Department of Information and Computer Science

Advances in Wireless Damage Detection for Structural Health Monitoring

Janne Toivola

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 80/2014

Advances in Wireless Damage Detection for Structural Health Monitoring

Janne Toivola

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 18 June 2014 at 12.

Aalto University
School of Science
Department of Information and Computer Science
Parsimonious Modelling

Supervising professor
Prof. Juho Rousu

Thesis advisor
Dr. Jaakko Hollmén

Preliminary examiners
Prof. Hannu Toivonen, University of Helsinki, Finland
Prof. Keith Worden, The University of Sheffield, United Kingdom

Opponent
Dr. Arno J. Knobbe, Leiden University, The Netherlands

Aalto University publication series
DOCTORAL DISSERTATIONS 80/2014

ISBN 978-952-60-5712-5
ISBN 978-952-60-5713-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5713-2

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Janne Toivola

Name of the doctoral dissertation
Advances in Wireless Damage Detection for Structural Health Monitoring

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 80/2014

Field of research Computer and Information Science

Manuscript submitted 16 October 2013 **Date of the defence** 18 June 2014

Permission to publish granted (date) 15 May 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

One of the fundamental tasks in structural health monitoring is to extract relevant information about a structure, such as a bridge or a crane, and reach statistical decisions about the existence of damages in the structure. Recent advances in wireless sensor network technology has offered new possibilities for acquiring and processing structural health monitoring data automatically.

The purpose of this dissertation is to explore various data processing methods for detecting previously unobserved deviation in measurements from accelerometer sensors, based on natural vibration of structures. Part of the processing is projected to be performed on resource constrained wireless sensors to ultimately reduce the cost of measurements.

Data processing in the proposed detection systems is divided into following general stages: feature extraction, dimensionality reduction, novelty detection, and performance assessment. Several methods in each of the stages are proposed and benchmarked in offline experiments with multiple accelerometer data sets. The methods include, for example, the Goertzel algorithm, random projection, tensor decomposition, collaborative filtering, nearest neighbor classification, and evaluating detection accuracy in terms of receiver operating characteristic curves.

Significant reductions are achieved in the amount of data transmitted from sensors and input to statistical classifiers, while maintaining some of the classification accuracy. However, the sensitivity and specificity in detection are worse than those of centralized methods operating on raw sensor data.

The work proposed and evaluated several combinations of data processing stages for wireless damage detection. While better than random detection accuracy can be achieved with very small amount of data per accelerometer sensor, challenges remain in reaching specificity required in practical applications.

Keywords accelerometer data, wireless sensor network, dimensionality reduction, collaborative filtering, novelty detection, structural health monitoring

ISBN (printed) 978-952-60-5712-5 **ISBN (pdf)** 978-952-60-5713-2

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 182 **urn** http://urn.fi/URN:ISBN:978-952-60-5713-2

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Janne Toivola

Väitöskirjan nimi
Edistysaskelia vaurioiden langattomaan ilmaisemiseen rakenteiden kunnonvalvonnassa

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietojenkäsittelytieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 80/2014

Tutkimusala Informaatiotekniikka

Käsikirjoituksen pvm 16.10.2013 **Väitöspäivä** 18.06.2014

Julkaisuluvan myöntämispäivä 15.05.2014 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Eräs rakenteiden kunnonvalvonnan keskeisistä tehtävistä on tuottaa merkityksellistä informaatiota rakenteesta, kuten sillasta tai nostokurjesta, ja tehdä tilastollisia päätelmiä mahdollisten vaurioiden olemassaolosta. Langattomien anturiverkkojen tekniikan kehitys on tarjonnut uusia mahdollisuuksia rakenteiden kunnonvalvontaan liittyvän tiedon automaattiseen mittaamiseen ja käsittelemiseen.

Tämän väitöskirjan tarkoituksena on tarkastella erilaisia tiedonkäsittelymenetelmiä aiemmin havaitsemattomien poikkeamien ilmaisemiseen kiihtyvyysanturimittauksissa, rakenteiden luonnolliseen värähtelyyn perustuen. Osa käsittelystä on suunniteltu suoritettavaksi resurssirajoitetuissa langattomissa antureissa, tavoitteena vähentää mittausten kustannuksia.

