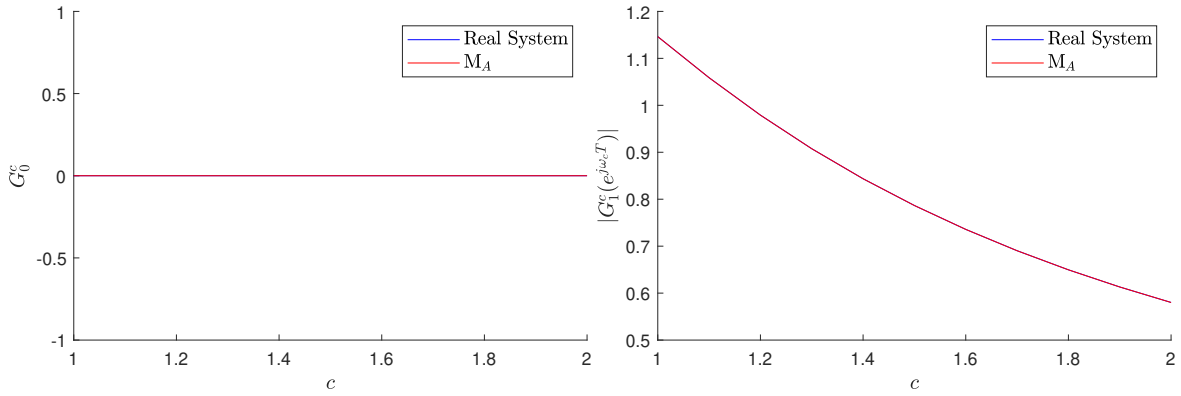


Figure 4.6: FRF of the real and identified system M_A.

4.4 Case study 2: a dual stable equilibria system

Consider the nonautonomous nonlinear system under the harmonic input which is given by

$$\ddot{y} + \dot{y} - y + y^2 + y^3 = u \tag{4.16}$$

where $u(t) = 0.1\cos(\omega_c t)$ and $\omega_c = 0.2\pi$. Rewrite the system (4.16) into the autonomous second-order system

$$\begin{aligned} \dot{y}_1 &= y_2 \\ \dot{y}_2 &= -y_1^3 - y_1^2 + y_1 - y_2. \end{aligned} \tag{4.17}$$

System (4.17) has three equilibria $(0, 0)$, $(0.618, 0)$, and $(-1.618, 0)$. According to Subsection 4.2.1, the Jacobian characteristics of the equilibria are given by Table 4.5. Then, the

Table 4.5: Characteristics of the nonlinear system (4.17) for each equilibrium.

	$(0,0)$	$(0.618,0)$	$(-1.618,0)$
Δ	-1	1.3820	3.6180
τ	-1	-1	-1
D	5	-4.5279	-13.4721

stability of the equilibria can be determined by Figure 4.1, which is shown in Table 4.6. The phase portrait of the nonlinear system is shown in Figure 4.7. Three red circles indicate the three equilibria, and all the trajectories are attracted to the two stable spirals. The following study is carried out on the two stable equilibria $A(0.618, 0)$ and $B(-1.618, 0)$.

Table 4.6: Classification of the equilibria of the nonlinear system (4.17).

		y	$\dot{y}$
	saddle	0	0
A	stable spiral	0.618	0
B	stable spiral	-1.618	0

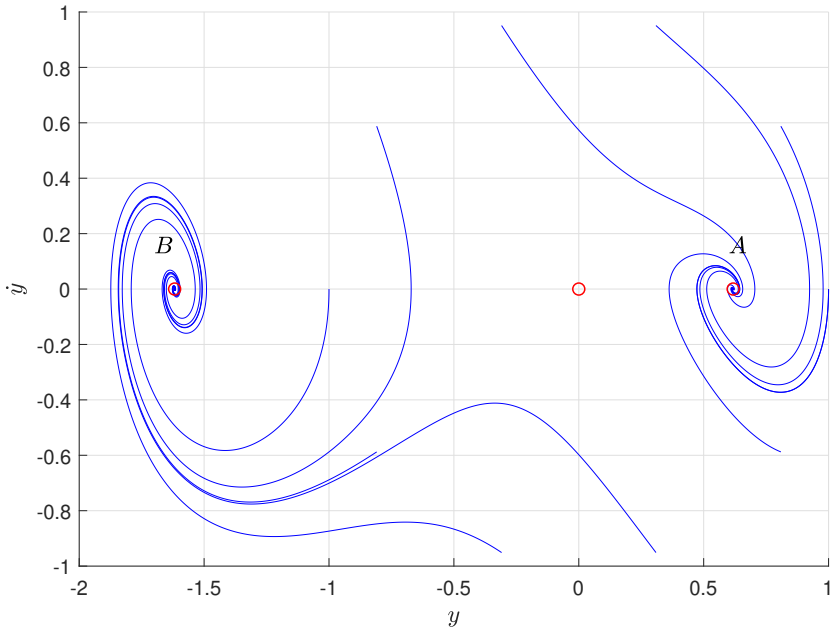
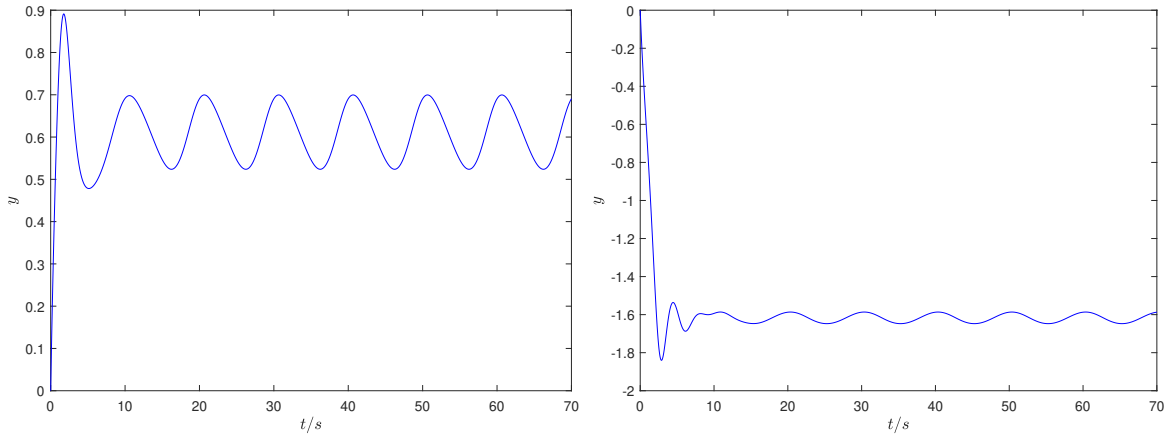


Figure 4.7: The phase portrait of the nonlinear system (4.16).

4.4.1 The model for single equilibrium

Two pairs of the input-output data sets $\mathcal{A}$ and $\mathcal{B}$ are sampled. The inputs in both sets are the same, i.e. $u(t) = 0.1\cos(\omega_c t)$ and $\omega_c = 0.2\pi$. The output in $\mathcal{A}$ is the harmonic response under the initial condition $(y, \dot{y}) = (0, 1)$, while the output in $\mathcal{B}$ is under the initial condition $(y, \dot{y}) = (0, -1)$. The data are sampled for 200 seconds, and the sampling period is 0.01 second, i.e. $T = 0.01$s. The outputs of the two data sets are shown in Figure 4.8. The output in $\mathcal{A}$ is attracted to the equilibrium $A(0.618, 0)$, while $\mathcal{B}$ is attracted to $B(-1.618, 0)$. It can also be observed that the outputs in both sets contain transient-state and steady-state. The transient-state data cannot be used for NOFRFs features extraction, but is important for NARX modelling.

After the data are sampled, $\mathcal{A}$ and $\mathcal{B}$ are used as training data for the modelling of two NARX models M_A and M_B, respectively. The terms of the NARX models are determined by the FROLS algorithm. In this cases, the nonlinearity of the model is 3 ($\ell = 3$), and the maximum delay is 3 for input and output ($n_u = n_y = 3$). The FROLS algorithm selects one term a time, until 5 terms are selected. The two models trained with $\mathcal{A}$ and $\mathcal{B}$ are given in Table 4.7.



(a) The output in $\mathcal{A}$. (b) The output in $\mathcal{B}$.

Figure 4.8: The harmonic responses used for model training.

To check whether the dynamics of the real system is reflected by the model correctly, the simulated output of the model is compared with the output of the real system. The simulated output $\hat{y}$ of the model can be easily computed using the models in Table 4.7. For example, M_A can be written into the form of the difference equation as

$$\begin{aligned}\hat{y}(k)&-1.9901\hat{y}(k-1)+9.8998\times10^{-1}\hat{y}(k-2)\\&+1.0122\times10^{-4}\hat{y}^3(k-2)+9.5678\times10^{-5}\hat{y}^2(k-1)=9.8608\times10^{-5}u(k).\end{aligned}\tag{4.18}$$

Then, the simulated output $\hat{y}(k)$ can be computed recursively when the input $u(k)$ and the initial condition $\big((y(-2),y(-1)\big)$ are known. It should be noticed that $\hat{y}(k)$ is not one-step-ahead prediction [21, p. 124], as no measured output y involving in (4.18). The comparison

Table 4.7: Terms of NARX models through FROLS

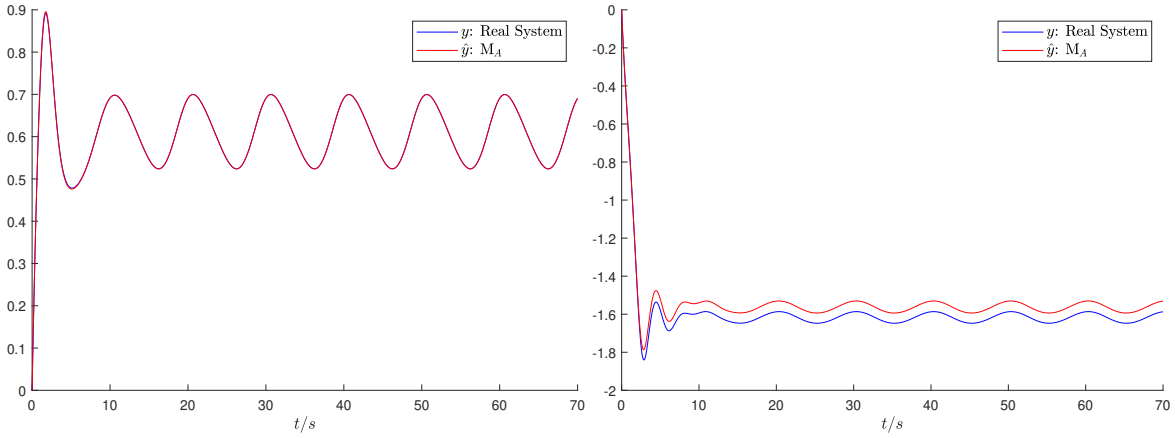
(a) The terms of M_A

Model term	Parameter	ERR
$y(k-1)$	1.9901	1.0000
$y(k-2)$	-9.8998×10^{-1}	1.2483×10^{-6}
$y^3(k-2)$	-1.0122×10^{-4}	4.2825×10^{-11}
$u(k)$	9.8608×10^{-5}	3.1661×10^{-11}
$y^2(k-1)$	-9.5678×10^{-5}	1.5930×10^{-13}

(b) The terms of M_B

Model term	Parameter	ERR
$y(k-1)$	1.9898	1.0000
$y(k-2)$	-9.8966×10^{-1}	2.6863×10^{-7}
$y^3(k-2)$	-9.4838×10^{-5}	1.3510×10^{-11}
$u(k)$	9.9477×10^{-5}	9.1010×10^{-12}
$y^2(k-2)$	-8.4296×10^{-5}	2.1262×10^{-13}

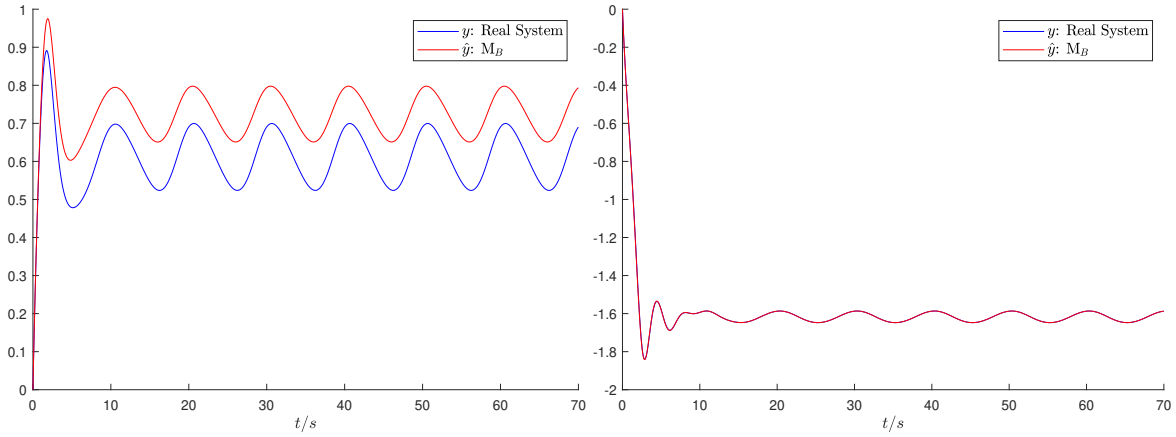
of the system harmonic responses from the simulated models and the real system models is shown in Figure 4.9 and Figure 4.10. It is known the dynamics of the nonlinear system (4.16) is dominated by the two equilibria $A(0.618, 0)$ and $B(-1.618, 0)$. Due to different initial conditions, the harmonic responses of the real system and the models are attracted to different equilibria. In Figure 4.9, it is found that M_A which are trained with $\mathcal{A}$ can correctly match the response around equilibrium A, but has bias in the response around B. In Figure 4.10, M_B can correctly match the response around equilibrium B, but has bias in the response around A. The equilibria of M_A and M_B can also be observed in the free oscillations (i.e. $u(k) = 0$) shown in Figure 4.11. The free oscillations start from the different initial conditions, and then the trajectories are attracted to the two stable equilibria. The trajectory stays at the unstable equilibrium $y = 0$ only when the initial condition is $\big((y(-2), y(-1)\big) = (0, 0)$. It can be seen that the locations of the equilibria for M_A and M_B are different.



(a) The responses around equilibrium A.

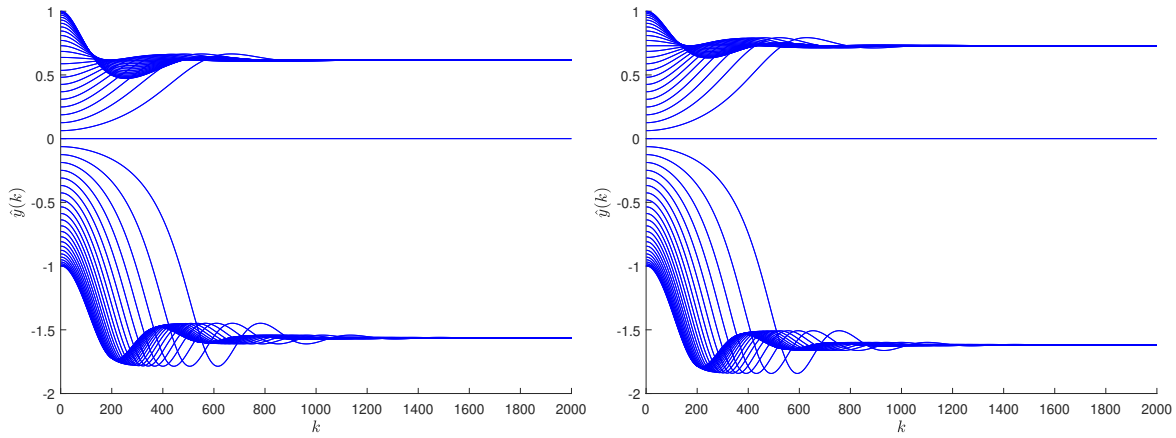
(b) The responses around equilibrium B.

Figure 4.9: The harmonic responses of M_A and the system (4.16).



(a) The responses around equilibrium A. (b) The responses around equilibrium B.

Figure 4.10: The harmonic responses of M_B and the system (4.16).



(a) The trajectories of M_A. (b) The trajectories of M_B.

Figure 4.11: The free oscillation trajectories of the models.

To quantitatively evaluate how the models describe the dynamics around the equilibria, the NOFRFs are adopted as the features of the nonlinear models. The dynamics around one stable equilibrium can be described by one set of NOFRFs. Therefore, each model or real system has two sets of NOFRFs.

To compute the NOFRFs, the harmonic inputs $\alpha_i \cos(\omega_c t)$ are given into the real system and models, where $\alpha_i = 0.01, 0.02, \ldots, 0.1$. The responses are generated, and the steady-state of the responses are used for the NOFRFs computation. It should be noticed that all the responses should be attracted around one equilibrium, so the NOFRFs can describe the dynamics around that equilibrium. Through choosing specific initial conditions, we can make the responses of the system or models to oscillate around a certain equilibrium. For the real system, the initial condition $(y, \dot{y}) = (0, 1)$ makes the responses to oscillate around $A(0.618, 0)$, and the initial condition $(0, -1)$ makes them around $B(-1.618, 0)$. For the models, the initial conditions are $\big(y(-2), y(-1)\big) = (0, 0.01)$ for A and $(0, -0.01)$ for B.

The NOFRFs for equilibrium A is given in Table 4.8. The nonlinearity of NOFRFs is infinity when the system or the NARX model is converted into Volterra series (2.18). The nonlinear terms higher than 4^{th} order are insignificance in this case, so the nonlinearity for the NOFRFs of is set as $N = 4$. Compare Table 4.8a and Table 4.8b, it is found that the NOFRFs of M_A match the real system, which implies that M_A can represent the dynamics of the real system around the equilibrium A. However, the NOFRFs of M_B cannot match the real system. In Table 4.9, the NOFRFs for the equilibrium B are computed. The NOFRFs of M_B match the real system, which implies M_B correctly reflects the dynamics around B. However, the NOFRFs of M_A cannot match the real system.

Table 4.8: The NOFRFs around the equilibrium $A(0.618, 0)$

(a) The NOFRFs of the real system (4.16)

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$	$\lvert G_2(e^{j\Omega T})\rvert$	$\lvert G_3(e^{j\Omega T})\rvert$	$\lvert G_4(e^{j\Omega T})\rvert$
0	6.1803×10^{-1}		1.5063		9.7888
ω_c		8.5460×10^{-1}		4.1949	
$2\omega_c$			1.6406		1.3076×10^{1}

(b) The NOFRFs of M_A

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$	$\lvert G_2(e^{j\Omega T})\rvert$	$\lvert G_3(e^{j\Omega T})\rvert$	$\lvert G_4(e^{j\Omega T})\rvert$
0	6.1808×10^{-1}		1.5142		9.7185
ω_c		8.5454×10^{-1}		4.1921	
$2\omega_c$			1.6425		1.3058×10^{1}

(c) The NOFRFs of M_B

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$	$\lvert G_2(e^{j\Omega T})\rvert$	$\lvert G_3(e^{j\Omega T})\rvert$	$\lvert G_4(e^{j\Omega T})\rvert$
0	7.2918×10^{-1}		9.2907×10^{-1}		3.8296
ω_c		7.1912×10^{-1}		2.1123	
$2\omega_c$			1.1927		5.5321

Table 4.9: The NOFRFs around the equilibrium $B(-1.618, 0)$

(a) The NOFRFs of the real system (4.16)

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$	$\lvert G_2(e^{j\Omega T})\rvert$	$\lvert G_3(e^{j\Omega T})\rvert$	$\lvert G_4(e^{j\Omega T})\rvert$
0	1.6180		9.8778×10^{-2}		4.8289×10^{-2}
ω_c		3.0451×10^{-1}		7.1560×10^{-2}	
$2\omega_c$			1.4922×10^{-1}		7.4213×10^{-2}

(b) The NOFRFs of M_A

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$	$\lvert G_2(e^{j\Omega T})\rvert$	$\lvert G_3(e^{j\Omega T})\rvert$	$\lvert G_4(e^{j\Omega T})\rvert$
0	1.5633		1.1011×10^{-1}		6.3083×10^{-2}
ω_c		3.1667×10^{-1}		8.7226×10^{-2}	
$2\omega_c$			1.7183×10^{-1}		9.6568×10^{-2}

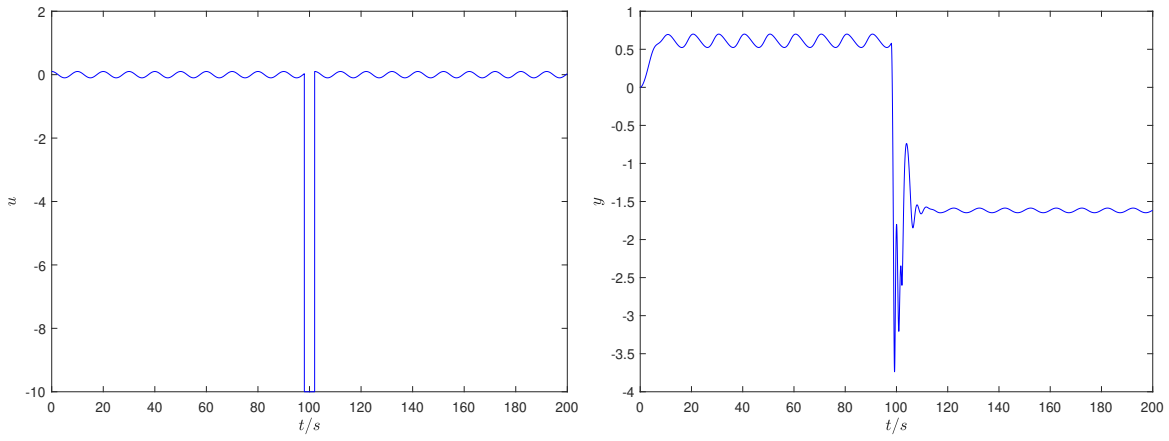
(c) The NOFRFs of M_B

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$	$\lvert G_2(e^{j\Omega T})\rvert$	$\lvert G_3(e^{j\Omega T})\rvert$	$\lvert G_4(e^{j\Omega T})\rvert$
0	1.6180		9.6726×10^{-2}		4.5351×10^{-2}
ω_c		3.0437×10^{-1}		6.8709×10^{-2}	
$2\omega_c$			1.4645×10^{-1}		6.9748×10^{-2}

4.4.2 The model for dual equilibria

The third input-output data set $\mathcal{C}$ is composed of dynamics around both the equilibrium A and B, which are shown in Figure 4.12. The system responses is firstly attracted around the equilibrium $A(0.618, 0)$. After 98 seconds, the large negative input is given to the system for 4 seconds. The system response is dragged to another equilibrium $B(-1.618, 0)$. After the transient process, the system response reaches steady-state at equilibrium B.

With the same FROLS configuration used in the last section, the NARX model M_C trained with $\mathcal{C}$ is given in Table 4.10. The free oscillation trajectories of M_C is shown in Figure 4.13. It is found the trajectories of the model M_C are attracted to the two stable equilibria around 0.618 and -1.618, except for one trajectory whose initial condition is 0 stays at the unstable equilibrium 0. Therefore, the positions of the equilibria match the equilibrium $A(0.618, 0)$ and $B(-1.618, 0)$ of the real system (4.17). The simulated harmonic responses of M_C around the two equilibria also match the responses of the real system, which are shown in Figure 4.14. To quantitatively evaluate the dynamics of M_C, the NOFRFs are computed with the same configuration as in the last section, which are shown in Table 4.11. By comparing Table 4.11 with the NOFRFs of the real system in Table 4.8a and Table 4.9a, it is found that M_C can reflect the system dynamics around both equilibria.



(a) The input in $\mathcal{C}$.

(b) The output in $\mathcal{C}$.

Figure 4.12: The input and the response used for model training.

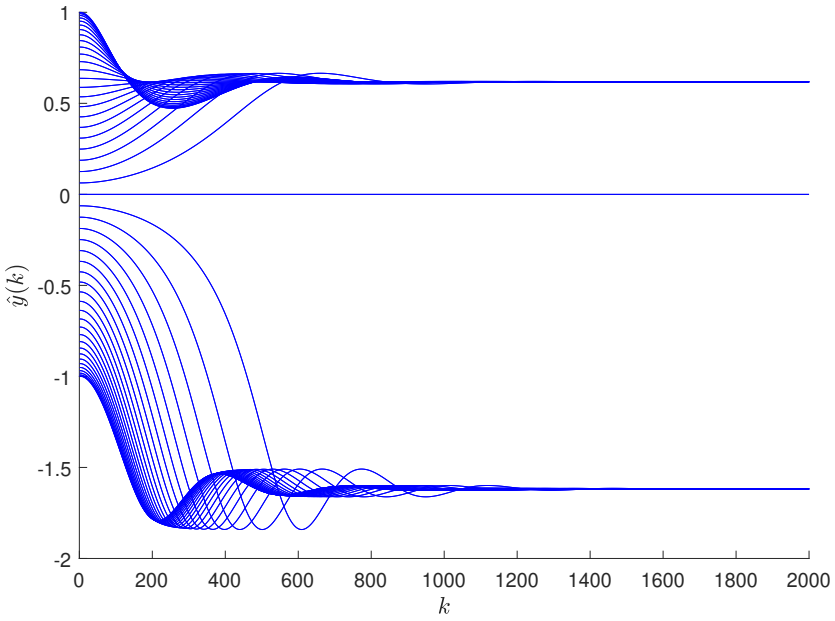
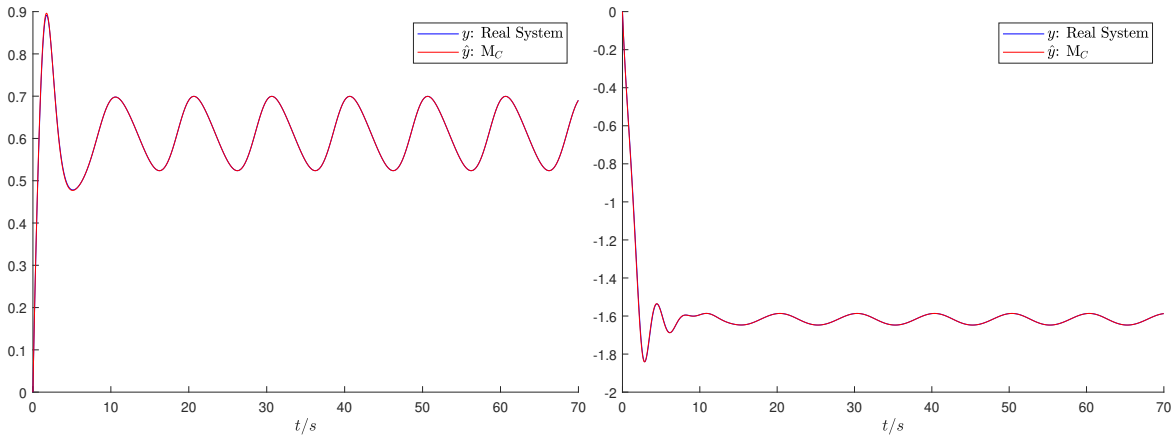


Figure 4.13: The free oscillation trajectories of M_C.

Table 4.10: Terms of NARX models M_C through FROLS

Model term	Parameter	ERR
$y(k-1)$	1.9902	9.9999×10^{-1}
$y(k-2)$	-9.9007×10^{-1}	1.0319×10^{-5}
$y^3(k-1)$	-9.9400×10^{-5}	4.1953×10^{-9}
$u(k-1)$	9.9306×10^{-5}	4.3337×10^{-9}
$y^2(k-1)$	-9.9416×10^{-5}	2.3879×10^{-9}



(a) The responses around equilibrium A. (b) The responses around equilibrium B.

Figure 4.14: The harmonic responses of M_C and the system (4.16).

Table 4.11: The NOFRFs of M_C for the two equilibria

(a) The NOFRFs around the equilibrium $A(0.618, 0)$

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$	$\lvert G_2(e^{j\Omega T})\rvert$	$\lvert G_3(e^{j\Omega T})\rvert$	$\lvert G_4(e^{j\Omega T})\rvert$
0	6.1796×10^{-1}		1.5054		9.7726
ω_c		8.5430×10^{-1}		4.1902	
$2\omega_c$			1.6402		1.3063×10^{1}

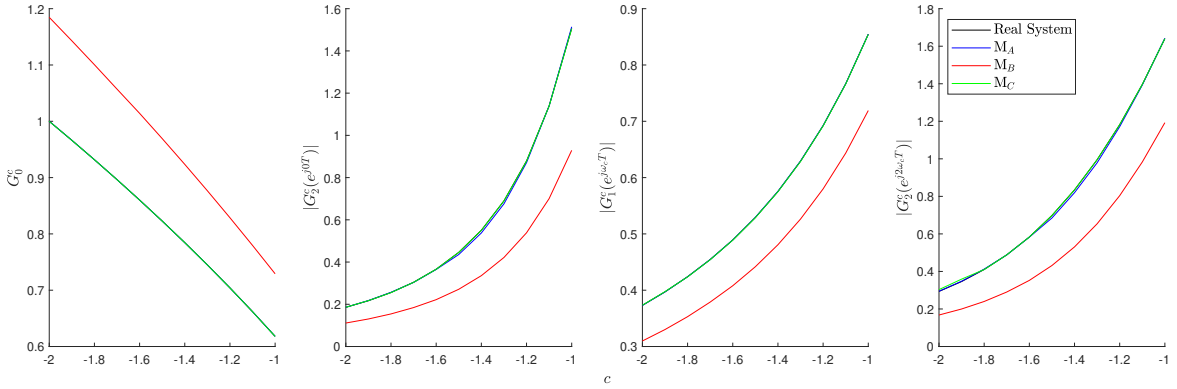
(b) The NOFRFs around the equilibrium $B(-1.618, 0)$

Ω \ $\lvert G\rvert$	$\lvert G_0(e^{j\Omega T})\rvert$	$\lvert G_1(e^{j\Omega T})\rvert$	$\lvert G_2(e^{j\Omega T})\rvert$	$\lvert G_3(e^{j\Omega T})\rvert$	$\lvert G_4(e^{j\Omega T})\rvert$
0	1.6181		9.8603×10^{-2}		4.8147×10^{-2}
ω_c		3.0425×10^{-1}		7.1404×10^{-2}	
$2\omega_c$			1.4907×10^{-1}		7.4010×10^{-2}

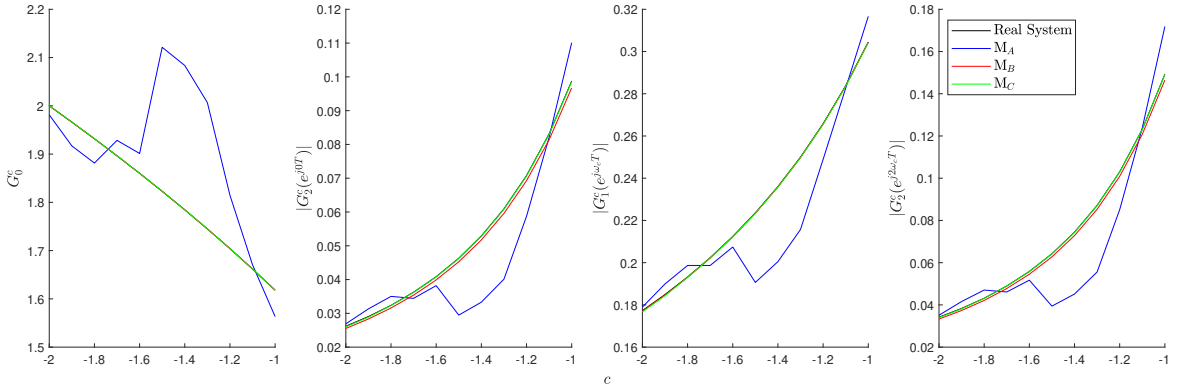
4.4.3 Evaluation of the changes in system parameters

Consider the stiffness of the system (4.16) varying gradually from -1 to -2, the investigation is carried out to examine whether the proposed modelling and model feature extraction method can track the parameter variation. When the stiffness changes, the input-output data are sampled to update the structure and parameters of the NARX model. The models are updated under three scenarios. For M_A, only the data around the equilibrium A are used for updating. For M_B, only the data around the equilibrium B are used for updating. For M_C, both of the data around the equilibrium A and B are used for updating. The three types of models are then used to monitor the variation in the system stiffness. Four NOFRFs features are extracted from each updated model. They are G_0^c, $|G_2^c(e^{j0T})|$, $|G_1^c(e^{j\omega_c T})|$, and $|G_2^c(e^{j2\omega_c T})|$, where c stands for the stiffness parameter given in the order $c = -1, -1.1, \ldots, -2$. Thus, the models are updated 11 times.

In Figure 4.15, the stiffness parameter change of the system modifies the system features around both the equilibrium A and B. For equilibrium A, the four NOFRF features of M_A and M_C all match that of the real system. Although the NOFRFs of M_B shows the trend of the stiffness change, the NOFRFs features given by the M_B gives significant bias. For equilibrium B, again the NOFRFs of M_B and M_C match that of the real system well. The error with the NOFRFs of M_A implies that M_A cannot reflect the real situation of the stiffness change about the equilibrium B. In addition, the NOFRFs of model M_C can be used to represent the system stiffness changes when the system works about either equilibrium A or B.



(a) NOFRFs features for equilibrium A.



(b) NOFRFs features for equilibrium B.

Figure 4.15: NOFRFs features of the system (4.16) when its stiffness changes.

4.5 Conclusions

This chapter demonstrates the capability of the modelling and model feature extraction method in the analysis of nonlinear systems around different stable equilibria. For modelling, three NARX models are build by the FROLS under different scenarios. For model feature extraction, the NOFRFs are adopted as the nonlinear model features. To check whether the model can reflect the situation of the real system, the harmonic responses of models and the real system are compared first. Then trajectories of the models in free oscillation are given to check the locations of the equilibria. Finally, the features of the models and the real system are compared quantitatively by evaluating the NOFRFs of both the identified models and the real system. Generally, the model built by the data dominated by one equilibrium cannot reflect the features of another equilibrium. When the data cover the dynamics of both equilibria,

the model have the ability to reflect the features of the two equilibria. In the evaluation of system parameter change, the results show that the system parameter change can be reflected by the system NOFRFs features in both equilibria. This study shows that the proposed modelling and model feature extraction can effectively reveal the features of linear and nonlinear systems and can, therefore, be a useful technique when applied in engineering system fault detection, where the NOFRF based system feature extraction can be exploited to reveal the system faulty conditions.

Chapter 5

Orthogonal Least Squares Based Fast Feature Selection for Classification

5.1 Introduction

The aim of the feature selection for classification is to select an optimal subset of features given the candidate features, which are continuous or categorical, and the response, which is categorical. The feature selection methods can be divided into three types: filter, wrapper, and embedded methods [50]. The filter methods rank the individual candidate features based on certain statistics, such as the correlation coefficient and the mutual information [54]. The wrapper methods train classifier by ranking the subsets of candidate features based on their classification performance. The embedded methods, e.g. Lasso [73] and CART [74], select optimal features during the training process of a specific classifier.

Comparing with the other two methods, a filter method is not based on a specific type of classifiers, so a filter method is more suitable to be used in the stage where the type of classifiers has not been decided. To rank the features by a filter method, it is desired that the features in the subset have the high relevance to the response, while the low redundancy between themselves. A straightforward way is to optimise the objective function constructed by the difference or the quotient between the relevance and the redundancy. For example, the well-known minimal-redundancy-maximal-relevance (mRMR) method adopt this idea, in which the relevance and redundancy are quantified by the mutual information [75]. The

second idea is to control the redundancy by orthogonalising the candidate features, and to find the maximum relevance between the orthogonalised features and the response. The second idea has been used in the term selection of time series models by Orthogonal Least Squares (OLS), where the relevance is defined by the error reduction ratio (ERR) [49]. The previous two ideas evaluate the relevance between the single feature and the response, and the relevance is analysed separately with the redundancy. The third idea uses the overall relevance between the subset features and the response. The definition of the overall relevance has taken the redundancy into consideration, e.g. the multiple correlation coefficient and the canonical correlation coefficient [76].

In this chapter, the third idea is adopted but based on a novel revelation and exploitation of a close relationship between the second and third idea. Based on OLS, the squared orthogonal correlation coefficients are defined and used to proposed a novel feature selection approach that can significantly improve the computational speed of the evaluation of the overall relevance. It is shown that the squared orthogonal correlation coefficients are especially useful in the greedy search for the best features. The relationship between the squared orthogonal correlation coefficients, the multiple correlation coefficient, the canonical correlation coefficient, and Fisher's criterion of the linear discriminant analysis are analysed showing the proposed OLS based feature selection method has following three advantages:

- fast in the greedy search;
- equivalent to the Canonical Correlation Analysis (CCA) and the Linear Discriminant Analysis (LDA);
- applicable to both continuous and categorical features.

The rest of the chapter is organised as follows. In Section 2, the definition of the squared orthogonal correlation coefficients is given. The relationships of the squared orthogonal correlation coefficients with the multiple correlation coefficient and the canonical correlation coefficient are analysed. Based on the two relationships, an OLS based feature selection method is developed for binomial classification (Section 3) and multinomial classification (Section 4), respectively. The speed advantage of the method in the greedy search is analysed for both binomial and multinomial classification cases, and the relationship of the proposed

method with LDA is studied in Section 4. In Section Section 5, a detailed example is provided to illustrate the procedure of the proposed method, and its relationship with CCA and LDA. In addition, a comparison of the proposed method with the mutual information based methods is carried out on two synthetic and two real world datasets. Conclusions are summarised in Section 6.

5.2 Squared orthogonal correlation coefficients

5.2.1 Definition

In the ordinary least-squares problem, the linear regression model with N observations is given by

$$\mathbf{y} = (\mathbf{1}, \mathbf{X}) \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} + \mathbf{e}, \tag{5.1}$$

where the response vector is

$$\mathbf{y} = (y_1, \ldots, y_N)^\top, \tag{5.2}$$

the design matrix of n independent variables with a constant term is

$$(\mathbf{1}, \mathbf{X}) = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \ldots & x_{N,n} \end{pmatrix}, \tag{5.3}$$

the estimated parameter vector is

$$\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} = (\beta_0, \beta_1, \ldots, \beta_n)^\top, \tag{5.4}$$

the error term is

$$\mathbf{e} = (e_1, \ldots, e_N)^\top. \tag{5.5}$$

The intercept β_0 satisfies the equation

$$\begin{aligned} \beta_0 &= \bar{\mathbf{y}} - \bar{\mathbf{X}}\boldsymbol{\beta} \\ &= \bar{\mathbf{y}} - (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_n)\, \boldsymbol{\beta}, \end{aligned} \tag{5.6}$$

where $\bar{\mathbf{y}}$ is the sample mean of $\mathbf{y}$, and $\bar{\mathbf{x}}_i$ is the sample mean of $\mathbf{x}_i$. Substituting (5.6) into (5.1), the linear model (5.1) is simplified to

$$\mathbf{y}_{\mathrm{C}} = \mathbf{X}_{\mathrm{C}}\boldsymbol{\beta} + \mathbf{e}, \tag{5.7}$$

where $\mathbf{y}_{\mathrm{C}}$ is the centred response variable given by

$$\mathbf{y}_{\mathrm{C}} = \begin{pmatrix} y_1 - \bar{\mathbf{y}} \\ \vdots \\ y_n - \bar{\mathbf{y}} \end{pmatrix}, \tag{5.8}$$

and $\mathbf{X}_{\mathrm{C}}$ is the matrix of the centred independent variables given by

$$\begin{aligned} \mathbf{X}_{\mathrm{C}} &= \left(\mathbf{x}_{\mathrm{C}1}, \ldots, \mathbf{x}_{\mathrm{C}n}\right) \\ &= \begin{pmatrix} x_{1,1} - \bar{\mathbf{x}}_1 & \ldots & x_{1,n} - \bar{\mathbf{x}}_n \\ \vdots & \ddots & \vdots \\ x_{N,1} - \bar{\mathbf{x}}_1 & \ldots & x_{N,n} - \bar{\mathbf{x}}_n \end{pmatrix}. \end{aligned} \tag{5.9}$$

As the parameter vector $\boldsymbol{\beta}$ satisfies the normal equation

$$\left(\mathbf{X}_{\mathrm{C}}^{\top}\mathbf{X}_{\mathrm{C}}\right)\boldsymbol{\beta} = \mathbf{X}_{\mathrm{C}}^{\top}\mathbf{y}_{\mathrm{C}}, \tag{5.10}$$

the least-squares problem with the intercept is transformed into the least-squares problem without the intercept.

When $\mathbf{X}_{\mathrm{C}}$ has full column rank, the reduced QR decomposition is performed on $\mathbf{X}_{\mathrm{C}}$ as

$$\mathbf{X}_{\mathrm{C}} = \mathbf{W}_{\mathrm{C}}\mathbf{A}, \tag{5.11}$$

where $\mathbf{A}$ is a $n \times n$ invertible upper triangular matrix and $\mathbf{W}_{\mathrm{C}}$ is a $N \times n$ matrix with the orthogonal columns $\mathbf{w}_{\mathrm{C}1}, \ldots, \mathbf{w}_{\mathrm{C}n}$. As $\mathbf{W}_{\mathrm{C}} = \mathbf{X}_{\mathrm{C}}\mathbf{A}^{-1}$, it can be seen that $\mathbf{w}_{\mathrm{C}i}$, which is the linear transformation of $\mathbf{x}_{\mathrm{C}1}, \ldots, \mathbf{x}_{\mathrm{C}n}$, has zero mean. Substituting (5.11) into (5.7), the linear regression model with the same error term $\mathbf{e}$ is given by

$$\mathbf{y}_{\mathrm{C}} = \mathbf{W}_{\mathrm{C}}\mathbf{g} + \mathbf{e}, \tag{5.12}$$

where $\mathbf{g} = \mathbf{A}\boldsymbol{\beta} = (g_1, \ldots, g_n)^{\top}$. The parameter vector $\mathbf{g}$ satisfies the normal equation

$$\mathbf{W}_{\mathrm{C}}^{\top}\mathbf{W}_{\mathrm{C}}\mathbf{g} = \mathbf{W}_{\mathrm{C}}^{\top}\mathbf{y}_{\mathrm{C}}, \tag{5.13}$$

which can be obtained by substituting (5.11) into (5.10). Thus, the ordinary least-squares problem (5.10) about $\mathbf{X}_\mathrm{C}$ and $\mathbf{y}_\mathrm{C}$ is transformed into the OLS problem (5.13) about $\mathbf{W}_\mathrm{C}$ and $\mathbf{y}_\mathrm{C}$.

The residual sum of squares for OLS is given by

$$\begin{aligned}\mathbf{e}^\top\mathbf{e} &= (\mathbf{y}_\mathrm{C} - \mathbf{W}_\mathrm{C}\mathbf{g})^\top (\mathbf{y}_\mathrm{C} - \mathbf{W}_\mathrm{C}\mathbf{g}) \\ &= \mathbf{y}_\mathrm{C}^\top\mathbf{y}_\mathrm{C} - 2\mathbf{g}^\top\mathbf{W}_\mathrm{C}^\top\mathbf{y} + \mathbf{g}^\top\mathbf{W}_\mathrm{C}^\top\mathbf{W}_\mathrm{C}\mathbf{g}.\end{aligned} \tag{5.14}$$

Because of (5.13), this equation becomes

$$\mathbf{e}^\top\mathbf{e} = \mathbf{y}_\mathrm{C}^\top\mathbf{y}_\mathrm{C} - \mathbf{g}^\top\mathbf{W}_\mathrm{C}^\top\mathbf{W}_\mathrm{C}\mathbf{g}. \tag{5.15}$$

As $\mathbf{W}_\mathrm{C}$ is orthogonal, the inner product $\mathbf{W}_\mathrm{C}^\top\mathbf{W}_\mathrm{C}$ is the diagonal matrix $\mathrm{diag}\,(\mathbf{w}_{\mathrm{C}1}^\top\mathbf{w}_{\mathrm{C}1}, \ldots, \mathbf{w}_{\mathrm{C}n}^\top\mathbf{w}_{\mathrm{C}n})$. Thus, (5.15) can be rewritten to

$$\mathbf{e}^\top\mathbf{e} = {\mathbf{y}_\mathrm{C}}^\top\mathbf{y}_\mathrm{C} - \sum_{i=0}^{n} g_i^2\mathbf{w}_{\mathrm{C}i}^\top\mathbf{w}_{\mathrm{C}i}. \tag{5.16}$$

To obtain ERRs, both sides of (5.16) are divided by ${\mathbf{y}_\mathrm{C}}^\top\mathbf{y}_\mathrm{C}$, that is

$$\frac{\mathbf{e}^\top\mathbf{e}}{{\mathbf{y}_\mathrm{C}}^\top\mathbf{y}_\mathrm{C}} = 1 - \sum_{i=1}^{n} \frac{g_i^2\mathbf{w}_{\mathrm{C}i}^\top\mathbf{w}_{\mathrm{C}i}}{{\mathbf{y}_\mathrm{C}}^\top\mathbf{y}_\mathrm{C}}. \tag{5.17}$$

Due to the orthogonality of $\mathbf{W}_\mathrm{C}$, the computation of the parameter vector $\mathbf{g}$ can be simplified as

$$g_i = \frac{\mathbf{w}_{\mathrm{C}i}^\top\mathbf{y}_\mathrm{C}}{\mathbf{w}_{\mathrm{C}i}^\top\mathbf{w}_{\mathrm{C}i}}. \tag{5.18}$$

Substituting (5.18) into (5.17),

$$\begin{aligned}\frac{\mathbf{e}^\top\mathbf{e}}{{\mathbf{y}_\mathrm{C}}^\top\mathbf{y}_\mathrm{C}} &= 1 - \sum_{i=1}^{n} \frac{\mathbf{y}_\mathrm{C}^\top\mathbf{w}_{\mathrm{C}i}\mathbf{w}_{\mathrm{C}i}^\top\mathbf{y}_\mathrm{C}}{\mathbf{w}_{\mathrm{C}i}^\top\mathbf{w}_{\mathrm{C}i}\mathbf{y}_\mathrm{C}^\top\mathbf{y}_\mathrm{C}} \\ &= 1 - \sum_{i=1}^{n} h_i,\end{aligned} \tag{5.19}$$

where h_i is the ERR of $\mathbf{w}_{\mathrm{C}i}$ given by

$$h_i = \frac{\mathbf{y}_\mathrm{C}^\top\mathbf{w}_{\mathrm{C}i}\mathbf{w}_{\mathrm{C}i}^\top\mathbf{y}_\mathrm{C}}{\mathbf{w}_{\mathrm{C}i}^\top\mathbf{w}_{\mathrm{C}i}\mathbf{y}_\mathrm{C}^\top\mathbf{y}_\mathrm{C}}, \qquad i = 1, \ldots, n. \tag{5.20}$$

which is the same as the squared Pearson correlation coefficient between $\mathbf{y}$ and $\mathbf{w}_{\mathrm{C}i}$ [57, p. 29], i.e.

$$r^2(\mathbf{y}, \mathbf{w}_{\mathrm{C}i}) = h_i. \tag{5.21}$$

In the following, we refer to h_i for $i = 1, \ldots, n$ as the *squared orthogonal correlation coefficients* between $\mathbf{X}$ and $\mathbf{y}$.

5.2.2 Relationship with multiple correlation coefficient

The multiple correlation coefficient is the measure of linear association between two or more independent variables and a dependent variable. If the n columns in the design matrix $\mathbf{X}$ are the samples of n independent variables and the response vector $\mathbf{y}$ is the samples of a dependent variable, the association between $\mathbf{X}$ and $\mathbf{y}$ can be measured by the multiple correlation coefficient $R(\mathbf{X}, \mathbf{y})$ or $R(\mathbf{y}, \mathbf{X})$. The multiple correlation analysis of $\mathbf{X}$ and $\mathbf{y}$ is to find a projection direction, so that the Pearson correlation coefficient between $\mathbf{y}_\mathrm{C}$ and the projected $\mathbf{X}_\mathrm{C}$ is maximised. The optimal projection direction is exactly the solution $\boldsymbol{\beta}$ of the normal equation (5.10). Then, the multiple correlation coefficient $R(\mathbf{X}, \mathbf{y})$ or $R(\mathbf{y}, \mathbf{X})$ is defined as

$$R(\mathbf{X}, \mathbf{y}) = R(\mathbf{y}, \mathbf{X}) = r(\hat{\mathbf{y}}_\mathrm{C}, \mathbf{y}_\mathrm{C}) = \frac{\hat{\mathbf{y}}_\mathrm{C}^\top \mathbf{y}_\mathrm{C}}{\sqrt{\hat{\mathbf{y}}_\mathrm{C}^\top \hat{\mathbf{y}}_\mathrm{C}} \sqrt{\mathbf{y}_\mathrm{C}^\top \mathbf{y}_\mathrm{C}}}, \tag{5.22}$$

where

$$\hat{\mathbf{y}}_\mathrm{C} = \mathbf{X}_\mathrm{C} \boldsymbol{\beta}. \tag{5.23}$$

It can be seen that the definition of the multiple correlation coefficient is based on the centred linear regression model (5.7). The squared multiple correlation coefficient (or called coefficient of determination) has the following relationship with the total sum of squares SST and the residual sum of squares SSR of the model (5.7)

$$R^2(\mathbf{X}, \mathbf{y}) = 1 - \frac{SSR(\mathbf{X}_\mathrm{C}, \mathbf{y}_\mathrm{C})}{SST(\mathbf{X}_\mathrm{C}, \mathbf{y}_\mathrm{C})}, \tag{5.24}$$

where

$$\begin{aligned} SST(\mathbf{X}_\mathrm{C}, \mathbf{y}_\mathrm{C}) &= \mathbf{y}_\mathrm{C}^\top \mathbf{y}_\mathrm{C} \\ SSR(\mathbf{X}_\mathrm{C}, \mathbf{y}_\mathrm{C}) &= (\mathbf{y}_\mathrm{C} - \hat{\mathbf{y}}_\mathrm{C})^\top (\mathbf{y}_\mathrm{C} - \hat{\mathbf{y}}_\mathrm{C}) = \mathbf{e}^\top \mathbf{e}. \end{aligned} \tag{5.25}$$

Comparing (5.19) and (5.24), it is found

$$R^2(\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} h_i. \tag{5.26}$$

Therefore, the sum of the squared orthogonal correlation coefficients between $\mathbf{X}$ and $\mathbf{y}$ is equal to the squared multiple correlation coefficient between $\mathbf{X}$ and $\mathbf{y}$.

5.2.3 Relationship with canonical correlation coefficient

The canonical correlation coefficient is the measure of linear association between two or more independent variables and two or more dependent variables. Given a response matrix as

$$\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m) = \begin{pmatrix} y_{1,1} & \cdots & y_{1,m} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,m} \end{pmatrix}, \tag{5.27}$$

if the n columns in the design matrix $\mathbf{X}$ are the samples of n independent variables and the m columns in the response matrix $\mathbf{Y}$ are the samples of m dependent variables, the association between $\mathbf{X}$ and $\mathbf{Y}$ can be measured by the canonical correlation coefficient $R(\mathbf{X}, \mathbf{Y})$. The Canonical Correlation Analysis (CCA) for $\mathbf{X}$ and $\mathbf{Y}$ is to find a pair of the projection directions $\mathbf{a}$ and $\mathbf{b}$, so that the Pearson correlation coefficient between $\mathbf{X}_{\mathrm{C}}\mathbf{a}$ and $\mathbf{Y}_{\mathrm{C}}\mathbf{b}$ is maximised, that is

$$\arg\max_{\mathbf{a},\mathbf{b}} r(\mathbf{X}_{\mathrm{C}}\mathbf{a}, \mathbf{Y}_{\mathrm{C}}\mathbf{b}), \tag{5.28}$$

where

$$\begin{aligned} \mathbf{Y}_{\mathrm{C}} &= \left(\mathbf{y}_{\mathrm{C}1}, \ldots, \mathbf{y}_{\mathrm{C}m}\right) \\ &= \begin{pmatrix} y_{1,1} - \bar{\mathbf{y}}_1 & \cdots & y_{1,m} - \bar{\mathbf{y}}_m \\ \vdots & \ddots & \vdots \\ y_{N,1} - \bar{\mathbf{y}}_1 & \cdots & y_{N,m} - \bar{\mathbf{y}}_m \end{pmatrix}, \end{aligned} \tag{5.29}$$

and $\bar{\mathbf{y}}_i$ is the sample mean of $\mathbf{y}_i$. The canonical correlation coefficient between $\mathbf{X}$ and $\mathbf{Y}$ can be computed by

$$R(\mathbf{X}, \mathbf{Y}) = r(\mathbf{X}_{\mathrm{C}}\mathbf{a}, \mathbf{Y}_{\mathrm{C}}\mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{R}_{\mathbf{X},\mathbf{Y}} \mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{R}_{\mathbf{X},\mathbf{X}} \mathbf{a}} \sqrt{\mathbf{b}^\top \mathbf{R}_{\mathbf{Y},\mathbf{Y}} \mathbf{b}}}, \tag{5.30}$$

where the correlation matrices are given by

$$\begin{aligned}
\mathbf{R}_{\mathbf{X},\mathbf{Y}} &= \begin{pmatrix} r_{\mathbf{x}_1,\mathbf{y}_1} & \cdots & r_{\mathbf{x}_1,\mathbf{y}_m} \\ \vdots & \ddots & \vdots \\ r_{\mathbf{x}_n,\mathbf{y}_1} & \cdots & r_{\mathbf{x}_n,\mathbf{y}_m} \end{pmatrix} \\
\mathbf{R}_{\mathbf{X},\mathbf{X}} &= \begin{pmatrix} r_{\mathbf{x}_1,\mathbf{x}_1} & \cdots & r_{\mathbf{x}_1,\mathbf{x}_n} \\ \vdots & \ddots & \vdots \\ r_{\mathbf{x}_n,\mathbf{x}_1} & \cdots & r_{\mathbf{x}_n,\mathbf{x}_n} \end{pmatrix} \\
\mathbf{R}_{\mathbf{Y},\mathbf{Y}} &= \begin{pmatrix} r_{\mathbf{y}_1,\mathbf{y}_1} & \cdots & r_{\mathbf{y}_1,\mathbf{y}_m} \\ \vdots & \ddots & \vdots \\ r_{\mathbf{y}_m,\mathbf{y}_1} & \cdots & r_{\mathbf{y}_m,\mathbf{y}_m} \end{pmatrix}.
\end{aligned} \tag{5.31}$$

The multiple correlation coefficient $R(\mathbf{X}, \mathbf{y})$ is a special case of the canonical correlation coefficient $R(\mathbf{X}, \mathbf{Y})$, when $\mathbf{Y}$ is a column vector $\mathbf{y}$. The CCA can be transformed to the eigenvalue problem given by [76, p. 173]

$$\mathbf{R}_{\mathbf{X},\mathbf{X}}^{-1}\mathbf{R}_{\mathbf{X},\mathbf{Y}}\mathbf{R}_{\mathbf{Y},\mathbf{Y}}^{-1}\mathbf{R}_{\mathbf{Y},\mathbf{X}}\mathbf{a} = R^2(\mathbf{X}, \mathbf{Y})\mathbf{a} \tag{5.32a}$$

$$\mathbf{R}_{\mathbf{Y},\mathbf{Y}}^{-1}\mathbf{R}_{\mathbf{Y},\mathbf{X}}\mathbf{R}_{\mathbf{X},\mathbf{X}}^{-1}\mathbf{R}_{\mathbf{X},\mathbf{Y}}\mathbf{b} = R^2(\mathbf{X}, \mathbf{Y})\mathbf{b}. \tag{5.32b}$$

The two projection directions $\mathbf{a}$ and $\mathbf{b}$ are the eigenvectors, and the eigenvalue is the square of the canonical correlation coefficient. If $\mathbf{X}_\mathrm{C}$ and $\mathbf{Y}_\mathrm{C}$ have full column rank, the number of the non-zero solutions of (5.32) is not more than $n \wedge m$, where the operator $\wedge$ returns the minimum of two values on both sides. Thus, in contrast with the multiple correlation coefficient which only has one value, there are $n \wedge m$ canonical correlation coefficients (which may contain zeros) for $\mathbf{X}$ and $\mathbf{Y}$, which are denoted as $R_1(\mathbf{X}, \mathbf{Y}), \ldots, R_{n\wedge m}(\mathbf{X}, \mathbf{Y})$.

It is known that the multiple correlation between $\mathbf{Y}$ and each $\mathbf{x}_i$ can be evaluated by [76, p. 174]

$$\begin{pmatrix} R^2(\mathbf{x}_1, \mathbf{Y}) \\ \vdots \\ R^2(\mathbf{x}_n, \mathbf{Y}) \end{pmatrix} = \mathrm{diag}\left(\mathbf{R}_{\mathbf{X},\mathbf{Y}}\mathbf{R}_{\mathbf{Y},\mathbf{Y}}^{-1}\mathbf{R}_{\mathbf{Y},\mathbf{X}}\right), \tag{5.33}$$

where the operator diag obtains the main diagonal of the matrix. According to (5.32a), the sum of the squared canonical correlation coefficients, i.e. the sum of the eigenvalues, are

given by

$$\sum_{k=1}^{n\wedge m} R_k^2(\mathbf{X},\mathbf{Y}) = \mathrm{tr}\left(\mathbf{R}_{\mathbf{X},\mathbf{X}}^{-1}\mathbf{R}_{\mathbf{X},\mathbf{Y}}\mathbf{R}_{\mathbf{Y},\mathbf{Y}}^{-1}\mathbf{R}_{\mathbf{Y},\mathbf{X}}\right), \tag{5.34}$$

where the operator tr obtains the trace of the matrix. If the columns of $\mathbf{X}$ are zero mean and orthogonal, the correlation matrix of $\mathbf{X}$ is identity matrix, so $\mathbf{R}_{\mathbf{X},\mathbf{X}}^{-1} = \mathbf{I}$. Therefore, according to (5.33) and (5.34), the following equation holds when the columns of $\mathbf{X}$ are centred and orthogonal.

$$\sum_{k=1}^{n\wedge m} R_k^2(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n} R^2(\mathbf{x}_i,\mathbf{Y}). \tag{5.35}$$

Through the reduced QR decomposition,

$$\begin{aligned} \mathbf{X}_\mathrm{C} &= \mathbf{W}_\mathrm{C}\mathbf{A} \\ \mathbf{Y}_\mathrm{C} &= \mathbf{V}_\mathrm{C}\mathbf{B} \end{aligned} \tag{5.36}$$

where $\mathbf{W}_\mathrm{C}$ is a $N \times n$ matrix with the centred orthogonal columns given by

$$\mathbf{W}_\mathrm{C} = (\mathbf{w}_{\mathrm{C}1},\ldots,\mathbf{w}_{\mathrm{C}n}), \tag{5.37}$$

$\mathbf{V}_\mathrm{C}$ is a $N \times m$ matrix with the centred orthogonal columns given by

$$\mathbf{V}_\mathrm{C} = (\mathbf{v}_{\mathrm{C}1},\ldots,\mathbf{v}_{\mathrm{C}m}), \tag{5.38}$$

$\mathbf{A}$ is a $n \times n$ invertible upper triangular matrix, and $\mathbf{B}$ is a $m \times m$ invertible upper triangular matrix. It is noticed that the transformation from $\mathbf{X}$ (or $\mathbf{Y}$) to $\mathbf{W}_\mathrm{C}$ (or $\mathbf{V}_\mathrm{C}$) is affine. As the canonical correlation coefficient is invariant under affine transformations,

$$R_k(\mathbf{X},\mathbf{Y}) = R_k(\mathbf{W}_\mathrm{C},\mathbf{V}_\mathrm{C}) \qquad k = 1,\ldots,n\wedge m. \tag{5.39}$$

As the columns of $\mathbf{W}_\mathrm{C}$ are centred and orthogonal, the following equation holds according to (5.35) and (5.39).

$$\sum_{k=1}^{n\wedge m} R_k^2(\mathbf{X},\mathbf{Y}) = \sum_{k=1}^{n\wedge m} R_k^2(\mathbf{W}_\mathrm{C},\mathbf{V}_\mathrm{C}) = \sum_{i=1}^{n} R^2(\mathbf{w}_{\mathrm{C}i},\mathbf{V}_\mathrm{C}). \tag{5.40}$$

Define squared orthogonal correlation matrix as

$$\mathbf{H} = \begin{pmatrix} h_{1,1} & \ldots & h_{1,m} \\ \vdots & \ddots & \vdots \\ h_{n,1} & \ldots & h_{n,m} \end{pmatrix}, \tag{5.41}$$

where

$$h_{i,j} = \frac{\mathbf{v}_{\mathrm{C}j}^{\top}\mathbf{w}_{\mathrm{C}i}\mathbf{w}_{\mathrm{C}i}^{\top}\mathbf{v}_{\mathrm{C}j}}{\mathbf{w}_{\mathrm{C}i}^{\top}\mathbf{w}_{\mathrm{C}i}\mathbf{v}_{\mathrm{C}j}^{\top}\mathbf{v}_{\mathrm{C}j}}. \tag{5.42}$$

Due to (5.26), the multiple correlation coefficient between $\mathbf{V}_{\mathrm{C}}$ and each $\mathbf{w}_{\mathrm{C}}$ can be evaluated by

$$\begin{aligned} R^2(\mathbf{w}_{\mathrm{C}1}, \mathbf{V}_{\mathrm{C}}) &= \sum_{j=1}^{m} h_{1,j} \\ &\vdots \\ R^2(\mathbf{w}_{\mathrm{C}n}, \mathbf{V}_{\mathrm{C}}) &= \sum_{j=1}^{m} h_{n,j}. \end{aligned} \tag{5.43}$$

Substituting (5.43) into (5.40), it is found that the sum of the squared canonical correlation coefficients between $\mathbf{X}$ and $\mathbf{Y}$ is equal to the sum of all entries of the squared orthogonal correlation matrix $\mathbf{H}$, that is

$$\sum_{k=1}^{n\wedge m} R_k^2(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} h_{i,j}, \tag{5.44}$$

which is a natural extension of (5.26) to the case where the response vector $\mathbf{y}$ becomes the response matrix $\mathbf{Y}$.

5.3 OLS based fast feature selection for binomial classification

If the N observations of $\mathbf{X}$ belong to two classes and the n variables in $\mathbf{X}$ represent n features, the feature selection problem for the binomial classification is to find the t features from the n features of $\mathbf{X}$, which is optimal to classify the N observations into the two classes. The two classes can be assigned values 0 and 1 to form a dummy response vector $\mathbf{y}$ for the N observations. Thus, the goodness of the features for the classification can be evaluated by the multiple correlation coefficient between the features of interest and the dummy response vector. In fact, the two classes can be assigned to any distinct values to form the response vector, which has no effect on the value of the multiple correlation coefficient. However, to be consistent to the multinomial classification case, the dummy encoding is adopted. The

optimal t features can be searched by comparing all $\binom{n}{t}$ feature combinations exhaustively, where

$$\binom{n}{t} = \frac{n!}{t!(n-t)!}. \tag{5.45}$$

In some cases, the *exhaustive search* is too expensive in computation. A realistic approach is to select only one feature in one step. In each step, the previously selected features will not be changed. For example, the three 'optimal' features can be selected in three steps as shown in Table 5.1. As each step selects the feature which maximises the multiple correlation, the search is referred to the *greedy search* [50].

Table 5.1: An example for selecting three features from n features by the greedy search for binomial classification, where $i = 1, \ldots, n$ for step 1, $i = 1, 2, 4, 5, \ldots, n$ for step 2, $i = 1, 2, 4, 6, 7, \ldots, n$ for step 3.

	Multiple Correlation	Selected Feature
Step 1	$R(\mathbf{x}_3, \mathbf{y}) \geq R(\mathbf{x}_i, \mathbf{y})$	$\mathbf{x}_3$
Step 2	$R((\mathbf{x}_3, \mathbf{x}_5), \mathbf{y}) \geq R((\mathbf{x}_3, \mathbf{x}_i), \mathbf{y})$	$\mathbf{x}_3, \mathbf{x}_5$
Step 3	$R((\mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_1), \mathbf{y}) \geq R((\mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_i), \mathbf{y})$	$\mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_1$

The multiple correlation coefficient can be obtained either using the definition (5.22) or the sum of the squared orthogonal correlation coefficients (5.26). In the greedy search, the OLS based feature selection method has the computational speed advantage over the definition based feature selection method. The computation complexity of the two feature selection methods can be explicitly compared by the asymptotic upper bound notation $\mathcal{O}$ [77, p. 47]. At Step k of the greedy search, the $k-1$ optimal features have been selected, and the rest of the $n-k+1$ features are the candidates of the k^{th} optimal feature. The candidate feature matrix is a $N \times k$ matrix composed of the $k-1$ selected features and a candidate feature. According to the normal equation (5.10), the definition based feature selection method is dominated by computing the inner product of the $N \times k$ centred candidate feature matrix. The computational complexity of the inner product of one centred candidate matrix is $\mathcal{O}(k^2 N)$. There are $n-k+1$ candidate features, so the complexity for Step k is

$$(n-k+1)\mathcal{O}(k^2 N) = \mathcal{O}(k^2 n N). \tag{5.46}$$

Thus, the overall complexity for t features selection is given by

$$\sum_{k=1}^{t} \mathcal{O}(k^2 nN) = \mathcal{O}\left(\sum_{k=1}^{t} k^2 nN\right) = \mathcal{O}(t^3 nN). \tag{5.47}$$

For OLS based feature selection, as the squared orthogonal correlation coefficients of the selected features (h_1 to h_{k-1}) have been computed in Step 1 to Step $k-1$, only the squared orthogonal correlation coefficients of the candidate feature (h_k) is required to compute. Thus, OLS based feature selection is dominated by the classical Gram-Schmidt orthogonalisation process. At Step k of the greedy search, the computational complexity of the orthogonalisation of one candidate feature is $\mathcal{O}(kN)$. There are $n-k+1$ candidate features, so the complexity for Step k is

$$(n-k+1)\mathcal{O}(kN) = \mathcal{O}(knN). \tag{5.48}$$

Thus, the overall complexity for t features selection is given by

$$\sum_{k=1}^{t} \mathcal{O}(knN) = \mathcal{O}\left(\sum_{k=1}^{t} knN\right) = \mathcal{O}(t^2 nN). \tag{5.49}$$

Consequently, compared to the definition based feature selection method, the OLS based feature selection method has a significant computational speed advantage in the greedy search.

5.4 OLS based fast feature selection for multinomial classification

If the N observations of $\mathbf{X}$ belong to c classes, where $c \leq N$, and the n columns in $\mathbf{X}$ represent n features, the feature selection problem for the multinomial classification is to find the t features from the n features of $\mathbf{X}$, which is optimal to classify the N observations into the c classes. Similar to the last section, the c classes can be encoded to certain values to form a response variable. The ordinal encoding is to assign $1, \ldots, c$ to the c labels to form a vector $\mathbf{y}$. Then, the multiple correlation coefficient between the features and $\mathbf{y}$ can be adopted to indicate the goodness of the features for the classification. The c labels can also be encoded to form a matrix $\mathbf{Y}$, e.g. c-label dummy encoding (or called one-hot encoding), $c-1$-label dummy encoding, effects encoding, and contrast encoding [57, Chapter 5]. When

the response is encoded as a matrix $\mathbf{Y}$, the canonical correlation coefficients between $\mathbf{X}$ and $\mathbf{Y}$ can be used as the feature selection criterion. Similar to the last section, an example of the greed search for multinomial classification is illustrated in Table 5.2, where the response is encoded as an $N \times c - 1$ matrix $\mathbf{Y}$ and the ranking criterion is the sum of the squared canonical correlation coefficients.

Table 5.2: An example for selecting three features from n features by the greedy search for multinomial classification, where $i = 1, \ldots, n$ for step 1, $i = 1, 2, 4, 5, \ldots, n$ for step 2, $i = 1, 2, 4, 6, 7, \ldots, n$ for step 3.

	Ranking Criterion	Selected Feature
1	$\sum_{k=1}^{1\wedge c-1} R_k^2(\mathbf{x}_3, \mathbf{Y}) \geq \sum_{k=1}^{1\wedge c-1} R_k^2(\mathbf{x}_i, \mathbf{Y})$	$\mathbf{x}_3$
2	$\sum_{k=1}^{2\wedge c-1} R_k((\mathbf{x}_3, \mathbf{x}_5), \mathbf{Y}) \geq \sum_{k=1}^{2\wedge c-1} R_k((\mathbf{x}_3, \mathbf{x}_i), \mathbf{Y})$	$\mathbf{x}_3, \mathbf{x}_5$
3	$\sum_{k=1}^{3\wedge c-1} R_k((\mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_1), \mathbf{Y}) \geq \sum_{k=1}^{3\wedge c-1} R_k((\mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_i), \mathbf{Y})$	$\mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_1$

In the following, the equivalence between CCA and a classical classifier linear discriminant analysis will first be demonstrated to reveal the implication of the canonical correlation coefficient in a classification problem. Then, the algorithm of the OLS based feature selection for multinomial classification is developed where the sum of the squared canonical correlation coefficients will be used as the feature ranking criterion. After that a version of the algorithm that can be used to deal with categorical features is presented.

5.4.1 Relationship with linear discriminant analysis

As the feature selection is used for the multinomial classification, it is reasonable to know the performance of the features in the Linear Discriminant Analysis (LDA), where the label encoding is not required. For the convenience of LDA, the feature matrix $\mathbf{X}$ is decomposed into $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(c)}$, where the $N_i \times n$ matrix $\mathbf{X}^{(i)}$ represents the N_i observations belonged to

the i^{th} class. The within-class scatter matrix of the samples is

$$\mathbf{S}_w = \sum_{i=1}^{c} \left(\mathbf{X}^{(i)} - \mathbf{1}^{(i)}\bar{\mathbf{X}}^{(i)}\right)^\top \left(\mathbf{X}^{(i)} - \mathbf{1}^{(i)}\bar{\mathbf{X}}^{(i)}\right), \tag{5.50}$$

where $\bar{\mathbf{X}}^{(i)}$ is the sample mean of each feature in $\mathbf{X}^{(i)}$ given by

$$\bar{\mathbf{X}}^{(i)} = \left(\bar{\mathbf{x}}_1^{(i)}, \ldots, \bar{\mathbf{x}}_n^{(i)}\right) \tag{5.51}$$

and $\mathbf{1}^i$ is $N_i \times 1$ vector of ones. The between-class scatter matrix of the samples is

$$\mathbf{S}_b = \sum_{i=1}^{c} N_i \left(\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}}\right)^\top \left(\bar{\mathbf{X}}^{(i)} - \bar{\mathbf{X}}\right), \tag{5.52}$$

where $\bar{\mathbf{X}}$ is the overall sample mean of each feature. The aim of LDA is to find a projection direction $\mathbf{d}$ for $\mathbf{X}$, so that the ratio between the projected between-class scatter and the projected within-class scatter is maximised. The ratio is called Fisher's criterion, which is given by

$$J = \frac{\mathbf{d}^\top \mathbf{S}_b \mathbf{d}}{\mathbf{d}^\top \mathbf{S}_w \mathbf{d}}. \tag{5.53}$$

The larger Fisher's criterion J implies the better the separation of the c classes. The LDA can be transformed to the eigenvalue problem given by [76, p. 246]

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{d} = J\mathbf{d}, \tag{5.54}$$

where the eigenvector is the optimal projection direction $\mathbf{d}$ and the eigenvalue is the maximised Fisher's criterion J.

The relationship between LDA and CCA can be found when $\mathbf{Y}$ is formed by c or $c-1$-label dummy encoding. Under the two encoding schemes, the eigenvalue problem (5.32a) can be rewritten as [78]

$$\left(\mathbf{S}_b + \mathbf{S}_w\right)^{-1} \mathbf{S}_b\mathbf{a} = R^2(\mathbf{X}, \mathbf{Y})\mathbf{a}, \tag{5.55}$$

or in the form of

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{a} = \frac{R^2(\mathbf{X}, \mathbf{Y})}{1 - R^2(\mathbf{X}, \mathbf{Y})}\mathbf{a}. \tag{5.56}$$

Comparing (5.54) and (5.56), it is found that LDA and CCA are equivalent, and Fisher's criterion of LDA can be evaluated by

$$J = \frac{R^2(\mathbf{X}, \mathbf{Y})}{1 - R^2(\mathbf{X}, \mathbf{Y})}. \tag{5.57}$$

To connect the feature selection criteria with LDA, the $c-1$-label dummy encoding and the canonical correlation coefficients are adopted, instead of the ordinal encoding and the multiple correlation coefficient. The $c-1$-label dummy encoding constructs a $N \times c-1$ matrix $Y = (y_{i,j})$, where

$$y_{i,j} = \begin{cases} 1 & i^{\text{th}} \text{ observation is belonged to } j^{\text{th}} \text{ class} \\ 0 & \text{otherwise.} \end{cases} \tag{5.58}$$

Thus, the dummy response in the last section is a special case of $c-1$-label dummy encoding where $c = 2$.

5.4.2 OLS based feature selection algorithm

The algorithm of OLS based feature selection for multinomial classification can be summarised in 5 steps as follows.

Input:

$\mathbf{X}$: $N \times n$ matrix containing N observations and n features.

$\mathbf{Y}$: $N \times c-1$ matrix formed by $c-1$-label dummy encoding.

t: The number of features is to be selected.

Step 1. First, centre $\mathbf{Y}$ into $\mathbf{Y}_{\text{C}}$. Second, orthogonalise $\mathbf{Y}_{\text{C}}$ into $\mathbf{V}_{\text{C}}$. Third, centre $\mathbf{X}$ into $\mathbf{X}_{\text{C}}$.

Step 2. Divide $\mathbf{X}$ into $(\mathbf{X}_{\text{s}}, \mathbf{X}_{\text{r}})$, where the selected feature matrix is given by

$$\mathbf{X}_{\text{s}} = (\mathbf{x}_{\text{s}1}, \ldots, \mathbf{x}_{\text{s}p}), \tag{5.59}$$

and the rest feature matrix is given by

$$\mathbf{X}_{\text{r}} = (\mathbf{x}_{\text{r}1}, \ldots, \mathbf{x}_{\text{r}q}), \tag{5.60}$$

p is the number of the selected features, and q is the number of the rest features. Correspondingly, divide $\mathbf{X}_{\text{C}}$ into $(\mathbf{X}_{\text{Cs}}, \mathbf{X}_{\text{Cr}})$, where

$$\begin{aligned} \mathbf{X}_{\text{Cs}} &= (\mathbf{x}_{\text{Cs}1}, \ldots, \mathbf{x}_{\text{Cs}p}) \\ \mathbf{X}_{\text{Cr}} &= (\mathbf{x}_{\text{Cr}1}, \ldots, \mathbf{x}_{\text{Cr}q}). \end{aligned} \tag{5.61}$$

Step 3. If no feature has been selected (i.e. $p = 0$), let

$$\begin{aligned} \mathbf{W}_{\text{Cr}} &= \mathbf{X}_{\text{Cr}} \\ \mathbf{w}_{\text{Cr}i} &= \mathbf{x}_{\text{Cr}i}, \quad i = 1, \ldots, q. \end{aligned} \tag{5.62}$$

If $p \neq 0$, first, orthogonalise $\mathbf{X}_{\text{Cs}}$ into $\mathbf{W}_{\text{Cs}}$, where

$$\mathbf{W}_{\text{Cs}} = (\mathbf{w}_{\text{Cs}1}, \ldots, \mathbf{w}_{\text{Cs}p}), \tag{5.63}$$

and $\mathbf{w}_{\text{Cs}i}^{\top}\mathbf{w}_{\text{Cs}j} = 0$ for $i \neq j$. Second, orthogonalise each feature in $\mathbf{X}_{\text{C}r}$ to $\mathbf{W}_{\text{Cs}}$ to form the matrix $\mathbf{W}_{\text{Cr}}$, where

$$\mathbf{W}_{\text{Cr}} = (\mathbf{w}_{\text{Cr}1}, \ldots, \mathbf{w}_{\text{Cr}q}), \tag{5.64}$$

and $\mathbf{w}_{\text{Cr}i}$ is obtained through the classical Gram-Schmidt process, which is given by

$$\mathbf{w}_{\text{Cr}i} = \mathbf{x}_{\text{Cr}i} - \sum_{j=1}^{p} \frac{\mathbf{x}_{\text{Cr}i}^{\top}\mathbf{w}_{\text{Cs}j}}{\mathbf{w}_{\text{Cs}j}^{\top}\mathbf{w}_{\text{Cs}j}}\mathbf{w}_{\text{Cs}j}, \quad i = 1, \ldots, q. \tag{5.65}$$

It should be noticed that $\mathbf{w}_{\text{Cr}i}$ is orthogonal to $\mathbf{W}_{\text{Cs}}$ but not to $\mathbf{W}_{\text{Cr}}$, that is $\mathbf{w}_{\text{Cr}i}^{\top}\mathbf{w}_{\text{Cs}j} = 0$ but $\mathbf{w}_{\text{Cr}i}^{\top}\mathbf{w}_{\text{Cr}j} \neq 0$.

Step 4. Compute $R^2(\mathbf{w}_{\text{Cr}i}, \mathbf{V}_{\text{C}})$ by

$$R^2(\mathbf{w}_{\text{Cr}i}, \mathbf{V}_{\text{C}}) = \sum_{j=1}^{c-1} h_{i,j}, \quad i = 1, \ldots, q, \tag{5.66}$$

where

$$h_{i,j} = \frac{\mathbf{v}_{\text{C}j}^{\top}\mathbf{w}_{\text{Cr}i}\mathbf{w}_{\text{Cr}i}^{\top}\mathbf{v}_{\text{C}j}}{\mathbf{w}_{\text{Cr}i}^{\top}\mathbf{w}_{\text{Cr}i}\mathbf{v}_{\text{C}j}^{\top}\mathbf{v}_{\text{C}j}}. \tag{5.67}$$

Step 5. First, find an i which maximises $R^2(\mathbf{w}_{\text{Cr}i}, \mathbf{V}_{\text{C}})$, that is

$$i_{\max} = \arg\max_{i} R^2(\mathbf{w}_{\text{Cr}i}, \mathbf{V}_{\text{C}}). \tag{5.68}$$

Second, remove $\mathbf{x}_{i_{\max}}$ from $\mathbf{X}_{\text{r}}$ and add it into $\mathbf{X}_{\text{s}}$. Correspondingly, q reduces by 1 and p increases by 1. Repeat **Step 2** to **Step 5**, until $p = t$.

Output $\mathbf{X}_{\text{s}}$ to complete the feature selection.

The speed advantage of the OLS based feature selection method is shown in **Step 4**. To evaluate the goodness of the candidate feature $\mathbf{x}_{\mathrm{r}i}$, CCA requires to compute the canonical correlation coefficient $R((\mathbf{X}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}i}), \mathbf{Y})$, while OLS only needs to compute the multiple correlation coefficient $R(\mathbf{w}_{\mathrm{Cr}i}, \mathbf{V}_{\mathrm{C}})$, because

$$\begin{aligned}\sum_{k=1}^{p+1\wedge c-1} R_k^2((\mathbf{X}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}i}), \mathbf{Y}) &= \sum_{k=1}^{p+1\wedge c-1} R_k^2((\mathbf{W}_{\mathrm{Cs}}, \mathbf{w}_{\mathrm{Cr}i}), \mathbf{V}_{\mathrm{C}}) \\ &= \sum_{k=1}^{p\wedge c-1} R_k^2(\mathbf{W}_{\mathrm{Cs}}, \mathbf{V}_{\mathrm{C}}) + R^2(\mathbf{w}_{\mathrm{Cr}i}, \mathbf{V}_{\mathrm{C}}).\end{aligned} \tag{5.69}$$

For each candidate feature $\mathbf{x}_{\mathrm{r}i}$, $R(\mathbf{W}_{\mathrm{Cs}}, \mathbf{V}_{\mathrm{C}})$ is the same. Thus, to find the maximal $R((\mathbf{X}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}i}), \mathbf{Y})$, only $R(\mathbf{w}_{\mathrm{Cr}i}, \mathbf{V}_{\mathrm{C}})$ is required to compute. In addition, although the multiple correlation coefficient $R(\mathbf{w}_{\mathrm{C}i}, \mathbf{Y})$, which is equal to $R(\mathbf{w}_{\mathrm{C}i}, \mathbf{V}_{\mathrm{C}})$, can be computed through the definition (5.22), OLS provides a faster way of computation. Equation (5.22) requires to solve the normal equation which is dominated by the inner product of the $N \times c-1$ matrix $\mathbf{Y}_{\mathrm{C}}$, whose computational complexity is $\mathcal{O}(c^2 N)$. For OLS, as the orthogonalisation of $\mathbf{Y}$ is only required once in **Step 1**, the dominant part is from the computation of (5.66) whose computational complexity is only $\mathcal{O}(cN)$.

As the above introduction of the OLS based algorithm is conceptual, some speed optimisation steps have been omitted. For example, in **Step 3**, $\mathbf{W}_{\mathbf{Cs}}$ computed for selecting the i^{th} optimal feature can be reused for selecting the $i + 1^{\text{th}}$ optimal feature. The further optimisation of the OLS speed can be found in the original paper of OLS based model term selection [49].

5.4.3 Dealing with categorical features

When the features are categorical, the feature encoding is required for OLS based feature selection. In the previous analysis, n features are represented by n column vectors in $\mathbf{X}$, but some encoding methods may encode the categorical features into matrices. In that case, the feature matrix is composed of n submatrices, that is

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n), \tag{5.70}$$

where the matrix $\mathbf{X}_i$ is the encoded i^{th} feature. An OLS based feature selection algorithm similar to the algorithm in Section 5.4.2 can be applied to the matrix encoded features. In this case, the candidate orthogonal feature matrix in **Step 3** of Section 5.4.2 is given by

$$\mathbf{W}_{\text{Cr}} = (\mathbf{W}_{\text{Cr}1}, \ldots, \mathbf{W}_{\text{Cr}q}), \tag{5.71}$$

where $\mathbf{W}_{\text{Cr}i}$ is a $N \times z_i$ matrix given by

$$\mathbf{W}_{\text{Cr}i} = \left(\mathbf{w}_{\text{Cr}i}^{[1]}, \ldots, \mathbf{w}_{\text{Cr}i}^{[z_i]}\right). \tag{5.72}$$

Besides being orthogonal to the selected orthogonal feature matrix $\mathbf{W}_{\text{Cs}}$, the submatrix $\mathbf{W}_{\text{Cr}i}$ should be column-wise orthogonal via an additional orthogonalisation process. In **Step 4**, the sum of the squared canonical correlation coefficients can be computed by

$$\sum_{k=1}^{z_i \wedge c-1} R_k^2(\mathbf{W}_{\text{Cr}i}, \mathbf{V}_{\text{C}}) = \sum_{j=1}^{c-1} \sum_{g=1}^{z_i} h_{i,j}^{[g]}, \quad i = 1, \ldots, q, \tag{5.73}$$

where

$$h_{i,j}^{[g]} = \frac{\mathbf{v}_{\text{C}j}^{\top} \mathbf{w}_{\text{Cr}i}^{[g]} \mathbf{w}_{\text{Cr}i}^{[g]\top} \mathbf{v}_{\text{C}j}}{\mathbf{w}_{\text{Cr}i}^{[g]\top} \mathbf{w}_{\text{Cr}i}^{[g]} \mathbf{v}_{\text{C}j}^{\top} \mathbf{v}_{\text{C}j}}. \tag{5.74}$$

Finally, the sum of the squared canonical correlation coefficients are used to rank the features for **Step 5**.

5.5 Empirical study

In this section, firstly, a simple example is used to illustrate the procedure of the OLS based feature selection method when applied to the Fisher's iris data [79]. The equivalence between the squared orthogonal correlation coefficients, canonical correlation coefficient, and Fisher's criterion is also demonstrated via this case study. Then, the OLS based feature selection methods are compared with mutual information based feature selection methods using both synthetic and real world datasets (i.e. Dexter and Gisette). The OLS method takes 401 ms for Dexter and 5109 ms for Gisette to select 20 features on a 2.6 GHz personal laptop, while the traditional definition based method takes 13200 ms and 134922 ms, respectively. The empirical studies are implemented in MATLAB R2019b, and the code will be published in GitHub[1].

[1] `https://github.com/MatthewSZhang/fs_ols`

5.5.1 An illustration of the OLS based feature selection

Table 5.3: Fisher's Iris Dataset.

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3	5.9	2.1	virginica

The Fisher's iris data are given in Table 5.3. The 7 observations have 4 features and 3 classes, so $N = 7$, $n = 4$, and $c = 3$. The objective of the feature selection is to find 3 optimal features for the 3 species classification.

The feature matrix is given by

$$\mathbf{X} = \begin{pmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3 & 1.4 & 0.2 \\ 7 & 3.2 & 4.7 & 1.4 \\ 6.4 & 3.2 & 4.5 & 1.5 \\ 6.3 & 3.3 & 6 & 2.5 \\ 5.8 & 2.7 & 5.1 & 1.9 \\ 7.1 & 3 & 5.9 & 2.1 \end{pmatrix}, \tag{5.75}$$

and the $c-1$-label dummy encoded response is

$$\mathbf{Y} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \tag{5.76}$$

where $(1,0)$ represents setosa, $(0,1)$ represents versicolor, and $(0,0)$ represents virginica. Following the algorithm introduced in Section 5.4.2, the procedure of the OLS based feature selection method is shown below.

Step 1. First, centre $\mathbf{Y}$ into $\mathbf{Y}_{\mathrm{C}}$, which is given by

$$\mathbf{Y}_{\mathrm{C}} = \begin{pmatrix} 0.7143 & -0.2857 \\ 0.7143 & -0.2857 \\ -0.2857 & 0.7143 \\ -0.2857 & 0.7143 \\ -0.2857 & -0.2857 \\ -0.2857 & -0.2857 \\ -0.2857 & -0.2857 \end{pmatrix}. \tag{5.77}$$

Second, orthogonalise $\mathbf{Y}_{\mathrm{C}}$ into $\mathbf{V}_{\mathrm{C}}$. Through the classical Gram-Schmidt process, use the first column of $\mathbf{Y}_{\mathrm{C}}$ as $\mathbf{v}_{\mathrm{C}1}$, then orthogonalise the second column to the first column. Thus, the centred orthogonalised response matrix is given by

$$\mathbf{V}_{\mathrm{C}} = \begin{pmatrix} 0.7143 & 0.0000 \\ 0.7143 & 0.0000 \\ -0.2857 & 0.6000 \\ -0.2857 & 0.6000 \\ -0.2857 & -0.4000 \\ -0.2857 & -0.4000 \\ -0.2857 & -0.4000 \end{pmatrix}. \tag{5.78}$$

Third, centre $\mathbf{X}$ into $\mathbf{X}_{\mathrm{C}}$, which is given by

$$\mathbf{X}_{\mathrm{C}} = \begin{pmatrix} -0.9857 & 0.3714 & -2.7429 & -1.2000 \\ -1.1857 & -0.1286 & -2.7429 & -1.2000 \\ 0.9143 & 0.0714 & 0.5571 & 0.0000 \\ 0.3143 & 0.0714 & 0.3571 & 0.1000 \\ 0.2143 & 0.1714 & 1.8571 & 1.1000 \\ -0.2857 & -0.4286 & 0.9571 & 0.5000 \\ 1.0143 & -0.1286 & 1.7571 & 0.7000 \end{pmatrix}. \tag{5.79}$$

Step 2. As no feature has been selected, $\mathbf{X}_{\mathrm{s}}$ is empty and $\mathbf{X}_{\mathrm{r}}$ is the same as $\mathbf{X}$. Correspondingly, $\mathbf{X}_{\mathrm{Cs}}$ is empty and $\mathbf{X}_{\mathrm{Cr}}$ is the same as $\mathbf{X}_{\mathrm{C}}$.

Step 3. In this step, the centred features in $\mathbf{X}_{\mathrm{Cr}}$ are required to be orthogonalised to $\mathbf{W}_{\mathrm{Cs}}$. As no feature has been selected, let $\mathbf{W}_{\mathrm{Cr}}$ equal to $\mathbf{X}_{\mathrm{Cr}}$.

Step 4. The multiple correlation coefficients between $\mathbf{w}_{\mathrm{Cr}i}$ and $\mathbf{V}_{\mathrm{C}}$ are given by

$$\begin{aligned} R^2(\mathbf{w}_{\mathrm{Cr}1}, \mathbf{V}_{\mathrm{C}}) &= h_{1,1} + h_{1,2} = 0.7386 + 0.0242 = 0.7628 \\ R^2(\mathbf{w}_{\mathrm{Cr}2}, \mathbf{V}_{\mathrm{C}}) &= h_{2,1} + h_{2,2} = 0.1047 + 0.1217 = 0.2264 \\ R^2(\mathbf{w}_{\mathrm{Cr}3}, \mathbf{V}_{\mathrm{C}}) &= h_{3,1} + h_{3,2} = 0.9184 + 0.0595 = 0.9779 \\ R^2(\mathbf{w}_{\mathrm{Cr}4}, \mathbf{V}_{\mathrm{C}}) &= h_{4,1} + h_{4,2} = 0.8331 + 0.1273 = 0.9604. \end{aligned} \tag{5.80}$$

Step 5. The third feature (i.e. petal length) has the highest multiple correlation. Thus, the petal length is selected into $\mathbf{X}_{\mathrm{s}}$, and the features contained in $\mathbf{X}_{\mathrm{r}}$ in order are sepal length, sepal width, and petal width.

Step 2. According to the new $\mathbf{X}_{\mathrm{s}}$ and $\mathbf{X}_{\mathrm{r}}$, the centred matrix $\mathbf{X}_{\mathrm{C}}$ is divided into $(\mathbf{X}_{\mathrm{Cs}}, \mathbf{X}_{\mathrm{Cr}})$.

Step 3. As only one feature is in $\mathbf{X}_{\mathrm{Cs}}$, let the orthogonalised feature $\mathbf{W}_{\mathrm{Cs}}$ equal to $\mathbf{X}_{\mathrm{Cs}}$. Through the classical Gram-Schmidt process, the features in $\mathbf{X}_{\mathrm{Cr}}$ are orthogonalised to $\mathbf{W}_{\mathrm{Cs}}$,

which is given by

$$\mathbf{W}_{\mathrm{Cr}} = \begin{pmatrix} 0.0288 & 0.2616 & 0.0401 \\ -0.1712 & -0.2384 & 0.0401 \\ 0.7082 & 0.0937 & -0.2519 \\ 0.1822 & 0.0857 & -0.0615 \\ -0.4727 & 0.2458 & 0.2604 \\ -0.6398 & -0.3902 & 0.0673 \\ 0.3643 & -0.0582 & -0.0944 \end{pmatrix}. \tag{5.81}$$

Step 4. The multiple correlation coefficients between $\mathbf{w}_{\mathrm{Cr}i}$ and $\mathbf{V}_{\mathrm{C}}$ are given by

$$\begin{aligned} R^2(\mathbf{w}_{\mathrm{Cr1}}, \mathbf{V}_{\mathrm{C}}) &= h_{1,1} + h_{1,2} = 0.0107 + 0.4352 = 0.4458 \\ R^2(\mathbf{w}_{\mathrm{Cr2}}, \mathbf{V}_{\mathrm{C}}) &= h_{2,1} + h_{2,2} = 0.0011 + 0.0830 = 0.0841 \\ R^2(\mathbf{w}_{\mathrm{Cr3}}, \mathbf{V}_{\mathrm{C}}) &= h_{3,1} + h_{3,2} = 0.0296 + 0.4348 = 0.4644. \end{aligned} \tag{5.82}$$

Step 5. The third feature (i.e. petal width) has the highest multiple correlation. Thus, the features contained in $\mathbf{X}_{\mathrm{s}}$ in order are petal length and petal width, and the features contained in $\mathbf{X}_{\mathrm{r}}$ in order are sepal length and sepal width.

Step 2. According to the new $\mathbf{X}_{\mathrm{s}}$ and $\mathbf{X}_{\mathrm{r}}$, the centred matrix $\mathbf{X}_{\mathrm{C}}$ is divided into $(\mathbf{X}_{\mathrm{Cs}}, \mathbf{X}_{\mathrm{Cr}})$.

Step 3. Keep the first column of $\mathbf{X}_{\mathrm{Cs}}$ unchanged, and orthogonalise the second column to the first column through the classical Gram-Schmidt process. The orthogonalised matrix is given by

$$\mathbf{W}_{\mathrm{Cs}} = \begin{pmatrix} -2.7429 & 0.0288 \\ -2.7429 & -0.1712 \\ 0.5571 & 0.7082 \\ 0.3571 & 0.1822 \\ 1.8571 & -0.4727 \\ 0.9571 & -0.6398 \\ 1.7571 & 0.3643 \end{pmatrix}. \tag{5.83}$$

Each feature in $\mathbf{X}_{\text{Cr}}$ is orthogonalised to $\mathbf{W}_{\text{Cs}}$, respectively, to obtain $\mathbf{W}_{\text{Cr}}$, which is given by

$$\mathbf{W}_{\text{Cr}} = \begin{pmatrix} 0.2563 & 0.0486 \\ -0.2072 & -0.0109 \\ -0.0354 & -0.0412 \\ 0.0525 & -0.0073 \\ 0.3320 & 0.1198 \\ -0.2736 & -0.1231 \\ -0.1247 & 0.0140 \end{pmatrix}. \tag{5.84}$$

Step 4. The multiple correlation coefficients between $\mathbf{w}_{\text{Cr}i}$ and $\mathbf{V}_{\text{C}}$ are given by

$$\begin{aligned} R^2(\mathbf{w}_{\text{Cr1}}, \mathbf{V}_{\text{C}}) &= h_{1,1} + h_{1,2} = 0.0105 + 0.0277 = 0.0382 \\ R^2(\mathbf{w}_{\text{Cr2}}, \mathbf{V}_{\text{C}}) &= h_{2,1} + h_{2,2} = 0.0004 + 0.1103 = 0.1108. \end{aligned} \tag{5.85}$$

Step 5. The second feature (i.e. sepal width), which has the highest multiple correlation, is selected into $\mathbf{X}_{\text{s}}$. Therefore, the 3 selected features are petal length, petal width, and sepal width.

The squared canonical correlation coefficients between the 3 features and $\mathbf{Y}$ are given by

$$\begin{aligned} R_1^2((\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_2), \mathbf{Y}) &= 0.9905 \\ R_2^2((\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_2), \mathbf{Y}) &= 0.5626. \end{aligned} \tag{5.86}$$

In LDA, the within-class scatter matrix is given by

$$\mathbf{S}_w = \begin{pmatrix} 0.5067 & 0.2367 & 0.2700 \\ 0.2367 & 0.1917 & 0.1800 \\ 0.2700 & 0.1800 & 0.3050 \end{pmatrix}, \tag{5.87}$$

and the between-class scatter matrix is given by

$$\mathbf{S}_b = \begin{pmatrix} 22.4305 & 10.1333 & -1.1886 \\ 10.1333 & 4.6483 & -0.5800 \\ -1.1886 & -0.5800 & 0.0893 \end{pmatrix}. \tag{5.88}$$

Through solving the eigenvalue problem (5.54), the Fisher's criteria of LDA are given by

$$\begin{aligned} J_1 &= 104.1481 \\ J_2 &= 1.2864. \end{aligned} \tag{5.89}$$

Comparing (5.86) and (5.89), it is verified that the relationship between the squared canonical correlation coefficients and Fisher's criterion of LDA is as described by (5.57). According to (5.44), the sum of the squared canonical correlation coefficients is equal to the sum of the multiple correlation coefficients of the selected features which are computed by hs, that is

$$\begin{aligned} R_1^2((\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_2), \mathbf{Y}) + R_2^2((\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_2), \mathbf{Y}) &= 0.9905 + 0.5626 \\ &= 0.9779 + 0.4644 + 0.1108 \\ &= 1.5531. \end{aligned} \tag{5.90}$$

5.5.2 Application to synthetic data for binomial classification

In this case study, the proposed feature selection method for a binomial classification is investigated. The $N \times n$ feature matrix is sampled from the multivariate normal distribution, which is given by

$$\mathbf{X} \sim \mathcal{M}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathcal{N}}), \tag{5.91}$$

where the mean values in the $n \times 1$ vector $\boldsymbol{\mu}$ is sampled from the normal distribution with mean 0 and standard deviation 0.1. The $n \times n$ covariance matrix $\boldsymbol{\Sigma}_{\mathcal{N}}$ is sampled from the Wishart distribution, which is given by

$$\boldsymbol{\Sigma}_{\mathcal{N}} \sim \frac{1}{N} \mathcal{W}(\boldsymbol{\Sigma}_{\mathcal{W}}, N), \tag{5.92}$$

where $\boldsymbol{\Sigma}_{\mathcal{W}}$ is a $n \times n$ diagonal matrix whose main diagonal is uniformly distributed on the interval $(0, 1)$. Let the number of the observations is 600, i.e. $N = 600$, and the number of the candidate features are 100, i.e. $n = 100$. The 5^{th}, 10^{th}, and 15^{th} features are used to construct the dummy response vector $\mathbf{y}$, which is sampled from the Bernoulli distribution (i.e. 1 trial binomial distribution) given by

$$\mathbf{y} \sim \mathcal{B}(\boldsymbol{\pi}), \tag{5.93}$$

where the probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ is generated by the binomial logistic regression model, that is

$$\pi_i = \frac{1}{1 + \exp\left(-(-2x_{i,5} - 3x_{i,10} + 4x_{i,15})\right)}, \quad i = 1, \ldots, N. \tag{5.94}$$

Given $\mathbf{X}$ and $\mathbf{y}$, the aim of the feature selection study is to find the 3 correct feature indices (i.e. 5, 10, and 15).

The proposed OLS based feature selection methods are compared with the mutual information based feature selection methods. All the feature selection approaches in the comparison are filter methods with different ranking criteria, and the features are selected via greedy search. The mutual information based feature selection methods in the comparison are summarised in [80]. The ranking criteria are the difference and quotient schemes minimal-Redundancy-Maximal-Relevance (mRMRd and mRMRq) [75], Mutual Information Maximisation (MIM) [81], Joint Mutual Information (JMI) [82], Conditional Mutual Info Maximisation (CMIM) [83], Conditional Infomax Feature Extraction (CIFE) [84], Interaction Capping (ICAP) [85], Double Input Symmetrical Relevance (DISR) [86].

For the mutual information based features selection methods, the continuous features are discretised into four categories by the mean values and 1 standard deviation. For the OLS based feature selection, the continuous features are treated in two ways. One (denoted by OLS) implements the algorithm in subsection 5.4.2 to use continuous features directly. Another one (denoted by OLSd) implements the algorithm in subsection 5.4.3, where the continuous features are discretised into four categories by the mean values and 1 standard deviation, and then encoded into matrices by $c-1$ dummy encoding.

The simulation study is repeated 100 times to check how many times the feature selection methods choose the correct 3 features, and the results are given by Table 5.4. In this comparison, two OLS based feature selection methods chooses the right features 95 times and 88 times in the 100 tests, which are the higher than the mutual information based feature selection methods.

5.5.3 Application to synthetic data for multinomial classification

In this case study, the feature selection for a 3-class multinomial classification is investigated. The $N \times n$ feature matrix is generated in the same way as in the last subsection. The number of the observations is 900, i.e. $N = 900$, and the number of the candidate features are 100, i.e. $n = 100$. We use the 5^{th}, 10^{th}, and 15^{th} features to construct the $N \times 3$ response matrix $\mathbf{Y}'$, which is c-label dummy encoded. $\mathbf{Y}'$ is sampled from the categorical distribution (i.e. 1

Table 5.4: The number of the right times for the different feature selection methods in the binomial classification simulation.

Method	Times	Method	Times
OLS	**95**	JMI	79
OLSd	88	CMIM	74
mRMRd	76	CIFE	76
mRMRq	77	ICAP	74
MIM	73	DISR	79

trial multinomial distribution) given by

$$\mathbf{Y}' \sim \mathcal{C}(\mathbf{\Pi}), \tag{5.95}$$

where the $N \times 3$ probability matrix $\mathbf{\Pi} = (\pi_{i,j})$ is composed of the probability vector for each class, that is

$$\mathbf{\Pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3)\,. \tag{5.96}$$

The probability vectors are generated by the multinomial logistic regression model [60, p. 270]. First, the probability ratios are given by

$$\begin{aligned} \frac{\pi_{i,1}}{\pi_{i,3}} &= \exp\left(-x_{i,5} - x_{i,10} + x_{i,15}\right) \\ \frac{\pi_{i,2}}{\pi_{i,3}} &= \exp\left(x_{i,5} - x_{i,10} - x_{i,15}\right), \quad i = 1, \ldots, N. \end{aligned} \tag{5.97}$$

Second, the probability of $\boldsymbol{\pi}_3$ is given by

$$\pi_{i,3} = \frac{1}{1 + \frac{\pi_{i,1}}{\pi_{i,3}} + \frac{\pi_{i,2}}{\pi_{i,3}}}, \quad i = 1, \ldots, N. \tag{5.98}$$

Finally, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ can be computed by substituting (5.98) into (5.97). To make the response matrix become $c-1$-label dummy encoded, the first column of $\mathbf{Y}'$ is removed to form $\mathbf{Y}$. Given $\mathbf{X}$ and $\mathbf{Y}$, the aim of the feature selection simulation is to find the 3 correct feature indices (i.e. 5, 10, and 15).

The task is repeated 100 times, and the number of times when a correct feature selection is achieved is shown in Table 5.5. Two OLS based methods still give the competitive results, especially OLS which uses the continuous features.

Table 5.5: The number of the right times for the different feature selection methods in the multinomial classification simulation.

Method	Times	Method	Times
OLS	**92**	JMI	80
OLSd	84	CMIM	82
mRMRd	83	CIFE	67
mRMRq	84	ICAP	82
MIM	82	DISR	80

5.5.4 Application to the datasets of NIPS feature selection challenge

Two datasets from the NIPS feature selection challenge [2] are used for the feature selection methods evaluation. The detail of the datasets are illustrated in Table 5.6. Dexter dataset is from Reuters text categorisation task and Gisette dataset is from a handwriting recoginition task. Both of the datasets have 2 classes. The features of the datasets are composed of real features and artificial features (called probes). As the probes do not carry information of the class labels, the desirable feature selection methods should avoid selecting them. The datasets are divided into training, validation, and testing data. The labels of the testing data are withheld by the data providers, and the performance on the testing data are obtained by uploading the results to the challenge website.

Table 5.6: Summary of the NIPS feature selection challenge datasets.

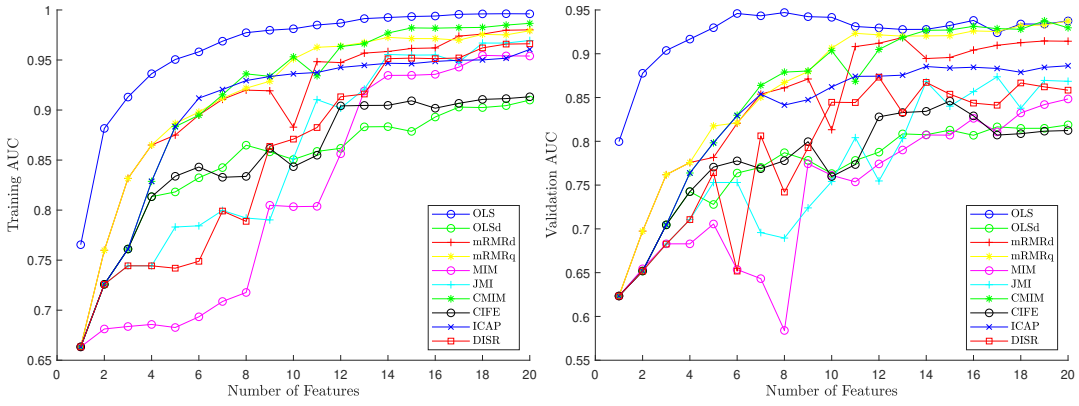
Name	Feature (Real/Probe)	Train/Validation/Test
Dexter	20000 (9947/10053)	300/300/2000
Gisette	5000 (2500/2500)	6000/1000/6500

The feature values in both datasets are quantised to 1000 levels, and the features are treated as continuous. For the mutual information based methods, the 1000 levels are discretised into 10 equal width bins. For OLSd, the discretised features are encoded into matrices

[2] `https://competitions.codalab.org/competitions/3931`

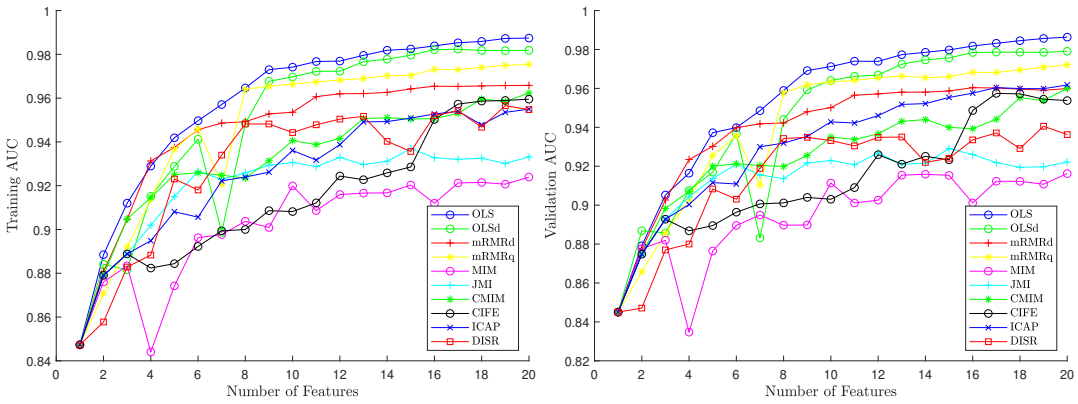
by $c-1$ dummy encoding. For OLS, the continuous features are used directly.

The experiment is implemented in the following steps. First, the optimal features are selected using the training data. Then, a linear Support Vector Machine (SVM) is trained with the training data. Finally, the prediction results are generated by the SVM model on the training, validation, and testing data, respectively. The classification performance is evaluated by the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. Each method selects 20 optimal features. The AUC results on the training and validation data are shown in Fig. 5.1 and Fig. 5.2. Generally, OLS which uses the continuous features gives the best classification performance. In Dexter dataset, OLS shows strikingly better results than other methods. The results on testing data are given in Table 5.7. Although OLS method selects 1 probe in Dexter dataset, the rest of 19 real features (especially the first 8 features according to Fig. 5.1b) selected by OLS are more informative for classification than 20 real features selected by other methods, showing that OLS method can achieve the best AUC results. In conclusion, the OLS based feature selection method shows better performance in linear classification when compared with the mutual information based methods and using continuous features.



(a) (b)

Figure 5.1: AUC results of the feature selection methods on (a) training and (b) validation Dexter dataset.



(a) (b)

Figure 5.2: AUC results of the feature selection methods on (a) training and (b) validation Gisette dataset.

Table 5.7: Results of the NIPS feature selection challenge using linear SVM with 20 features.

Method	Dexter		Gisette	
	AUC	Probe	AUC	Probe
OLS	**0.9551**	1	**0.9873**	0
OLSd	0.8413	4	0.9824	0
mRMRd	0.9246	0	0.9662	0
mRMRq	0.9355	1	0.9776	0
MIM	0.8774	0	0.9324	0
JMI	0.8917	0	0.9352	0
CMIM	0.9444	1	0.9667	0
CIFE	0.8367	14	0.9605	0
ICAP	0.8848	2	0.9580	0
DISR	0.8908	0	0.9490	0

5.6 Conclusions

This chapter proposes a novel OLS based feature selection method for classification. The squared orthogonal correlation coefficients are defined and their relationship with the multiple correlation coefficient and the canonical correlation coefficient have been revealed. Utilising the relationships, the OLS based feature selection method is developed where either the multiple correlation coefficient (for binomial classification) or the canonical correlation coefficient (for multinomial classification) is used as the feature ranking criterion. The equivalence between CCA and LDA is analysed to demonstrate the statistical implication of the canonical correlation coefficient in classification problem. The speed advantage of the OLS based feature selection method in the greedy search has been analysed conceptually. In empirical studies, a simple example has been used to illustrate the procedure of the OLS based feature selection method, and to demonstrate the equivalence between the squared orthogonal correlation coefficients, canonical correlation coefficient, and Fisher's criteria. The synthetic and real world datasets have been used to compare the mutual information based methods with new OLS based methods, showing that the OLS method can achieve the best AUC results in both the synthetic and real data analysis. It is concluded that, when continuous features are used, compared with the mutual information based methods, the OLS based feature selection method can produce a better performance for linear classification.

5.7 Summaries

The proposed frequency feature extraction method in Chapter 4 and the OLS based feature selection method in Chapter 5 are applied into two applications, which is shown in Figure 5.3. For wind turbine fault detection, the nonlinear models are trained by SCADA data, and NOFRFs features are extracted from the models for fault detection. For spontaneous pre-term birth (sPTB) diagnosis, the linear models are trained by the magnetic impedance spectroscopy (MIS) data. After the FRF features are extracted from the models, the optimal features for sPTB diagnosis are selected by the proposed OLS based feature selection method.

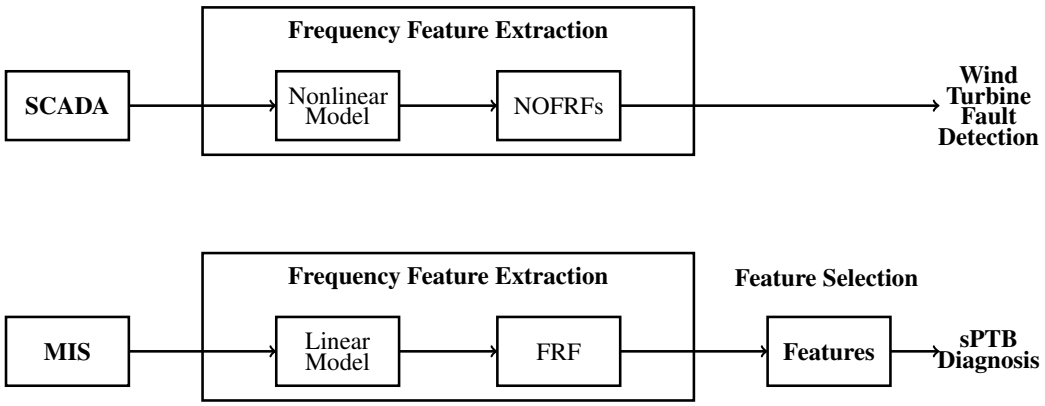


Figure 5.3: Applications of the novel frequency feature extraction and feature selection methods.

Chapter 6

Dynamic Model Sensor and Its Application to Wind Turbine Fault Detection

6.1 Introduction

The wind power generation, as a mainstream option for sustainable energy, requires timely fault detection for reducing the cost of operation and maintenance (O&M) [87, 88, 89]. However, many advanced fault detection approaches are difficult to be implemented in practice due to the need of additional equipment that may incur considerable costs [90]. Therefore, fault detection based on data from supervisory control and data acquisition (SCADA) systems, which have been installed in most MW-scale wind turbines, has attracted extensive research attention [91, 92, 93, 94, 95].

From the raw data of traditional hardware sensor measurements, simple signal processing techniques are often used to detect wind turbine faults by checking whether the values of some measurements have exceeded a threshold [96], or whether the trend of the measurements with a particular wind turbine is significantly different from that with the neighboring wind turbines [97]. However, hardware sensors cannot directly measure some physical variables such as bedding moments and drive-train torques, which are important indicators for wind turbine failures. Consequently, the techniques of soft sensors have been applied to esti-

mate the immeasurable information from the measurable physical variables [98, 99]. A soft sensor is basically a predictive model that is used to infer critical but difficult-to-measure physical variables [100, 101]. For example, the wind turbine shaft torque is vital for bearing fatigue life prognostic but difficult to measure directly; thus, soft sensors were used to estimate shaft torque from the measured generator power output and shaft rotational speed [92, 99]. Principle component analysis (PCA) techniques can also be used as soft sensors to estimate the damage sensitive latent variables. Jia et al. [102] used the standard deviation of the secondary principle component (PC2) derived from measurable SCADA parameters as the indicator of wind turbine failures.

However, many damage sensitive features cannot be revealed by individual measurements but are embedded in the relationship between these measurements. For example, the power curve which shows the relationship between the wind speed and wind turbine power output has been used for wind turbine fault detections [103, 104, 105, 3]. Hereafter, we will refer to such a relationship as the model sensor where the features of the relationship between measurements rather than the measurements themselves are used to evaluate the health conditions of underlying systems.

An illustration of the traditional hardware sensors, soft sensors, and model sensors is shown in Figure 6.1. The system in Figure 6.1 can be a whole wind turbine or a subsystem of the wind turbine such as the generator and gearbox. The inputs are the external environmental conditions such as wind speed and ambient temperature, and the outputs are the physical variables affected by wind turbines operation such as power output and generator temperature. Traditional soft sensors are often built off-line using first principles or data-driven methods [106, 107], while model sensors are required to be built on-line to reveal the changes of the relationship between the input and output measurements in real time.

The sensor measurements are used to represent health conditions of underlying systems or components and damage sensitive features need to be extracted from the measurements for fault diagnosis. The outputs of hardware and soft sensors are individual signals where the features are extracted by signal processing techniques, e.g., Fourier Transform and wavelet analysis [89]. Instead, model sensors use the models to represent system health conditions, and exploit the model analysis techniques to extract the damage sensitive features [91, 22].

Some researchers have already adopted the idea of the model sensor method for SCADA-data-based wind turbine fault detection. For example, Gill et al. [108] used the SCADA data in a normal wind turbine to generate a baseline copula-power curve, which is the power curve transformed by the copula estimation; then the similarity between the copula-power curves in the actual turbine operating condition and the baseline case is evaluated for the purpose of wind turbine fault diagnosis. Yang et al. [91] trained 4^{th} degree polynomial models to describe the relationship between the SCADA parameters under both healthy and faulty conditions. As this relationship varies with turbine health conditions, the model coefficients can be used as indicators for wind turbine fault diagnosis. In these cases, the copula-power curves and polynomial models can both be regards as model sensors for monitoring the health conditions of wind turbines. The damage sensitive model features (the characteristics of the copula-power curves and the coefficients of the polynomial models) rather than individual signals are used for the purpose of fault diagnosis.

However, these existing model sensors are all static model-based, which cannot be used to reveal the damage related changes in dynamic characteristics of wind turbine systems and components. The present study aims to develop a dynamic model sensor which is based on SCADA data and able to detect incipient wind turbine generator faults providing wind farm operators an early alarm when a failure is about to take place. The idea is to establish a dynamic model representation for the relationship between the wind speed, turbine ambient temperature, and generator temperature. It is expected that some changes in this relationship will be able to indicate the development of damage and the occurrence of failure. For these objectives, it is proposed that the structure of the dynamic relationship between wind speed, turbine ambient temperature, and generator temperature is first derived using first principles. Then, the parameters in the dynamic relationship are updated every month using a parameter estimation procedure to produce a model that can reflect the changes in this dynamic relationship. After that, damage sensitive features are extracted monthly from the updated dynamic model to perform fault diagnosis using a novel nonlinear system frequency analysis known as NOFRFs (Nonlinear Output Frequency Response Functions) approach. This novel approach is then applied to process the SCADA data from an operating wind turbine over 5 years when a generator failure had taken place once. The results show that the new approach can not

only detect the occurrence of the failure but also reveal incipient fault occurring well before the failure, demonstrating significant potential of the new dynamic model sensor approach in SCADA-data-based wind turbine fault detection.

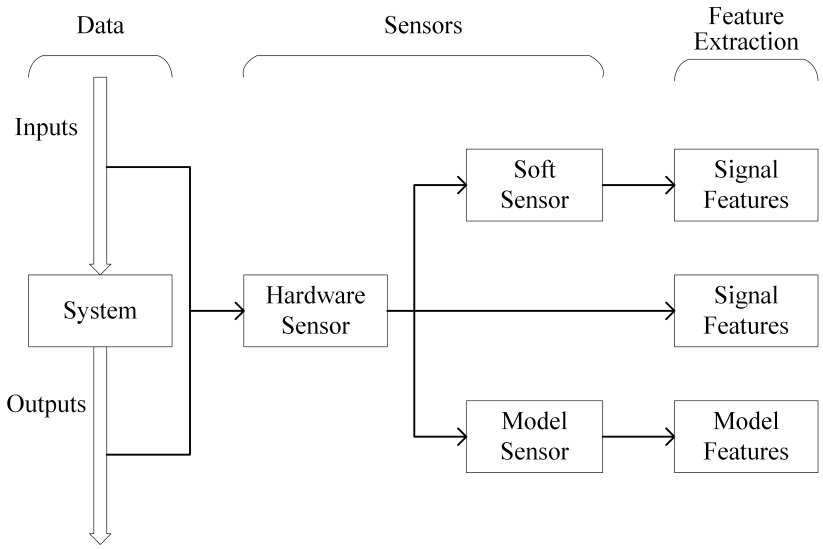


Figure 6.1: Schematic diagram showing a comparison of hardware sensor, soft sensor and model sensor.

6.2 Dynamic model sensor for wind turbine fault detection

Figure 6.2 illustrates the principle of the dynamic model sensor for the SCADA data-based wind turbine fault detection. Here, the time series data are the data of the wind speed, ambient temperature and generator temperature. These data are regularly collected from the SCADA system and used to update the parameters of a model sensor. The model sensor represents the dynamic relationship between the wind speed, ambient temperature, and generator temperature over the time when the data are collected. Therefore, from the analysis of the model sensor characteristics, the operational status of a wind turbine can be evaluated, and potential faults with the turbine system and components can be detected from a damage sensitive index as illustrated at the bottom of Figure 6.2. The implementation of these ideas requires to address three issues which are the model sensor design, model sensor parameter updating, and model sensor analysis, respectively. The model sensor design and parameter updating

are concerned with the determination of the model structure and parameters while the model sensor analysis is to extract model features and evaluate an index which is sensitive to wind turbine system and component damage for potential fault detection.

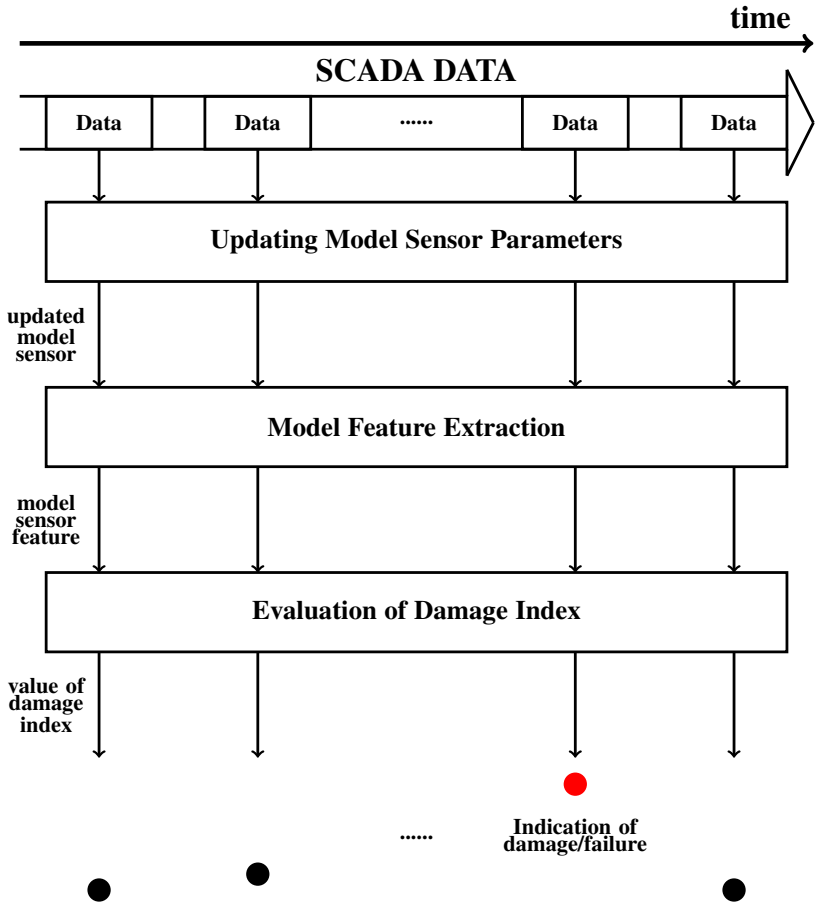


Figure 6.2: Procedural of the model sensor method to detect the system changes.

6.3 Design and updating of model sensors for SCADA-data-based wind turbine generator fault detection

6.3.1 Model sensor design

First principles will be applied in the following to find the relationship between generator winding temperature, wind speed and ambient temperature to determine the model sensor structure. According to [109], the relationship between the temperature change of the wind

turbine generator winding $\Delta T_g(^\circ\mathrm{C})$ and associated energy $Q(\mathrm{J})$ is given by

$$Q = C_g \Delta T_g = C_g(T_g(k) - T_g(k-1)) \tag{6.1}$$

where C_g is the thermal capacitance $(\mathrm{J}/^\circ\mathrm{C})$, $k \in \mathbb{Z}^+$ denotes the discrete time, and $T_g(k)$ is the generator winding temperature at the k^{th} time instant.

The energy Q can be determined by the "energy in" Q_{in} caused by copper loss [110], and the "energy out" Q_{out} caused by cooling, that is

$$Q = Q_{in} - Q_{out}. \tag{6.2}$$

Copper loss is the heat produced by the current in generator windings, and Q_{in} generally has a nonlinear relationship with the wind speed denoted as V_w. It is found that a third-degree polynomial is sufficient to approximate this relationship [111]. as there is no copper loss when wind speed is 0, the constant term in the polynomial is 0; therefore, the polynomial is given by

$$Q_{in} = f(V_w) = f_3 {V_w}^3 + f_2 {V_w}^2 + f_1 V_w. \tag{6.3}$$

For Q_{out}, we only consider the conduction effect between the external environment and the generator, and the thermal resistance $(^\circ\mathrm{C}/\mathrm{W})$ between them is denoted as R_{ga}. The equation of the heat conduction is given by [109]

$$Q_{out} = T_s q_{out} = T_s \frac{T_g - T_a}{R_{ga}} \tag{6.4}$$

where q_{out} is the heat flow rate (W), $T_s = 600\mathrm{s}$ is the time interval of the SCADA data collection, and T_a is the ambient temperature $(^\circ\mathrm{C})$.

The generators often have two cooling systems. One is the passive cooling system which is composed of the blades directly mounted onto the generator rotor shaft. Another is the active cooling system, which is composed of two electrical fans; the two fans work with a constant power (say 3 kW in our case study) when a certain trigger temperature ($40^\circ\mathrm{C}$ in the case study) is surpassed. To simplify the analysis, in this study, only the passive control system is taken into consideration. Therefore, the thermal resistance R_{ga} is determined by generator rotor speed. The fan speed is generally proportional to $1/R_{ga}$, and the relationship

between the rotor speed and the wind speed can be approximated by a third-degree polynomial [112]. Thus, the relationship between the thermal resistance and the wind speed is described by

$$\frac{1}{R_{ga}} = h(V_w) = h_3 {V_w}^3 + h_2 {V_w}^2 + h_1 V_w + h_0. \tag{6.5}$$

According to (6.1)-(6.5), the dynamic model representing the generator winding temperature can be written as

$$T_g(k) = \frac{C_g^{-1} f(V_w(k)) + T_g(k-1) - T_a(k)}{1 + C_g^{-1} T_s h(V_w(k))} + T_a(k). \tag{6.6}$$

6.3.2 Model sensor parameter updating

In order to apply the dynamic model (6.6) to SCADA data for wind turbine fault detection in real time, the parameters of the model need to be updated regularly. In this study, the parameters are updated every month.

The prediction error minimization (PEM) method [53] is applied to update the parameters of model (6.6). The use of the PEM method is based on the relationship

$$\begin{aligned} T_g^*(k) &= \frac{C_g^{-1} f(V_w^*(k)) + T_g^*(k-1) - T_a^*(k)}{1 + C_g^{-1} T_s h(V_w^*(k))} + T_a^*(k) + e(k) \\ &\triangleq \hat{T}_g(k) + e(k) \end{aligned} \tag{6.7}$$

where $T_g^*(k)$, $V_w^*(k)$ and $T_a^*(k)$ are the generator temperature, wind speed and ambient temperature measured by the SCADA system, $e(k)$ is the modelling error. Considering (6.3) and (6.5), $\hat{T}_g(k)$ can be written as

$$\hat{T}_g(k) = \frac{C_g^{-1}(f_3 {V_w^*}^3(k) + f_2 {V_w^*}^2(k) + f_1 V_w^*(k)) + T_g^*(k-1) - T_a^*(k)}{1 + C_g^{-1} T_s (h_3 {V_w^*}^3(k) + h_2 {V_w^*}^2(k) + h_1 V_w^*(k) + h_0)} + T_a^*(k). \tag{6.8}$$

In order to estimate the parameters of the model sensor, PEM minimizes the square of the prediction error such that

$$\min_{\theta} \sum_{k=1}^{N_s} F_k^2(\theta) \tag{6.9}$$

where N_s is the sample size,

$$\begin{aligned} F_k(\theta) =& \hat{T}_g(k) - T_g^*(k) \\ =& \frac{C_g^{-1}(f_3 {V_w^*}^3(k) + f_2 {V_w^*}^2(k) + f_1 V_w^*(k)) + T_g^*(k-1) - T_a^*(k)}{1 + C_g^{-1} T_s (h_3 {V_w^*}^3(k) + h_2 {V_w^*}^2(k) + h_1 V_w^*(k) + h_0)} \\ & - \left(T_g^*(k) - T_a^*(k)\right) \end{aligned} \tag{6.10}$$

and

$$\begin{aligned}\theta =&[\theta_1, \ldots, \theta_7]^{\mathrm{T}} \\ =&C_g^{-1}\left[f_3, f_2, f_1, T_s h_3, T_s h_2, T_s h_1, 1/C_g^{-1} + T_s h_0\right]^{\mathrm{T}}\end{aligned} \tag{6.11}$$

is the vector of model sensor parameters to be updated every month for the purpose of wind turbine fault detection. After the parameter vector θ has been obtained, the dynamic model sensor, according to (6.8), is given by

$$y(k) = \frac{\theta_1 {u_1}^3(k) + \theta_2 {u_1}^2(k) + \theta_3 u_1(k) + y(k-1) - u_2(k)}{\theta_4 {u_1}^3(k) + \theta_5 {u_1}^2(k) + \theta_6 u_1(k) + \theta_7} + u_2(k) \tag{6.12}$$

where $u_1 = V_w^*$ and $u_2 = T_a^*$ are model inputs and $y = \hat{T}_g$ is the model output.

6.4 Extraction of damage sensitive model sensor features using NOFRFs

To quantitatively evaluate the characteristics of the model sensors for fault detection, the damage sensitive features should be extracted from each updated model sensor. The model parameters can be the features in some cases, but the number of the parameters involved in model sensor (6.12) implies the parameters are hard to be used to produce a simple index for the fault detection objective. The frequency features of a system model are often very effective features for the representation of system properties [113]. When a model is nonlinear, the NOFRFs have been demonstrated to be effective for the frequency feature extraction and analysis [44, 22]. Therefore, the NOFRFs will be exploited here for the extraction of the model sensor features for wind turbine fault detection.

6.4.1 The effects of the system operating point

For model sensor (6.12), when wind speed varies around V_w^0 and turbine ambient temperature varies around T_a^0, the model sensor is known as being working around the operating point (V_w^0, T_a^0). If the variation of the turbine ambient temperature about T_a^0 is negligible, the extraction of the NOFRFs features of model sensor (6.12) can be achieved by considering the

case where system (6.12) is subject to inputs

$$\begin{aligned} u_1(k) &= u(k) + V_w^0 \\ u_2(k) &= T_a^0. \end{aligned} \tag{6.13}$$

As one input has been set a constant, system (6.12) can be regarded as a single input and single output system which has input $u(k)$ and output $y(k)$ and can be represented by the Volterra series (2.18) where $h_n(\tau_1, \ldots, \tau_n)$, $n = 0, \ldots, N$ are dependent on (V_w^0, T_a^0). Figure 6.3 illustrates this representation of model sensor system (6.12). Because $h_n(\tau_1, \ldots, \tau_n)$, $n = 0, \ldots, N$, are now the functions of (V_w^0, T_a^0), the NOFRFs evaluated using the method in Section 2.3.2 above will be affected by (V_w^0, T_a^0). Therefore, the evaluated NOFRFs should be compared with the baseline NOFRFs at the same operating point as (V_w^0, T_a^0) for the purpose of wind turbine fault diagnosis.

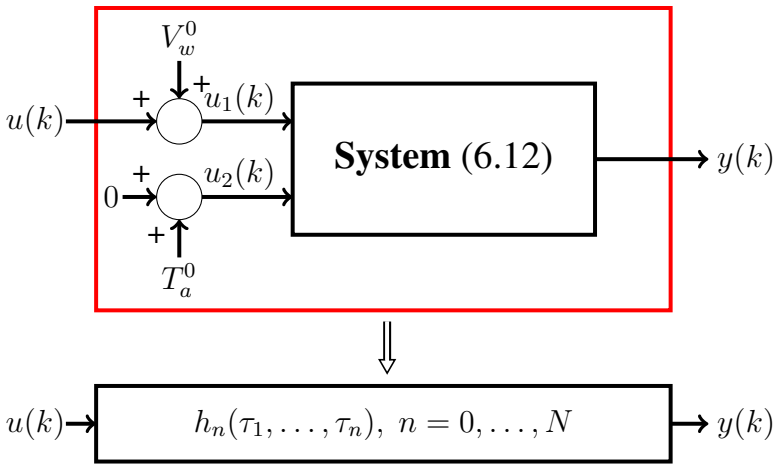


Figure 6.3: The effects of the operating points on the Volterra series expansions of model sensor.

6.4.2 Damage sensitive indices

Denote the NOFRFs of model sensor (6.12) and their baseline that represent the current and normal conditions of a wind turbine working about the operating point (V_w^0, T_a^0) as $G_n(j\omega)$ and $G_n^b(j\omega)$, $n = 0, \ldots, N$, respectively. Then, the NOFRFs based damage sensitive indices for the wind turbine can be defined as

$$I_n(j\omega) = |G_n(j\omega)| - \left|G_n^b(j\omega)\right|. \tag{6.14}$$

In principle, $G_n^b(j\omega), n = 0, \ldots, N$ are the NOFRFs determined from a benchmark wind turbine which is similar to the evaluated turbine system and is working normally in a similar environment.

In practical applications, $M(M \geq 2)$ wind turbines in the same wind farm and located near the wind turbine of concern can be used as the benchmark and, each time, sets of the values of indices (6.14) are evaluated. Obviously, in most cases, when an alarm is raised from the evaluated values of (6.14), there are two possible situations. One is the benchmark is normal, so the alarm correctly indicated there exists a fault with the wind turbine of concern. Another is the benchmark turbine is of fault so the alarm may be wrong. However, it is reasonable to assume that at the same time lower than 50% of the benchmark wind turbines can be in fault. Therefore, if less than 50% of the evaluated sets of the values of index (6.14) indicate there exists a fault, the turbine of concern can be considered to be normal. Otherwise, the turbine can be of fault. Thus, the SCADA data based wind turbine fault detection can be achieved.

6.5 Application to fault detection of three operating wind turbines

6.5.1 The SCADA data from operating wind turbines

The data used in the present study were collected from the SCADA of three operating wind turbines which are referred to as A101, A102 and A103, respectively. The three wind turbines are of the same model and located close to each other in a wind farm in Spain. The data were collected from 08/2009 to 12/2014. Every 10 minutes, 40 measurements were obtained from each turbine. These measurements include the wind speed, the temperature of various components, the vibration of the tower, and the power output, etc. The maximum, minimum, standard deviation and average of the measurements over the 10 minutes were recorded.

In 01/2013, the generator of turbine A103 was replaced for a serious rotor winding failure, which is shown in Figure 6.4 This generator failure was detected by the SCADA system on 21/01/2013 after the generator had totally broke down. Then the wind turbine was recovered

after a new generator was installed on 22/01/2013.

In this application study, the model senor technique proposed in Sections 6.2 to 6.4 will be applied to process the SCADA data in order to demonstrate how to use the proposed technique to detect the generator failure in advance. For this purpose, the SCADA data were first pre-processed to clean the data by 1) Removing the data sets which have missing data and outliers, and 2) setting wind speed as 0 when wind turbine stopped.

A considerable proportion of missing values were found from the SCADA data. For example, 2.15% power data during 08/2009 to 12/2014 from wind turbine A103 are missing. Data set with missing data cannot be used for model sensor parameter updating, so have to be removed. The data set with outliers may interfere the accuracy of model parameter estimation so should also be removed. The data measured when wind turbines stopped have been kept, since the data can help to model the cooling processes. Wind turbines may shut down due to wind speed is outside the cut-in (3 m/s) and cut-off (25 m/s) wind speed or curtailment. In both cases, no power will be produced, and the generator will cool down. As the wind speed has no effect on power output or generator temperature when a wind turbine stops, the wind speed is treated as 0 in the data analysis.

6.5.2 Power curve analysis

The traditional power curve analysis is applied to the SCADA data of wind turbine A103, which is shown in Fig. 6.5. Through comparing the power curves of the three years before the failure happened, no significant change is found when the time approaches the failure.

6.5.3 Comparison with the constant thermal resistance model in the cooling process

The similar thermal dynamic model structure can also be found in [114], where the thermal resistant is assumed as a constant. The model fitting performance of the constant thermal resistance model in [114] and the varying thermal resistance model (6.6) is compared in Fig. 6.6. There are two cooling processes shown in the two rows of the figure. In both processes, the wind turbine is shut down from about 1300 minutes, and then the generator starts cooling

Figure 6.4: Rotor winding short circuit causing the generator failure of wind turbine A103 in 01/2013.

down. It is found for the model with the constant thermal resistance, the model predicted cooling process is slower than the real system at the beginning, but faster at the end. The reason of the mismatch is because that the cooling system still works at the beginning of the cooling process, until the generator rotor shaft is totally stopped. The thermal resistance of the real system unceasingly increases with the rotor speed slowdown in the cooling process. The proposed model (6.6) successfully describes the varying thermal resistance, and the fitting performance is improved significantly.

6.5.4 Application of the model sensor method to the wind turbine fault detection

In this application, model sensor (6.12) was adopted. From SCADA data collected from 08/2009 to 12/2014, after cleansing, about 4000 data were used each month for updating

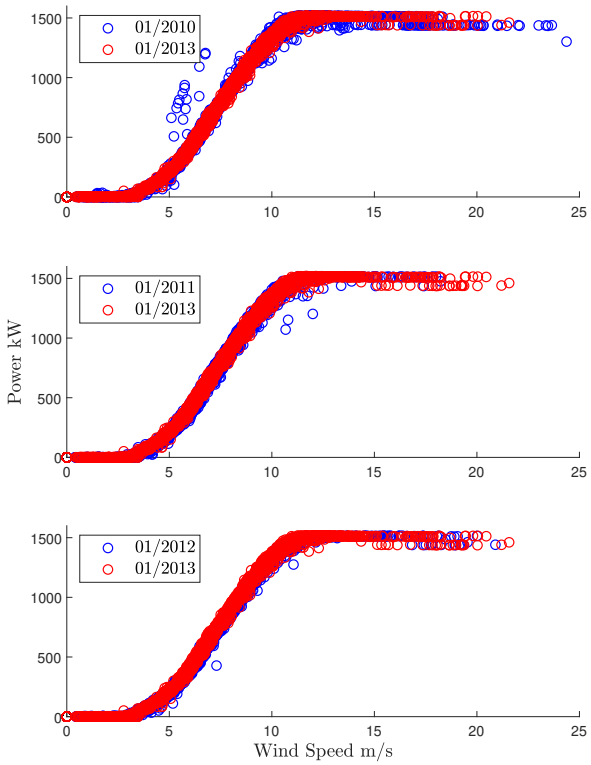


Figure 6.5: Comparison of the power curve of 01/2013 with the power curves of 01/2010, 01/2011, and 01/2012.

the parameters of model sensor (6.12) for turbine A101, A102, and A103, respectively. For A103, as it was broken down on about 21/01/2013, the data used for updating its model sensor in January 2013 were collected during the period from 01/01/2013 to 19/01/2013. From the data collected from the three turbines every month, three model sensors of the form of equation (6.12) are determined and used to represent the operating conditions of the three wind turbines, respectively. For example, the model sensor for A102 in 09/2009 was obtained as follows.

$$y(k) = \frac{0.0068{u_1}^3(k) - 0.0714{u_1}^2(k) + 0.7190u_1(k) + y(k-1) - u_2(k)}{1.7943 \times 10^{-4}{u_1}^3(k) - 0.0042{u_1}^2(k) + 0.0348u_1(k) + 1.0059} + u_2(k). \quad (6.15)$$

In order to evaluate the NOFRFs of the model sensors, the operating point of $\left(6.3, \bar{T}_a\right)$ was used where $\bar{T}_a$ is the average ambient temperature in each month and 6.3 m/s is the average wind speed over 08/2009 to 12/2014. So, the inputs of model sensor (6.12) that were used

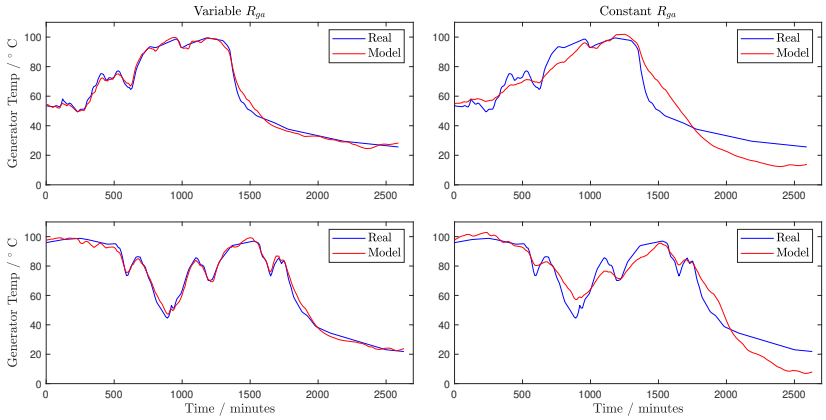


Figure 6.6: Model fitting performance of the constant and varying thermal resistance models in the cooling processes.

for the evaluation of the NOFRFs are given by

$$\begin{aligned} u_1(k) &= u(k) + 6.3 \\ u_2(k) &= \bar{T}_a \end{aligned} \tag{6.16}$$

where

$$u(k) = \alpha u^*(k) \tag{6.17}$$

$u^*(k) = \cos(\omega_c k T_s)$, and $\omega_c = 7.27 \times 10^{-5}$rad/s, which corresponds to the period of one day and is an important frequency component in the spectra of wind turbine SCADA measurements.

The NOFRFs of model sensor (6.12) determined from every month's SCADA data from a wind turbine were evaluated using (2.32) from $u(k)$ given by (6.17) and corresponding output $y(k)$ of the model sensor. The NOFRFs up to the 2nd order, i.e. G_0, $G_2(0)$, $G_1(j\omega_c)$ and $G_2(j2\omega_c)$) were used to assess the wind turbine operating conditions.

For example, in order to evaluate $G_1(j\omega_c)$, the maximum order of system nonlinearity was taken as $N = 6$ which implies the input $u^*(k)$ should be scaled by at least 3 different α (i.e. $\bar{N} \geq 3$). In this case, following the procedure in Section 2.3.2, three responses of model sensor (6.12) to $u(k) = \alpha_1 u^*(k) = 0.25u^*(k)$, $u(k) = \alpha_2 u^*(k) = 0.35u^*(k)$ and

$u(k) = \alpha_3 u^*(k) = 0.5u^*(k)$, respectively, were used to determine $G_1(j\omega_c)$ as

$$\begin{aligned} G_1(j\omega_c) &= [1,0,0]\begin{bmatrix} G_1(j\omega_c) \\ G_3(j\omega_c) \\ G_5(j\omega_c) \end{bmatrix} \\ &= [1,0,0]\Big[\mathbf{AU}(j\omega_c)^{\mathrm{T}}\mathbf{AU}(j\omega_c)\Big]^{-1}\mathbf{AU}(j\omega_c)^{\mathrm{T}}\mathbf{Y}(j\omega_c) \end{aligned} \tag{6.18}$$

where

$$\begin{aligned} \mathbf{Y}(j\omega_c) &= \begin{bmatrix} Y^1(j\omega_c) \\ Y^2(j\omega_c) \\ Y^3(j\omega_c) \end{bmatrix} \\ \mathbf{AU}(j\omega_c) &= \begin{bmatrix} U_1^1(j\omega_c) & U_3^1(j\omega_c) & U_5^1(j\omega_c) \\ U_1^2(j\omega_c) & U_3^2(j\omega_c) & U_5^2(j\omega_c) \\ U_1^3(j\omega_c) & U_3^3(j\omega_c) & U_5^3(j\omega_c) \end{bmatrix} \end{aligned} \tag{6.19}$$

Similarly, G_0, $G_2(0)$, and $G_2(j2\omega_c)$) can also be obtained.

In this study, A101 and A102 were used as the benchmark turbines. Therefore $M = 2$, and the M sets of damage sensitive indices for A103 are given by

$$\begin{cases} I_0^{A101} = \left|G_0^{A103}\right| - \left|G_0^{A101}\right| \\ I_1^{A101}(j\omega_c) = \left|G_1^{A103}(j\omega_c)\right| - \left|G_1^{A101}(j\omega_c)\right| \\ I_2^{A101}(0) = \left|G_2^{A103}(0)\right| - \left|G_2^{A101}(0)\right| \\ I_2^{A101}(j2\omega_c) = \left|G_2^{A103}(j2\omega_c)\right| - \left|G_2^{A101}(j2\omega_c)\right| \end{cases} \tag{6.20}$$

and

$$\begin{cases} I_0^{A102} = \left|G_0^{A103}\right| - \left|G_0^{A102}\right| \\ I_1^{A102}(j\omega_c) = \left|G_1^{A103}(j\omega_c)\right| - \left|G_1^{A102}(j\omega_c)\right| \\ I_2^{A102}(0) = \left|G_2^{A103}(0)\right| - \left|G_2^{A102}(0)\right| \\ I_2^{A102}(j2\omega_c) = \left|G_2^{A103}(j2\omega_c)\right| - \left|G_2^{A102}(j2\omega_c)\right| \end{cases} \tag{6.21}$$

The values of the two sets of indices evaluated by applying the proposed model sensor technique to the SCADA data of A101, A102, and A103 over the period from 08/2009 to 12/2014 are shown in Figure 6.7.

It can be observed from Figure 6.7 that both I_0^{A101} and I_0^{A102} start from a point higher than 0, and then follow a trend which slowly increases with time until the generator failure of

A103 takes place in January 2013. After the time point when A103 generator was replaced, both I_0^{A101} and I_0^{A102} reduce back to about zero. These exactly reflect the actual operating conditions of wind turbine A103 and indicate that I_0 can be used as an excellent index for the SCADA data based wind turbine fault diagnosis.

In addition, Figure 6.7 shows that $I_1(j\omega_c)$ can also be a good index for the purpose of SCADA data based wind turbine fault diagnosis. However, the trend with I_0, which slowly increases with time until the point when A103's failure took place, cannot be very obviously observed from $I_1(j\omega_c)$.

Also it is worth pointing out that although $G_2(0)$ and $G_2(j2\omega_c)$ cannot, as clearly as I_0, indicate the trend of change of the wind turbine operating status, it is still necessary to take the effect of system nonlinearity into account in the proposed analysis. If (6.3) and (6.5) are simplified as

$$\begin{aligned} f(V_w) &= f_1 V_w \\ g(V_w) &= g_0 \end{aligned} \tag{6.22}$$

the dynamic model (6.6) become linear, which is given by

$$T_g(k) = \frac{C_g^{-1} f_1 V_w(k) + T_g(k-1) - T_a(k)}{1 + C_g^{-1} T_s h_0} + T_a(k). \tag{6.23}$$

Fig. 6.8 shows the results of I_0 and $I_1(j\omega_c)$ determined when assuming the linear model structure (6.23) so $N = 1$. It can be observed from Fig. 6.8 that I_0 thus obtained is not able to clearly show the trend of turbine winding ageing so as to properly issue an alarm before the winding failure takes place. In addition, $I_1(j\omega_c)$ obtained in this way is also obviously no longer a good index for the turbine operating conditions.

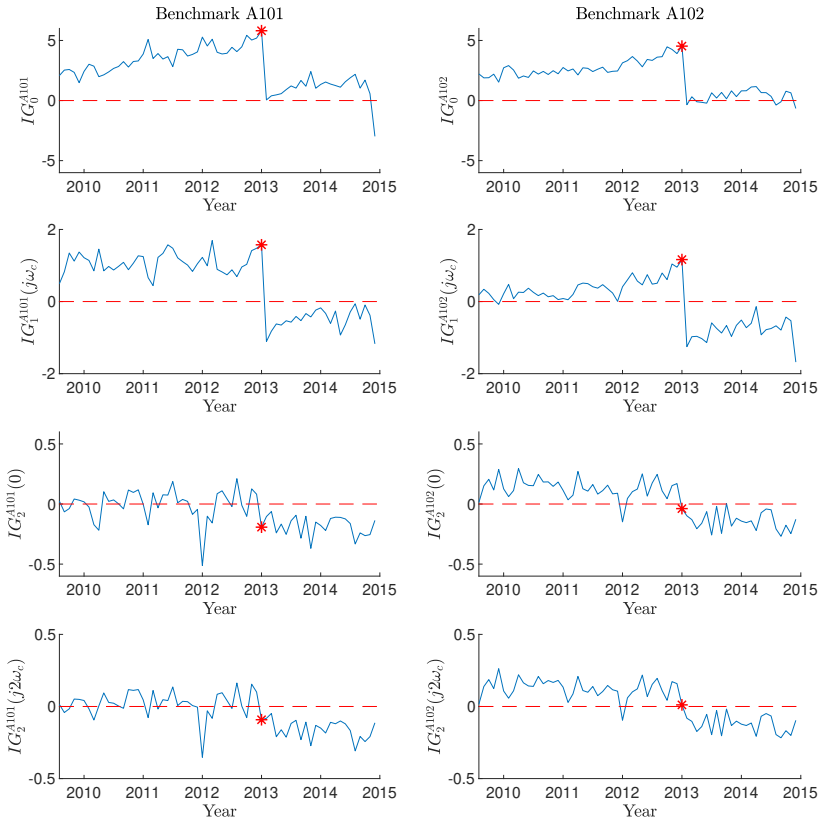


Figure 6.7: The values of A103 damage sensitive indices I_0, $I_2(0)$, $I_1(j\omega_c)$ and $I_2(j2\omega_c)$) and of the model sensor (6.12) evaluated from the SCADA data collected from 08/2009 to 12/2014 with * indicating when the generator failure took place.

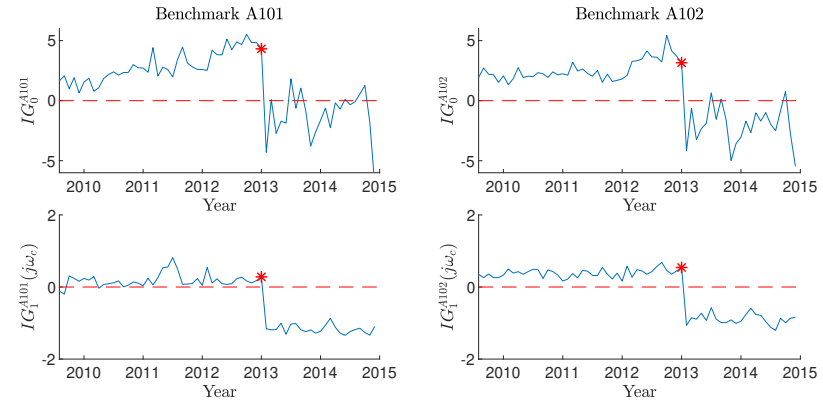


Figure 6.8: The values of A103 damage sensitive indices I_0 and $I_1(j\omega_c)$ extracted from a linear approximation of model sensor (6.12).

6.6 Conclusions

In this chapter, a novel dynamic model sensor method is proposed for the detection of faults in wind turbines from the SCADA data. The model sensor represents the dynamic relationship between the generator temperature, wind speed, and ambient temperature with the model structure derived from first principles. When applied to SCADA data to conduct turbine fault detection, the parameters of the model sensor are updated every month by a parameter estimation process so that the model can timely represent the turbine operating conditions. Then, a NOFRFs based frequency analysis for the model sensor characteristics is carried out to extract damage sensitive indices and to perform fault detection based on the values of these indices. The application of the new model sensor method to 5 years' SCADA data of three operating wind turbines in Spain has shown that the new method can not only correctly detect a generator failure with one of the three turbines but also reveal the trend of ageing with the turbine's winding insulations. The key idea with the proposed method is to use the changes in the properties of inspected systems to conduct system fault detection. The field data analysis in the present study has demonstrated the effectiveness of this novel idea and its potential applications in wind energy industry.

Chapter 7

Spontaneous Preterm Birth Prediction and Diagnosis Based on Magnetic Impedance Spectroscopy

7.1 Introduction

Infants are born at less than 37 weeks of gestational age are preterm births, in which about 65-75% are spontaneous preterm births (sPTB) [115, 116]. Preterm birth is the leading cause of the perinatal mortality in developed countries, and many of the surviving preterm infants suffer serious morbidity [117, 118, 119]. Accurate prediction of spontaneous preterm birth helps the perinatal health care, which is effective to reduce the associated complications [116].

The most effective risk factors for sPTB so far are cervical length (CL) and quantitative fetal fibronectin (fFN) [115, 120, 121, 122]. However, the recent research showed that CL and fFN had low predictive accuracy for sPTB among nulliparous women with singleton pregnancies [4]. In addition, the risk factor like fFN is only feasible in predicting the sPTB within 7-14 days [123, 124]. The short prediction time horizon makes the perinatal health care be difficult. Therefore, other risk factors are necessary to be discovered to combine with CL and fFN to improve the sPTB prediction accuracy.

In this chapter, Magnetic impedance spectroscopy (MIS) features are used as the new risk factors for sPTB prediction. MIS is used to describe the relationship between the induced

magnetic field and the excitation magnetic field under different frequency. The induced magnetic field is related to the dielectric properties of the environment around the induction coil [125]. Thus, if the coil for producing induced magnetic field is surrounded by the investigated object, the change of the induced magnetic field can be used to indicate the change of the characteristics of the object. The MIS technique has been used in many areas, such as food quality assessment [126], nonferrous metal classification [127], and conductive fluid imaging [128].

The traditional MIS measurement device converts the raw time domain data to frequency domain data by the Field-Programmable Gate Array (FPGA) at several discrete frequencies, and only provide frequency domain data. However, only the steady-state data are utilised to generate frequency domain data, while the transient-state data between the frequency switching are discarded. To make better use of the transient-state data, the time series model is trained by both transient and steady-state data, and the frequency domain features are generated from the model indirectly. Compared to the traditional frequency domain measurements, the proposed method can obtain the frequency features over a continuous range rather than a few discrete points.

After the features have been obtained, the feature selection method introduced in Chapter 5 are applied to select the most useful features, which are then used in logistic regression models for sPTB prediction. The results demonstrate the feasibility of the MIS in sPTB prediction, and show the potential of the proposed time domain data modelling method in carrying out the MIS measurement.

7.2 MIS measurement and data description

The data are collected measured by MIS2 device which is designed to safely take measurements of the impedance of the cervix tissue in pregnant women. The impedance is defined by (2.33), where the input of the system is the alternating current (AC) and the output is the induced voltage which is affected by the cervix tissue. The AC is applied at 15 different frequencies sweeping from 21 kHz to 1.013 MHz. Each frequency takes 12 ms, and the highest frequency (i.e. 1.013 MHz) will repeat once. Therefore, the total time of a frame, i.e. a sweep

of 15 frequencies, is 192 ms. As the sampling frequency is 40 MHz, 480,000 data (in 12 ms) are sampled at each frequency. Figure 7.1 shows an example of the input and the output of a frame. Generally, the amplitude of AC decreases with the increase of the frequency to control the induced voltage within a safe range to pregnant women.

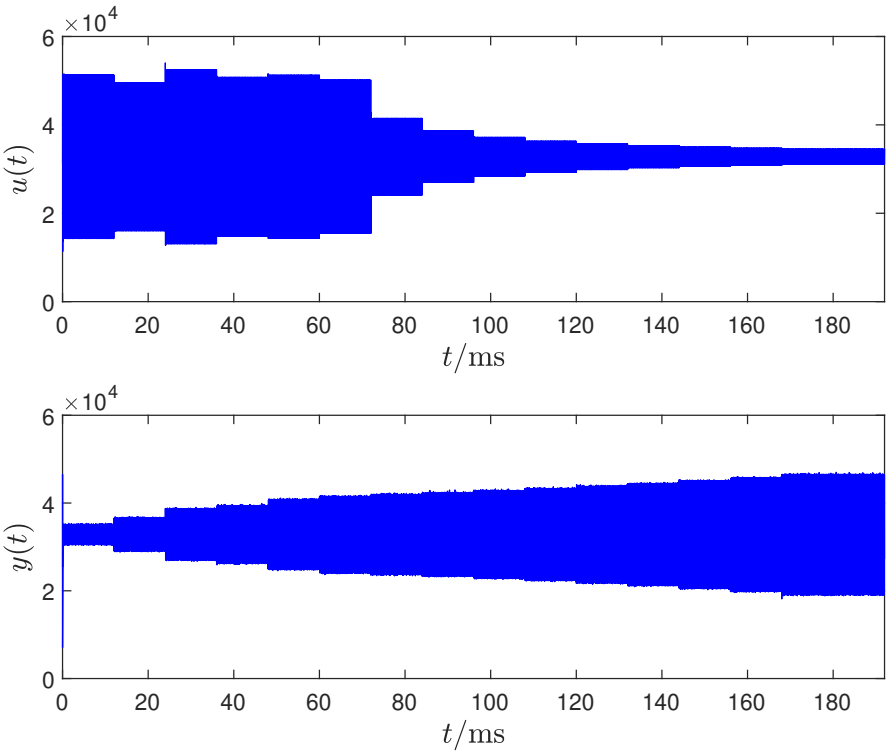


Figure 7.1: The time history of a frame.

After the nurse switches on the probe, the measurement for a pregnant woman is taken in the following order.

- Ferrite target: the nurse puts the probe into ferrite solution and press the footswitch.
- Air: the nurse takes out the probe and presses the footswitch. Then the probe is wiped.
- Cervix 1: the probe is put on the cervix and the footswitch is pressed.
- Vaginal 1: the probe leave the cervix a short distance but still in the vagina. Then the nurse presses the footswitch.
- Cervix 2: repeat Cervix 1.

- Vaginal 2: repeat Vaginal 1.
- Cervix 3: repeat Cervix 1.

In this process, the MIS2 continuously generate the data in frames. The whole measurement for a woman needs 2000 to 5000 frames. In each time the nurse presses the footswitch, an event flag is generated and the following 4 data frames are stored. There are 7 events in total for a woman's measurement and 28 raw data frames are stored. Except the 28 raw data frames, all other frames are converted into frequency domain by FPGA and stored. For each frame, both of the input AC and the output induced voltage will be converted to 16 complex numbers by Fourier transform to represent the amplitude and phase at 16 frequencies. Taking 21 kHz as an example, the first frequency in a frame is 21 kHz, which takes 12 ms. The leading 1 ms and ending 1 ms is transient process which is shown in Figure 7.2. Only take the middle of 10 ms (i.e. 40,000 samples) to do Fourier transform. Then take the complex number at the 21 kHz as results. The same procedure is repeated in other frequencies, so 16 complex numbers are generated in one frame. In the end, the frequency domain data (or FPGA data) for all frames are given and time domain data (or raw data) for only 28 frames are generated. Figure 7.3 shows the whole measurement of the imaginary part at 1.013 MHz for a woman. There are 5174 frames for both the input and the output in this measurements. The 7 events are marked as the red triangles in order. The raw data are collected from each red triangle to the following 4 consecutive frames.

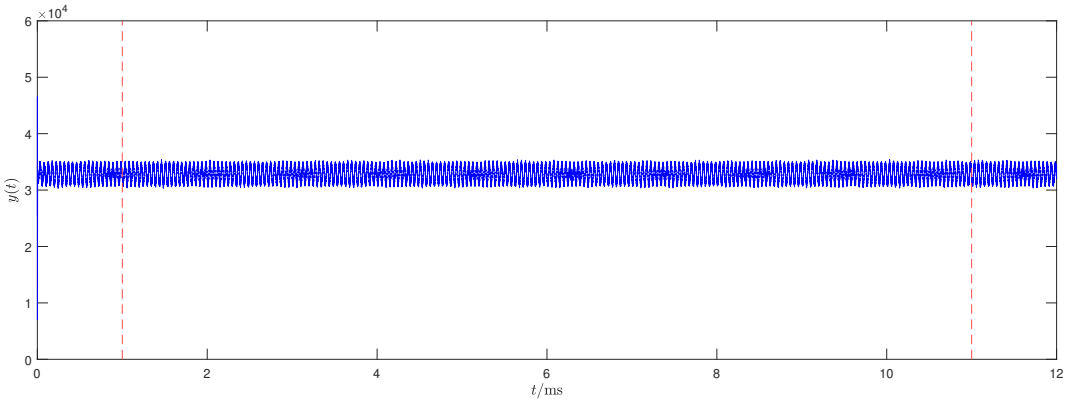


Figure 7.2: The time history of the 21 kHz measurement. The two red dash lines split the time history into 3 parts. The head and end parts are the transient states, and the middle part is the steady state.

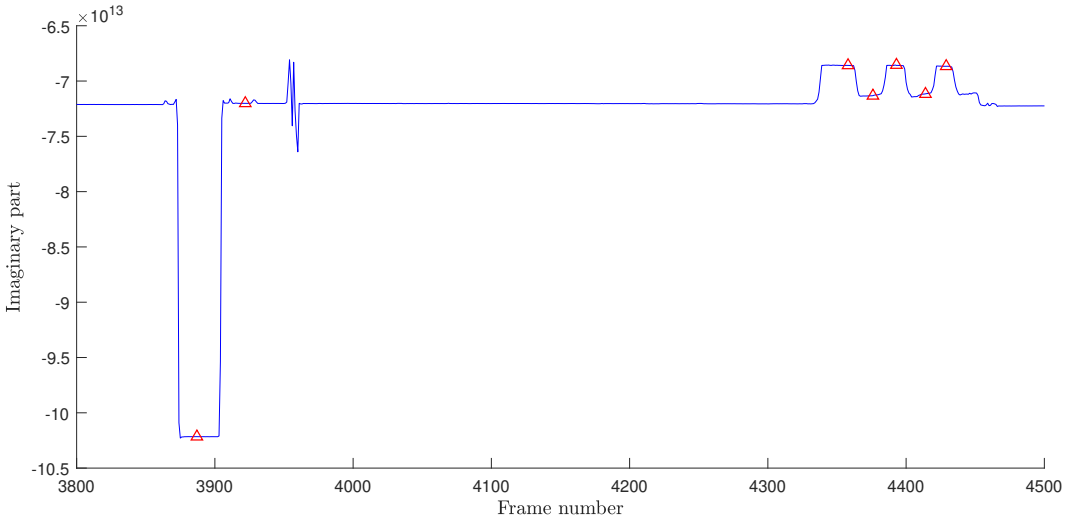


Figure 7.3: The imaginary part of the output spectrum at 1.013 MHz in a measurement. The 7 events are marked by the red triangle.

7.3 System identification using MIS data

The relationship between the input current u and the induced output voltage y is represented by an ARX model. The model terms are determined by FROLS algorithm shown in Algorithm 1. It is found that the sampling frequency is too high, which leads to the neighbour terms such as $y(k-1)$ and $y(k-2)$ are nearly linearly dependent. Thus, after the orthogonalisation, the FROLS algorithm tends to reject the terms which are close, but select the terms which have large time delay difference. The large time delay terms lead to a high order ARX model which is overly complicated for this study. Therefore, the original MIS data are down sampled to 5 MHz.

Generally, the FROLS algorithm selects similar terms for different women's MIS measurement, so the ARX model structure is fixed as

$$\begin{aligned} y(k) =& \theta_0 + \theta_1 y(k-1) + \theta_2 y(k-2) \\ &+ \theta_3 u(k) + \theta_4 u(k-1) + \theta_5 u(k-5) + e(k), \end{aligned} \tag{7.1}$$

where $\theta_0, \ldots, \theta_5$ are the model parameters and e is the noise. The parameters are trained by a frame of the MIS data, so each frame corresponds to a set of parameters. It is noticed that both u and y vary around 32800. If we centre u and y to make them vary around 0, the constant term θ_0 can be eliminated. Thus, the ARX model structure change to

$$\begin{aligned} y_c(k) =& \theta_1 y_c(k-1) + \theta_2 y_c(k-2) \\ &+ \theta_3 u_c(k) + \theta_4 u_c(k-1) + \theta_5 u_c(k-5) + e(k), \end{aligned} \tag{7.2}$$

where y_c and u_c are the centred u and y.

The parameters $\theta_1, \ldots, \theta_5$ are determined by least-squares which minimises the cost function

$$R(\boldsymbol{\theta}) = \frac{1}{N_s - 5} \sum_{k=6}^{N_s} \left[(y_c(k) - \mathbf{x}_c \boldsymbol{\theta} \right]^2, \tag{7.3}$$

where

$$\begin{aligned} \boldsymbol{\theta} &= [\theta_1, \ldots, \theta_5]^\top \\ \mathbf{x}_c(k) &= [y_c(k-1), y_c(k-2), u_c(k), u_c(k-1), u_c(k-5)], \end{aligned} \tag{7.4}$$

and N_s is the number of samples. It is found the value of the parameters are effected by all samples. However, the most of the data are in the steady states which are sinusoidal wave at

a specific frequency, while the transient states which cover more frequency information only account for a small part. Therefore, the least squares make the model intend to fit the steady states but ignore the transient states. To make the least squares focus on fitting transient states, the steady states in the data are removed. Thus, the cost function becomes

$$R(\boldsymbol{\theta}) = \frac{1}{|S|} \sum_{k \in S} \left[(y_c(k) - \mathbf{x}_c(k)\boldsymbol{\theta} \right]^2, \tag{7.5}$$

where $S = \{k | y(k-1), y(k-2), u(k), u(k-1), u(k-5) \notin \text{steady states}\}$ and $|S|$ is the cardinality of the set S.

After the parameters are evaluated, which are denoted as $\hat{\boldsymbol{\theta}}$, the model fitting performance should be validated. In time domain, as $\mathbf{x}_c(k)\hat{\boldsymbol{\theta}}$ is the model predicted output, $R(\hat{\boldsymbol{\theta}})$ is the mean square error (MSE) of the model, which can used for model validation. MSE ranges from 0 (perfect fitting) to infinity (bad fitting). The models in this study can also be validated in frequency domain. Firstly, the frequency response of the ARX model is obtained. The ARX model is trained by a frame of the raw data. Secondly, the frequency responses obtained from the steady state of the raw data. The impedance spectrum or the frequency response is computed by

$$Z(e^{j\omega_i T}) = \frac{Y_s(e^{j\omega_i T})}{U_s(e^{j\omega_i T})}, \qquad i = 1, \ldots, 15, \tag{7.6}$$

where $Y_s(e^{j\omega_i T})$ and $U_s(e^{j\omega_i T})$ are the spectra of the input and output in steady state at frequency ω_i. The subscript 1 stands for lowest frequency (i.e. 21 kHz) and 15 stands for highest frequency (i.e. 1.013 MHz). Thirdly, the impedance spectrum or the frequency response from the FPGA is obtained. An example is shown in Figure 7.4, where 2.5×10^6 Hz is the Nyquist frequency. The frequency responses obtained in the three approaches generally match each other, so the model is validated.

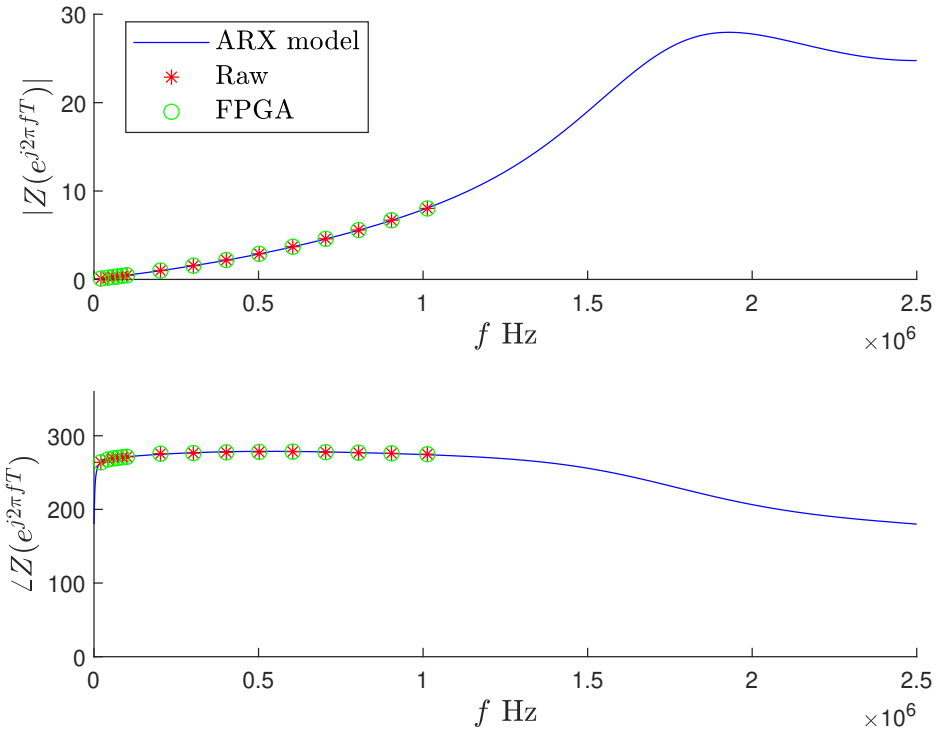


Figure 7.4: The comparison of the impedance computed from the ARX model, the spectrum of raw data, and the FPGA direct results.

7.4 Impedance calibration

To remove the effect of the ambient conditions and the device itself, the impedance obtained in the last section should be calibrated. Three references, i.e. Air1, Air2, and Ferrite1, are used for the calibration. In Figure 7.5, the 3 references and the 7 events are shown using the impedance provided by FPGA. The 10 consecutive frames started from Air event are used as Air1, and the 10 consecutive frames started from Ferrite target event are used as Ferrite1. There is no event flag triggered for Air2. We simply use the 10 consecutive frames which are 100 frames before Cervix 1 event triggering. The two air references are used for zeroing the device, where Air1 is for the zeroing of Ferrite1 and Air2 is for the zeroing of the following events, like Cervix 1, Virginal 1, etc. The Ferrite1 is used for phase reference, as the ferrite target generates a known phase shift ($-\pi/2$). The calibration procedure is listed as the following.

- Calibrate the impedance of Ferrite1 Z_{f1} to Air1 Z_{a1} by

$$Z_{fa} = Z_{f1} - Z_{a1}. \tag{7.7}$$

- As the real phase for Z_p is $-\pi/2$, the impedance measured by this device should shift phase by $\Delta\theta$.

$$\Delta\theta = -\frac{\pi}{2} - \angle Z_{fa}. \tag{7.8}$$

- The phase adjusted impedance is given by

$$Z_\theta = \cos\Delta\theta + i\sin\Delta\theta. \tag{7.9}$$

- Calibrate the impedance of other events Z_e to Air2 Z_{a2} by

$$Z_{ea} = Z_e - Z_{a2}. \tag{7.10}$$

- Shift the phase of the calibrated impedance Z_{ea} by $-\pi/2$.

$$Z_{ea\theta} = Z_{ea}Z_\theta. \tag{7.11}$$

It is noticed that the 3 references from FPGA measurements are only available at the 15 frequencies. However, the impedance obtained by the ARX model is continuous over a frequency range (i.s. from 0 to Nyquist frequency). To solve this problem, a 3^{rd} degree polynomial curve is fitted into the 15 frequencies, and the 3 references over the frequency range is approximated. The real part and the imaginary part are fitted separately, which are shown in Figure 7.6. Through the 3^{rd} degree polynomial models of Air1, Air2, and Ferrite1, the continuous references are obtained.

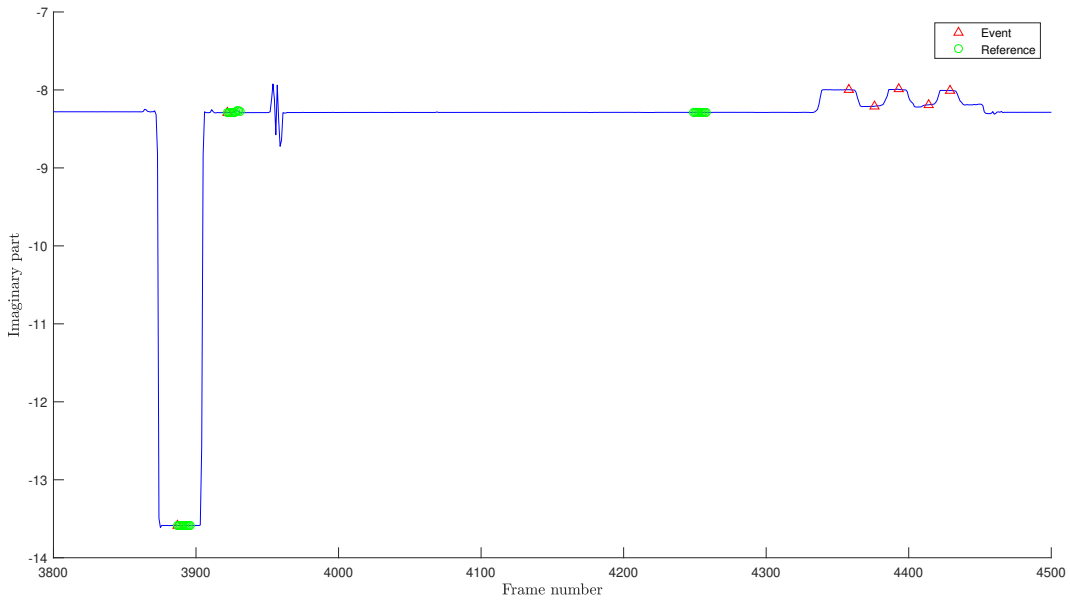


Figure 7.5: The frames used for calibration. The blue line is the imaginary part of the impedance at 1.013 MHz given by FPGA. The 3 green areas from left to right are Ferrite1, Air1, and Air2. Each green area has 10 consecutive frames.

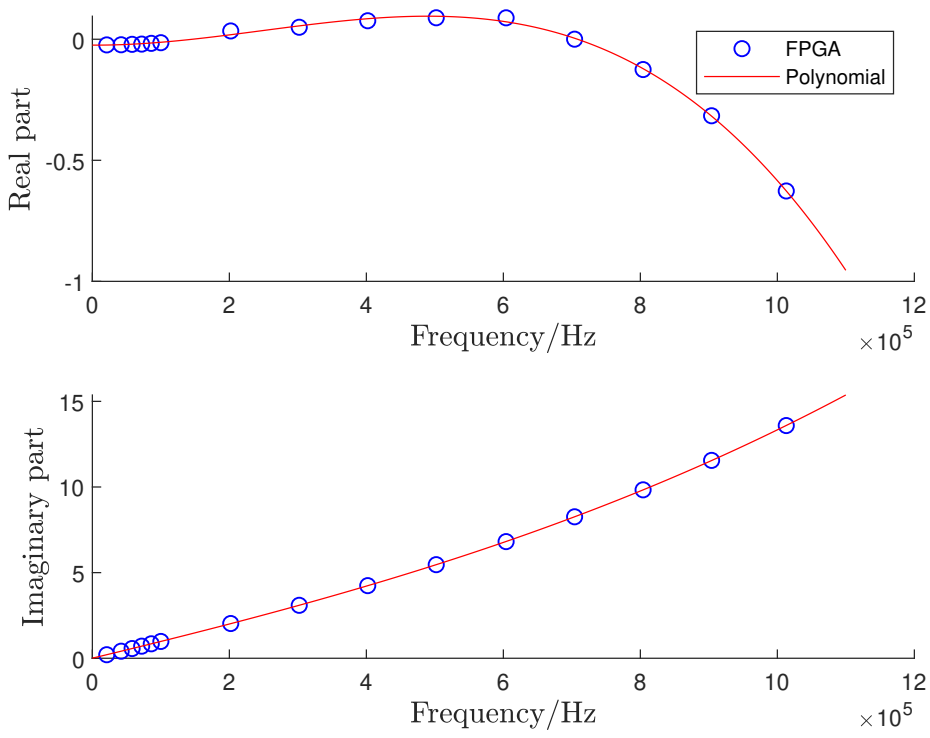


Figure 7.6: The impedance of Ferrite1 reference from 0 to 1.1 MHz.

7.5 Feature extraction

After the impedance calibration, we average the 4 consecutive frames of Cervix 1, Cervix 2, and Cervix 3, and the three averaged impedance spectra are used for the sPTB prediction. For the FPGA measured impedance, there are only 15 discrete points on the spectra. For the ARX model predicted impedance, the completed spectrum from 0 to 2.5 MHz (i.e. Nyquist frequency) is obtained. However, in Figure 7.7, we notice that the frequency range for the training data is rather narrow, although the transient states are used in the training. The 15 transient states used as training data are shown in Figure 7.8. The highest frequency last 24 ms, while the other frequencies occupy 12 ms each. Except the first transient state is 1 ms, the rest transient state is 2 ms. Therefore, instead of using the completed spectrum, the model predicted impedance from 0 to 1.1 MHz is used for sPTB prediction.

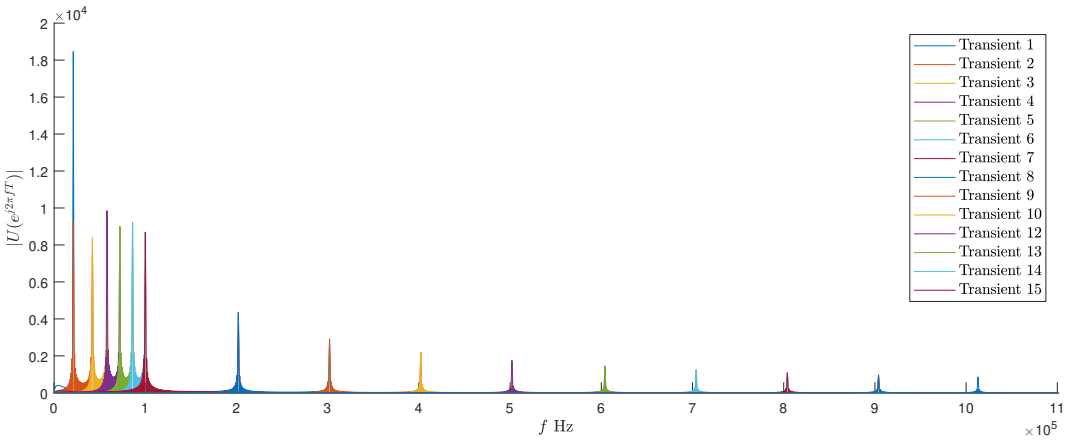


Figure 7.7: The spectra of the training input u.

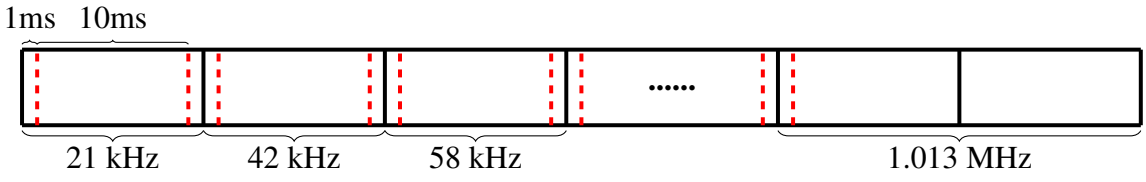


Figure 7.8: The transient and steady states of a frame.

It is found that the impedance in the low frequencies cannot effectively reflect the characteristics of the cervix tissue. An example of 21 kHz impedance is shown in Figure 7.9. It is hard to distinguish the difference among the air event, the cervix events, and virginal events.

Therefore, it is suggested that only use the impedance at the high frequencies (i.e. from 502.1 kHz to 1.1 MHz) for sPTB.

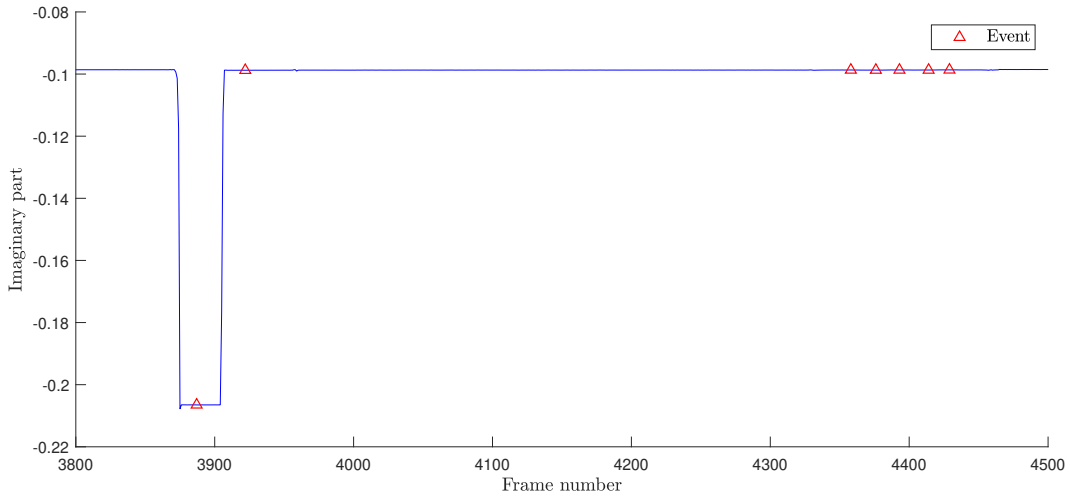


Figure 7.9: The imaginary part of the impedance at 21 kHz.

The absolute value of the impedance is effected by the device, so the gradient value computed by (7.12) is used as the feature.

$$F(\omega_2) = 2\pi \frac{Z(e^{j\omega_2 T}) - Z(e^{j\omega_1 T})}{\omega_2 - \omega_1}, \tag{7.12}$$

where ω_1 and ω_2 are neighboured frequencies, and $\omega_2 > \omega_1$. From 502.1 kHz to 1.013 MHz, only 6 impedance points are measured by FPGA impedance, so 5 gradient impedance can be computed. Thus, there are 10 features including real parts and imaginary parts of the 5 impedance gradients. The difference $\omega_2 - \omega_1$ used for the FPGA impedance gradient evaluation is about 100 kHz. For the ARX model predicted impedance, the difference $\omega_2 - \omega_1$ can be smaller to obtain more accurate gradient, and 100 Hz difference is used in this study. If the frequency interval of $F(\omega)$ is 10 kHz, 60 impedance from 502.1 kHz to 1.1 MHz can be used for sPTB prediction. Therefore, there are 120 features from the ARX model including real parts and imaginary parts.

7.6 Results and discussion

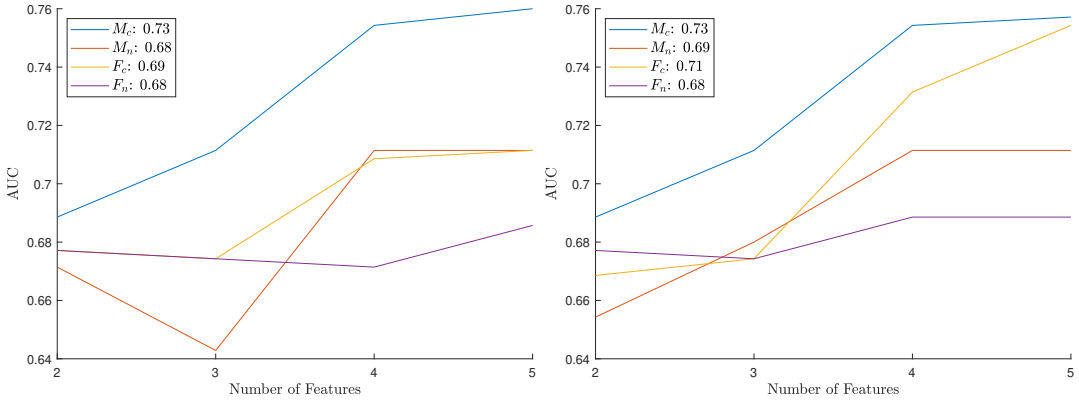
The MIS data are measured from 39 pregnant women, of which 25 are full term birth and 14 are sPTB. In the measurement, each woman has 7 events flagged, in which the impedance

spectra of Cervix 1, Cervix 2, and Cervix 3 are used for sPTB prediction. The impedance spectra are obtained from both the FPGA and the ARX model. After the calibration and gradient impedance extraction, 10 features are obtained from the FPGA measurements, and 120 features are obtained from the ARX model prediction. The OBCC method introduced in Section 3.3.2 and OLS method proposed in the last section are applied to select the significant features. Then, the selected features are given into the logistic regression model. The AUC of the ROC is used to evaluate the classification performance of the logistic regression models. In this analysis, the features are categorised in three ways and compared.

1. The FPGA features and the ARX model features.
2. The calibrated features and the non-calibrated features.
3. Cervix 1 features, Cervix 2 features, and Cervix 3 features.

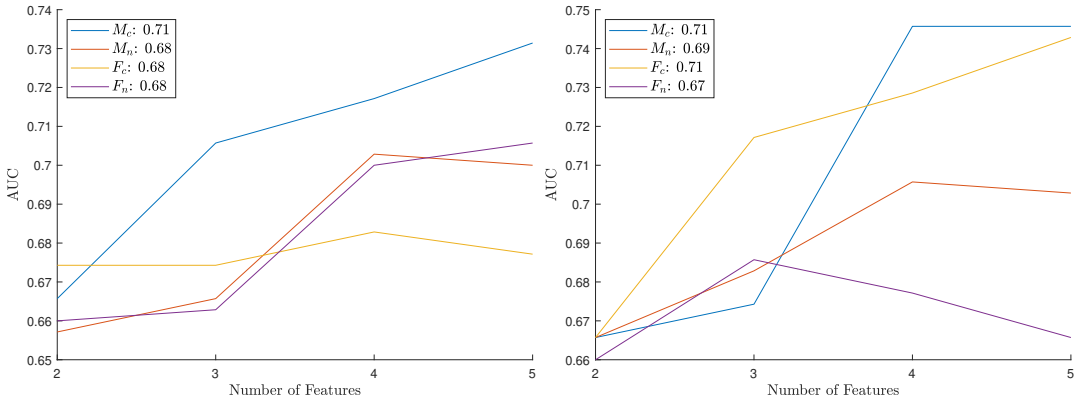
The results are given in the following figures, where M_c denotes the calibrated ARX model features, M_n denotes the non-calibrated ARX model features, F_c denotes the calibrated FPGA features, and F_n denotes the non-calibrated FPGA features. The mean value of the AUC over 2 to 5 features are shown in the legends. Through comparing the results in the figures, the following conclusions can be obtained.

1. The number of features are from 2 to 5. Generally, more features gives better (higher) AUC.
2. The ARX model features are generally better than the FPGA features. The reason could be that the gradient features obtained from the ARX model features are more accurate due to the smaller $\omega_2 - \omega_1$ in (7.12). Another obvious reason is that the feature pool for the ARX model features (120 features) are much more that the FPGA features (10 features).
3. The calibrated features are generally better than the non-calibrated features. This result confirms the effectiveness of the impedance calibration.
4. Generally, the classification performance becomes worse when the features move from Cervix 1 to Cervix 3. The reason could be that the Cervix 3 are effected more by the temperature drift than the Cervix 1 [129].



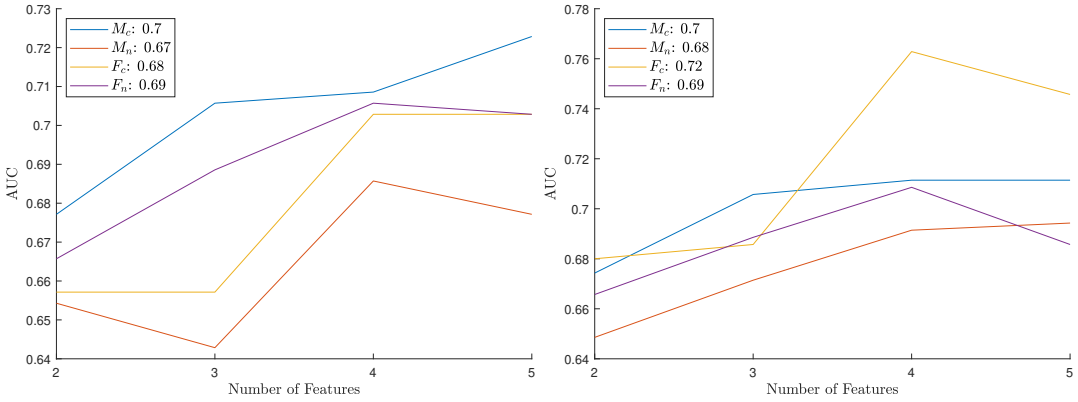
(a) Features selected by OBCC. (b) Features selected by OLS.

Figure 7.10: The AUC of the logistic regression model for Cervix 1 features.



(a) Features selected by OBCC. (b) Features selected by OLS.

Figure 7.11: The AUC of the logistic regression model for Cervix 2 features.



(a) Features selected by OBCC.

(b) Features selected by OLS.

Figure 7.12: The AUC of the logistic regression model for Cervix 3 features.

7.7 Conclusions

In this chapter, the MIS data are analysed for the sPTB prediction. Compared to the traditional method which only makes use of the frequency domain measurement, this chapter provides a new idea of obtaining the impedance spectra from the time domain measurement. The transient part of the time domain data are applied to train the ARX models. Then, the impedance spectra are evaluated by the frequency responses of the models. The new method provides the entire continue impedance spectra rather than discrete impedance points on the spectra obtained from the traditional method. Through the results of this chapter, it is demonstrated that the impedance calibration is necessary. In addition, we also suggest that the future study can focus on the impedance from Cervix 1 event, which is affected less by the temperature drift. The analysis in this chapter shows the potential capability of the MIS data in sPTB prediction, and provides a guidance for the future study.

Chapter 8

Conclusions and Future Works

This chapter lists the main contributions in Chapters 4 to 7 and the direction of the future works.

8.1 Contributions

The PhD research work aims to develop the new methods for system frequency feature extraction and selection of features for machine learning oriented classification and to apply these methods in the fault detection and medical diagnosis. The four contribution chapters are Chapters 4 to 7. The key contributions are as follows.

- In Chapter 4, a new modelling and model feature extraction method is introduced for the system characteristics monitoring. The NARX models are built with the system input-output data, and then the NOFRF features are extracted from both the data driven and real system models. As the NOFRFs of the data driven and real system models match each other, it is verified that the proposed method effectively can reflect the characteristics of the real system and tracks the characteristics change of the real system. Specifically, the proposed method is examined in a nonlinear system with two stable equilibria A and B, and three NARX models are built to represent the system dynamics around A, B, and both A and B. It is found that the NARX model built for A (or B) can only reveal the characteristics around A (or B), while the NARX model

built for both A and B can effectively reflect the characteristics over both equilibria of the original system.

- Motivated by the work in Chapter 4, the model sensor method is proposed in Chapter 6 to detect the wind turbine rotor winding failure with SCADA data. The model sensor is expected to monitor the change of the relationship between the generator temperature, the wind speed, and the ambient temperature. The structure of the model sensor is designed by the first principles. The parameters of the model sensor is updated monthly by the PEM method. After the model sensor is built, the NOFRFs features are extracted from each model sensor. The method to evaluate NOFRFs under different operating points is provided. Based on the proposed model sensor method, the strategy of the alarm system which can be applied in the real wind farms is developed. The results show the proposed model sensor method and NOFRFs analysis can not only indicate the time of the failure happen, but also reveal the ageing process of the winding insulation.

- In Chapter 5, a OLS based feature selection method for classification is proposed. The properties of the squared orthogonal correlation coefficients are analysed, and the relationship with canonical correlation coefficient and Fisher's criterion is revealed. For a greedy search, it has been demonstrated that computing the multiple correlation coefficient or the canonical correlation coefficient using the OLS based method is faster than that directly using the definitions. Besides the fast speed and the statistical meaning, the proposed method can also be used to deal with both continuous and categorical features. In addition, as no requirement of tuning hyperparameters and discretising features, the proposed method is more convenient to be applied in the preprocessing stage of classification.

- In Chapter 7, the MIS data and the feature selection technique proposed in Chapter 5 are used for sPTB prediction. The new idea of extracting the impedance spectra from the time series model is proposed. It is found that the new model features generally give better prediction performance than the traditional FPGA features. This chapter also compares the calibrated features with the non-calibrated features. The results show

that the impedance calibration improves the sPTB prediction performance significantly.

These application studies demonstrate that the proposed methods have great potential to be used in many engineering system fault detection and medical diagnosis related applications.

8.2 Future works

The future work of this thesis can be carried out in the following aspects.

- The NOFRFs and FRF analysis proposed in the thesis highly rely on the model sensor method. The core of the model sensor method is system identification. However, the stability of the nonlinear system identification is still an open problem. The stability of the model sensors directly affects the success of the NOFRFs or FRF extraction. To overcome this issue, using simulation error minimisation (SEM) method rather than PEM method for model parameter estimation might be more likely to generate stable models [130]. In addition, the constraints on the parameters, which is often used to guarantee the stability of linear system identification, might help to ensure the stability of nonlinear model identification [53].

- The accuracy of the NOFRFs and FRF features extracted from model sensors depends on the prediction error of the model sensors. The uncertainty of the NOFRFs and FRF features are important for the fault detection and diagnosis. Through the prediction error, the uncertainty of the NOFRFs and FRF features might be estimated by error propagation law [131] or Bayesian estimation [132].

- The NOFRFs can be estimated under general inputs rather than harmonic inputs only [10]. The fault sensitivity of the NOFRFs under different inputs, such as impulse, step, harmonic inputs, should be investigated.

- Volterra series based nonlinear frequency response functions mentioned in this thesis can not describe sub-harmonic phenomena, which in some fields are key features for

fault detection [133]. It will be investigated whether the idea of GFRF and NOFRF can be extended to depict the sub-harmonic of nonlinear systems.

- In Chapter 5, the correlation based criteria are used for ranking the features. However, the correlation can only describe the linear association between the features and the responses. The feature selection method for the nonlinear features is necessary to be developed in the future.

- Based on the results in Chapter 7, the future MIS analysis can focus on the Cervix 1 event. In addition, the feature used in Chapter 7 is only impedance gradient. With the help of the time series model, the more advanced techniques, such as the bin method [134, 135] and the curve kurtosis [136], can be applied for the feature extraction in the future.

These future works will further develop the proposed methods and help to explore wider application scenarios.

Bibliography

[1] Andrew KS Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510, 2006.

[2] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

[3] Shahab Shokrzadeh, Mohammad Jafari Jozani, and Eric Bibeau. Wind turbine power curve modeling using advanced parametric and nonparametric methods. *IEEE Transactions on Sustainable Energy*, 5(4):1262–1269, 2014.

[4] M Sean Esplin, Michal A Elovitz, Jay D Iams, Corette B Parker, Ronald J Wapner, William A Grobman, Hyagriv N Simhan, Deborah A Wing, David M Haas, Robert M Silver, et al. Predictive accuracy of serial transvaginal cervical lengths and quantitative vaginal fetal fibronectin levels for spontaneous preterm birth among nulliparous women. *Jama*, 317(10):1047–1056, 2017.

[5] MP O'Connell, J Tidy, SJ Wisher, NJ Avis, BH Brown, and SW Lindow. An in vivo comparative study of the pregnant and nonpregnant cervix using electrical impedance measurements. *BJOG: An International Journal of Obstetrics & Gynaecology*, 107(8):1040–1041, 2000.

[6] Rolf Isermann. *Fault-diagnosis systems: an introduction from fault detection to fault tolerance*. Springer Science & Business Media, 2006.

[7] Andrew D Dimarogonas. Vibration of cracked structures: a state of the art review. *Engineering fracture mechanics*, 55(5):831–857, 1996.

[8] Rolf Isermann. Model-based fault-detection and diagnosis–status and applications. *Annual Reviews in control*, 29(1):71–85, 2005.

[9] SA Billings and MI Yusof. Decomposition of generalized frequency response functions for nonlinear systems using symbolic computation. *International Journal of Control*, 65(4):589–618, 1996.

[10] ZQ Lang and SA Billings. Energy transfer properties of non-linear systems in the frequency domain. *International Journal of Control*, 78(5):345–362, 2005.

[11] A.V. Oppenheim, A.S. Willsky, and S.H. Nawab. *Signals and systems*. Prentice-Hall signal processing series. Prentice Hall, 1997.

[12] Rolf Isermann and Marco Münchhof. *Identification of dynamic systems: an introduction with applications*. Springer Science & Business Media, 2010.

[13] RB Randall. Cepstrum analysis and gearbox fault-diagnosis. *Maintenance Management International*, 3(3):183–208, 1982.

[14] Robert B Randall. A history of cepstrum analysis and its application to mechanical problems. *Mechanical Systems and Signal Processing*, 97:3–19, 2017.

[15] Robert B Randall and Jerome Antoni. Rolling element bearing diagnostics—a tutorial. *Mechanical systems and signal processing*, 25(2):485–520, 2011.

[16] PD McFadden and JD Smith. Model for the vibration produced by a single point defect in a rolling element bearing. *Journal of sound and vibration*, 96(1):69–82, 1984.

[17] N Tandon and A Choudhury. A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings. *Tribology international*, 32(8):469–480, 1999.

[18] Scott W Doebling, Charles R Farrar, Michael B Prime, et al. A summary review of vibration-based damage identification methods. *Shock and vibration digest*, 30(2):91–105, 1998.

[19] Richard C Dorf and Robert H Bishop. *Modern control systems*. Pearson (Addison-Wesley), 1998.

[20] Benjamin C. Kuo. *Digital Control Systems*. Oxford University Press, Inc., New York, NY, USA, 2nd edition, 1992.

[21] Stephen A Billings. *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons, 2013.

[22] ZK Peng, ZQ Lang, and SA Billings. Crack detection using nonlinear output frequency response functions. *Journal of Sound and Vibration*, 301(3-5):777–788, 2007.

[23] Martin Hemmer and Tor I Waag. a comparison of acoustic emission and vibration measurements for condition monitoring of an offshore drilling machine. *International Journal of Prognostics and Health Management*, page 8, 2017.

[24] PD McFadden and JD Smith. Acoustic emission transducers for the vibration monitoring of bearings at low speeds. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 198(2):127–130, 1984.

[25] M Kendig and J Scully. Basic aspects of electrochemical impedance application for the life prediction of organic coatings on metals. *Corrosion*, 46(1):22–29, 1990.

[26] R Hirayama and S Haruyama. Electrochemical impedance for degraded coated steel having pores. *Corrosion*, 47(12):952–958, 1991.

[27] AA Oskuie, T Shahrabi, A Shahriari, and E Saebnoori. Electrochemical impedance spectroscopy analysis of x70 pipeline steel stress corrosion cracking in high ph carbonate solution. *Corrosion Science*, 61:111–122, 2012.

[28] MC Li and YF Cheng. Corrosion of the stressed pipe steel in carbonate–bicarbonate solution studied by scanning localized electrochemical impedance spectroscopy. *Electrochimica Acta*, 53(6):2831–2836, 2008.

[29] Konstantinos A Sierros, Nicholas J Morris, Karpagavalli Ramji, and Darran R Cairns. Stress–corrosion cracking of indium tin oxide coated polyethylene terephthalate for flexible optoelectronic devices. *Thin Solid Films*, 517(8):2590–2595, 2009.

[30] MP O'connell, NJ Avis, BH Brown, SR Killick, and SW Lindow. Electrical impedance measurements: an objective measure of prelabor cervical change. *The Journal of Maternal-Fetal & Neonatal Medicine*, 14(6):389–391, 2003.

[31] Mark E Orazem and Bernard Tribollet. *Electrochemical impedance spectroscopy*, volume 48. John Wiley & Sons, 2011.

[32] J Ross Macdonald and E Barsoukov. *Impedance spectroscopy: theory, experiment, and applications*. John Wiley & Sons, 2005.

[33] Y-T Su and S-J Lin. On initial fault detection of a tapered roller bearing: frequency domain analysis. *Journal of Sound and Vibration*, 155(1):75–84, 1992.

[34] PD McFadden and JD Smith. Vibration monitoring of rolling element bearings by the high-frequency resonance technique—a review. *Tribology international*, 17(1):3–10, 1984.

[35] Abdullah M Al-Ghamd and David Mba. A comparative experimental study on the use of acoustic emission and vibration analysis for bearing defect identification and estimation of defect size. *Mechanical systems and signal processing*, 20(7):1537–1571, 2006.

[36] Yaguo Lei, Jing Lin, Ming J Zuo, and Zhengjia He. Condition monitoring and fault diagnosis of planetary gearboxes: A review. *Measurement*, 48:292–305, 2014.

[37] G Dalpiaz, A Rivola, and Riccardo Rubini. Effectiveness and sensitivity of vibration processing techniques for local fault detection in gears. *Mechanical systems and signal processing*, 14(3):387–412, 2000.

[38] Eric Bechhoefer and Michael Kingsley. A review of time synchronous average algorithms. In *Annual Conference of the Prognostics and Health Management Society, San Diego, CA, Sept*, pages 24–33, 2009.

[39] Francois Combet and Leonid Gelman. An automated methodology for performing time synchronous averaging of a gearbox signal without speed sensor. *Mechanical systems and signal processing*, 21(6):2590–2606, 2007.

[40] Frédéric Bonnardot, Mohamed El Badaoui, RB Randall, J Daniere, and François Guillet. Use of the acceleration signal of a gearbox in order to perform angular resampling (with limited speed fluctuation). *Mechanical Systems and Signal Processing*, 19(4):766–785, 2005.

[41] Mitchell Lebold, Katherine McClintic, Robert Campbell, Carl Byington, and Kenneth Maynard. Review of vibration analysis methods for gearbox diagnostics and prognostics. In *Proceedings of the 54th meeting of the society for machinery failure prevention technology*, volume 634, page 16, 2000.

[42] A P Bovsunovsky and C Surace. Considerations regarding superharmonic vibrations of a cracked beam and the variation in damping caused by the presence of the crack. *Journal of Sound and Vibration*, 288(4-5):865–886, 2005.

[43] ZK Peng, ZQ Lang, and FL Chu. Numerical analysis of cracked beams using nonlinear output frequency response functions. *Computers & Structures*, 86(17-18):1809–1818, 2008.

[44] ZK Peng, ZQ Lang, C Wolters, SA Billings, and K Worden. Feasibility study of structural damage detection using narmax modelling and nonlinear output frequency response function based analysis. *Mechanical Systems and Signal Processing*, 25(3):1045–1061, 2011.

[45] Joseph Wang. Electrochemical biosensors: towards point-of-care cancer diagnostics. *Biosensors and Bioelectronics*, 21(10):1887–1892, 2006.

[46] Brian H Brown, John A Tidy, Karen Boston, Anthony D Blackett, Rod H Smallwood, and Frank Sharp. Relation between tissue structure and imposed electrical current flow in cervical neoplasia. *The Lancet*, 355(9207):892–895, 2000.

[47] Lingyan Feng, Yong Chen, Jinsong Ren, and Xiaogang Qu. A graphene functionalized electrochemical aptasensor for selective label-free detection of cancer cells. *Biomaterials*, 32(11):2930–2937, 2011.

[48] Ruimin Wang, Jing Di, Jie Ma, and Zhanfang Ma. Highly sensitive detection of cancer cells by electrochemical impedance spectroscopy. *Electrochimica Acta*, 61:179–184, 2012.

[49] Sheng Chen, Stephen A Billings, and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, Nov. 1989.

[50] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, Mar. 2003.

[51] Yuzhu Guo, L.Z. Guo, S.A. Billings, and Hua-Liang Wei. An iterative orthogonal forward regression algorithm. *International Journal of Systems Science*, 46(5):776–789, 2015.

[52] Long Zhang, Kang Li, Er-Wei Bai, and George W Irwin. Two-stage orthogonal least squares methods for neural network construction. *IEEE transactions on neural networks and learning systems*, 26(8):1608–1621, 2015.

[53] Torsten Söderström and Petre Stoica. *System identification*. Prentice Hall, 1989.

[54] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, Aug. 2005.

[55] Thomas M Cover and Jan M Van Campenhout. On the possible orderings in the measurement selection problem. *IEEE transactions on systems, man, and cybernetics*, 7(9):657–661, 1977.

[56] Jose Roberto Ayala Solares, Hua-Liang Wei, and Stephen A Billings. A novel logistic-narx model as a classifier for dynamic binary classification. *Neural Computing and Applications*, pages 1–15, 2017.

[57] Jacob Cohen and Patricia Cohen. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Wiley, Hillsdale, MI, USA, 1975.

[58] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[59] Cèsar Ferri, José Hernández-Orallo, and Peter A Flach. A coherent interpretation of auc as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 657–664, 2011.

[60] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*. Wiley, Hoboken, NJ, USA, 3rd edition, 2013.

[61] Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

[62] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010.

[63] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[64] Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley. Model selection: Beyond the bayesian/frequentist divide. *Journal of Machine Learning Research*, 11(Jan):61–87, 2010.

[65] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[66] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

[67] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[68] George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57, 2010.

[69] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331, 1983.

[70] Paul Glendinning. *Stability, instability and chaos: an introduction to the theory of nonlinear differential equations*, volume 11. Cambridge university press, 1994.

[71] Ivana Kovacic and Michael J Brennan. *The Duffing equation: nonlinear oscillators and their behaviour*. John Wiley & Sons, 2011.

[72] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press, 2018.

[73] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, Jan. 1996.

[74] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, Boca Raton, FL, USA, 1984.

[75] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205, Jun. 2005.

[76] W. W. Cooley and P. R. Lohnes. *Multivariate Data Analysis*. Wiley, New York, NY, USA, 1971.

[77] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, USA, 3rd edition, 2009.

[78] Tingkai Sun and Songcan Chen. Class label versus sample label-based cca. *Applied Mathematics and Computation*, 185(1):272–283, Feb. 2007.

[79] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, Sep. 1936.

[80] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of machine learning research*, 13:27–66, 2012.

[81] David D Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*, pages 212–217, Harriman, New York, USA, 1992.

[82] Howard Hua Yang and John Moody. Data visualization and feature selection: New algorithms for nongaussian data. In *Advances in Neural Information Processing Systems*, volume 12, pages 687–693, Denver, Colorado, USA, 1999.

[83] François Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

[84] Dahua Lin and Xiaoou Tang. Conditional infomax learning: an integrated framework for feature extraction and fusion. In *European Conference on Computer Vision*, pages 68–82, Graz, Austria, 2006.

[85] Aleks Jakulin. *Machine learning based on attribute interactions*. PhD thesis, University of Ljubljana, Slovenia, 2005.

[86] Patrick E Meyer and Gianluca Bontempi. On the use of variable complementarity for feature selection in cancer classification. In *Workshops on Applications of Evolutionary Computation*, pages 91–102, Budapest, Hungary, 2006.

[87] Steve Sawyer, Sven Teske, and Morten Dyrholm. Global wind energy outlook 2016. *Global Wind Energy Council*, 2016.

[88] Steve Sawyer and Morten Dyrholm. Global wind report 2016. *Global Wind Energy Council*, 2016.

[89] Bin Lu, Yaoyu Li, Xin Wu, and Zhongzhou Yang. A review of recent advances in wind turbine condition monitoring and fault diagnosis. In *Power Electronics and Machines in Wind Applications*, pages 1–7, Lincoln, NE, USA, 2009. IEEE.

[90] Peng Sun, Jian Li, Caisheng Wang, and Xiao Lei. A generalized model for wind turbine anomaly identification based on scada data. *Applied Energy*, 168:550–567, 2016.

[91] Wenxian Yang, Richard Court, and Jiesheng Jiang. Wind turbine condition monitoring by the approach of scada data analysis. *Renewable Energy*, 53:365–376, 2013.

[92] Christopher S Gray and Simon J Watson. Physics of failure approach to wind turbine condition based maintenance. *Wind Energy*, 13(5):395–405, 2010.

[93] Yanhui Feng, Yingning Qiu, Christopher J Crabtree, Hui Long, and Peter J Tavner. Use of scada and cms signals for failure detection and diagnosis of a wind turbine gearbox. In *European Wind Energy Conference and Exhibition 2011, EWEC 2011*, pages 17–19. Sheffield, 2011.

[94] Meik Schlechtingen, Ilmar Ferreira Santos, and Sofiane Achiche. Wind turbine condition monitoring based on scada data using normal behavior models. part 1: System description. *Applied Soft Computing*, 13(1):259–270, 2013.

[95] Meik Schlechtingen and Ilmar Ferreira Santos. Wind turbine condition monitoring based on scada data using normal behavior models. part 2: Application examples. *Applied Soft Computing*, 14:447–460, 2014.

[96] Yingning Qiu, Yanhui Feng, Peter Tavner, Paul Richardson, Gabor Erdos, and Bindi Chen. Wind turbine scada alarm analysis for improving reliability. *Wind Energy*, 15(8):951–966, 2012.

[97] Michael Wilkinson, Brian Darnell, Thomas Van Delft, and Keir Harman. Comparison of methods for wind turbine condition monitoring with scada data. *IET Renewable Power Generation*, 8(4):390–397, 2014.

[98] Braulio Barahona, Cyprien Hoelzl, and Eleni Chatzi. Applying design knowledge and machine learning to scada data for classification of wind turbine operating regimes. In *2017 IEEE Symposium Series on Computational Intelligence*, pages 1–8, Honolulu, HI, USA, 2017. IEEE.

[99] Eduardo J Alvarez and Adrijan P Ribaric. An improved-accuracy method for fatigue load analysis of wind turbine gearbox based on scada. *Renewable Energy*, 115:391–399, 2018.

[100] Petr Kadlec, Bogdan Gabrys, and Sibylle Strandt. Data-driven soft sensors in the process industry. *Computers and Chemical Engineering*, 33(4):795–814, 2009.

[101] Petr Kadlec and Bogdan Gabrys. Local learning-based adaptive soft sensor for catalyst activation prediction. *AIChE Journal*, 57(5):1288–1301, 2011.

[102] Xiaodong Jia, Chao Jin, Matt Buzza, Wei Wang, and Jay Lee. Wind turbine performance degradation assessment based on a novel similarity metric for machine performance curves. *Renewable Energy*, 99:1191–1201, 2016.

[103] Onder Uluyol, Girija Parthasarathy, Wendy Foslien, and Kyusung Kim. Power curve analytic for wind turbine performance monitoring and prognostics. In *Annual Conference of the Prognostics and Health Management Society*, volume 2, pages 1–8, Montreal, Canada, 2011.

[104] M Lydia, S Suresh Kumar, A Immanuel Selvakumar, and G Edwin Prem Kumar. A comprehensive review on wind turbine power curve modeling techniques. *Renewable and Sustainable Energy Reviews*, 30:452–460, 2014.

[105] C Carrillo, AF Obando Montaño, J Cidrás, and E Díaz-Dorado. Review of power curve modelling for wind turbines. *Renewable and Sustainable Energy Reviews*, 21:572–581, 2013.

[106] Zhiqiang Ge and Zhihuan Song. A comparative study of just-in-time-learning based methods for online soft sensor modeling. *Chemometrics and Intelligent Laboratory Systems*, 104(2):306–317, 2010.

[107] Chao Shang, Fan Yang, Dexian Huang, and Wenxiang Lyu. Data-driven soft sensor development based on deep learning technique. *Journal of Process Control*, 24(3):223–233, 2014.

[108] Simon Gill, Bruce Stephen, and Stuart Galloway. Wind turbine condition assessment through power curve copula modeling. *IEEE Transactions on Sustainable Energy*, 3(1):94–101, 2012.

[109] Ramin S Esfandiari and Bei Lu. *Modeling and analysis of dynamic systems*. CRC Press, Boca Raton, 2014.

[110] Ola Aglen. Loss calculation and thermal analysis of a high-speed generator. In *Electric Machines and Drives Conference*, volume 2, pages 1117–1123, Madison, 2003. IEEE.

[111] Junji Tamura. Calculation method of losses and efficiency of wind generators. In *Wind Energy Conversion Systems*, pages 25–51. Springer, 2012.

[112] Donghwa Shin, Sung Woo Chung, Eui-Young Chung, and Naehyuck Chang. Energy-optimal dynamic thermal management: computation and cooling power co-optimization. *IEEE Transactions on Industrial Informatics*, 6(3):340–351, 2010.

[113] Paul M Frank. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: A survey and some new results. *automatica*, 26(3):459–474, 1990.

[114] Yingning Qiu, Yanhui Feng, Juan Sun, Wenxiu Zhang, and David Infield. Applying thermophysics for wind turbine drivetrain fault diagnosis using scada data. *IET Renewable Power Generation*, 10(5):661–668, 2016.

[115] Robert L Goldenberg, Jennifer F Culhane, Jay D Iams, and Roberto Romero. Epidemiology and causes of preterm birth. *The lancet*, 371(9606):75–84, 2008.

[116] Honest Honest, Lucas M Bachmann, Janesh K Gupta, Jos Kleijnen, and Khalid S Khan. Accuracy of cervicovaginal fetal fibronectin test in predicting risk of spontaneous preterm birth: systematic review. *Bmj*, 325(7359):301, 2002.

[117] The Lancet. Preterm birth: what can be done? *The Lancet*, 371(9606):2, 2008.

[118] Dieter Wolke and Renate Meyer. Cognitive status, language attainment, and prereading skills of 6-year-old very preterm children and their peers: the bavarian longitudinal study. *Developmental medicine and child neurology*, 41(2):94–109, 1999.

[119] AL Stewart, L Rifkin, PN Amess, V Kirkbride, JP Townsend, DH Miller, SW Lewis, DPE Kingsley, IF Moseley, O Foster, et al. Brain structure and neurocognitive and behavioural function in adolescents who were born very preterm. *The lancet*, 353(9165):1653–1657, 1999.

[120] Mozziyar Etemadi, Philip Chung, J Alex Heller, Jonathan A Liu, Larry Rand, and Shuvo Roy. Towards birthalert—a clinical device intended for early preterm birth detection. *IEEE Transactions on Biomedical Engineering*, 60(12):3484–3493, 2013.

[121] Jay D Iams, Robert L Goldenberg, Paul J Meis, Brian M Mercer, Atef Moawad, Anita Das, Elizabeth Thom, Donald McNellis, Rachel L Copper, Francee Johnson, et al. The length of the cervix and the risk of spontaneous premature delivery. *New England Journal of Medicine*, 334(9):567–573, 1996.

[122] Robert L Goldenberg, Brian M Mercer, Paul J Meis, Rachel L Copper, Anita Das, Donald McNellis, The NICHD Material Fetal Medicine Units, et al. The preterm prediction study: fetal fibronectin testing and spontaneous preterm birth. *Obstetrics & Gynecology*, 87(5):643–648, 1996.

[123] Harry M Georgiou, Megan KW Di Quinzio, Michael Permezel, and Shaun P Brennecke. Predicting preterm labour: current status and future prospects. *Disease markers*, 2015, 2015.

[124] Jigna Shah, Bhavya Baxi, et al. Identification of biomarkers for prediction of preterm delivery. *Journal of Medical society*, 30(1):3, 2016.

[125] Camelia Gabriel, Sami Gabriel, and y E Corthout. The dielectric properties of biological tissues: I. literature survey. *Physics in medicine & biology*, 41(11):2231, 1996.

[126] Michael D O'Toole, Liam A Marsh, John L Davidson, Yee Mei Tan, David W Armitage, and Anthony J Peyton. Non-contact multi-frequency magnetic induction spectroscopy system for industrial-scale bio-impedance measurement. *Measurement Science and Technology*, 26(3):035102, 2015.

[127] Michael D O'Toole, Noushin Karimian, and Anthony J Peyton. Classification of non-ferrous metals using magnetic induction spectroscopy. *IEEE Transactions on Industrial Informatics*, 14(8):3477–3485, 2017.

[128] Jinxi Xiang, Yonggui Dong, Maomao Zhang, and Yi Li. Design of a magnetic induction tomography system by gradiometer coils for conductive fluid imaging. *IEEE Access*, 7:56733–56744, 2019.

[129] Fritz Primdahl. Temperature compensation of fluxgate magnetometers. *IEEE Transactions on Magnetics*, 6(4):819–822, 1970.

[130] Lennart Ljung. *System identification: theory for the user*. Prentice-hall, 1987.

[131] John Taylor. *Introduction to error analysis, the study of uncertainties in physical measurements*. University Science Books, 1997.

[132] William R Jacobs, Tara Baldacchino, Tony J Dodd, and Sean R Anderson. Sparse bayesian nonlinear system identification using variational inference. *IEEE Transactions on Automatic Control*, 2018.

[133] Fangji Wu and Liangsheng Qu. Diagnosis of subharmonic faults of large rotating machinery based on emd. *Mechanical Systems and Signal Processing*, 23(2):467–475, 2009.

[134] A Llombart, SJ Watson, D Llombart, and JM Fandos. Power curve characterization i: improving the bin method. In *International conference on renewable energies and power quality, Zaragoza*, 2005.

[135] Julia Gottschall and Joachim Peinke. How to improve the estimation of power curves for wind turbines. *Environmental Research Letters*, 3(1):015005, 2008.

[136] Andrew Kusiak and Anoop Verma. Monitoring wind farms with performance curves. *IEEE transactions on sustainable energy*, 4(1):192–199, 2012.