Ehdotettujen ilmaisujärjestelmien tiedonkäsittely on jaettu seuraaviin yleisen tason vaiheisiin: piirreirrotus, ulotteisuuden vähentäminen, poikkeavuuden ilmaisu ja suorituskyvyn arviointi. Useita kullekin vaiheelle ehdotettuja menetelmiä vertaillaan keskitettyinä eräajoina suoritetuissa kokeissa usealla kiihtyvyysanturiaineistolla. Menetelminä ovat mm. Goertzelin algoritmi, satunnaisprojektiot, tensorihajotelma, yhteisöllinen suodatus, lähimmän naapurin luokittelu ja ilmaisutarkkuuden arviointi toimintaominaiskäyrien avulla.

Antureilta lähetetyn ja tilastollisille luokittimille syötetyn datan määrässä saavutetaan merkittäviä säästöjä, säilyttäen samalla osa luokittelutarkkuudesta. Vertailussa täysin keskitettyihin, anturiverkossa käsittelemättömään tietoon perustuviin menetelmiin nähden, ilmaisun herkkyys ja spesifisyys kuitenkin heikkenee.

Työssä on ehdotettu ja arvioitu useita langattoman vaurionilmaisun tiedonkäsittelyvaiheiden yhdistelmiä. Satunnaista ilmaisua parempi tarkkuus on saavutettavissa hyvinkin pienellä määrällä mittauksia kiihtyvyysanturia kohden, mutta käytännön sovelluksien vaatiman spesifisyyden saavuttaminen jää haasteeksi.

Avainsanat kiihtyvyysmittaus, langaton anturiverkko, ulotteisuuden vähentäminen, yhteisöllinen suodatus, poikkeavuuden ilmaisu, rakenteiden kunnonvalvonta

ISBN (painettu) 978-952-60-5712-5 **ISBN (pdf)** 978-952-60-5713-2

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2014

Sivumäärä 182 **urn** http://urn.fi/URN:ISBN:978-952-60-5713-2

Preface

"Our knowledge can only be finite, while
our ignorance must necessarily be infinite." – Karl Popper

This research has been funded by Helsinki University of Technology (TKK) / Aalto University School of Science, its Department of Information and Computer Science (ICS), the MIDE research program via the ISMO project, and the HIIT collaboration & Algodan CoE. The doctoral program Hecse provided funding also.

Jaakko Hollmén spent a decade instructing me, while also Jyrki Kullaa and Mikael Björkbom led the way in SHM and WSN. At the end of the process, professors Juho Rousu, Hannu Toivonen, and Keith Worden have kindly reviewed this book.

Besides Jaakko and Jyrki, also Miguel A. Prada and Maurizio Bocca have spent significant amount of time juggling the ideas, performing the experiments, and co-authoring the publications, together with many others in the ISMO project and the WSN group. As a member of HIIT, I had the priviledge of attending a PARAFAC presentation by Aapo Hyvärinen, and a CF presentation by Patrik O. Hoyer, with apparent consequences.

I would also like to thank my past co-workers: Mikko Korpela, Mika Sulkava, Luis Gabriel De Alba Rivera, Prem Raj Adhikari, Jaakko Talonen, and the former Pattern Discovery group.

Last but not the least, I mention the support from Mikko & guys on `#turska`, the amateur radio club OH2TI, and my family: Tiina and Eino.

Laajalahti, Espoo, May 22, 2014,

Janne Toivola

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Janne Toivola and Jaakko Hollmén. Feature Extraction and Selection from Vibration Measurements for Structural Health Monitoring. In *Advances in Intelligent Data Analysis VIII*, Lecture Notes in Computer Science 5772:213–224, Springer, August 2009.

II Janne Toivola, Miguel Á. Prada, and Jaakko Hollmén. Novelty Detection in Projected Spaces for Structural Health Monitoring. In *Advances in Intelligent Data Analysis IX*, Lecture Notes in Computer Science 6056:208–219, Springer, May 2010.

III Miguel Á. Prada, Janne Toivola, Jyrki Kullaa, and Jaakko Hollmén. Three-way analysis of structural health monitoring data. *Neurocomputing*, 80:119–128, Elsevier, March 2012.

IV Janne Toivola and Jaakko Hollmén. Collaborative Filtering for Coordinated Monitoring in Sensor Networks. In *ICDMW 2011 11th IEEE International Conference on Data Mining Workshops*, pages 987–994, IEEE, December 2011.

V Janne Toivola, Jyrki Kullaa, and Jaakko Hollmén. Evaluation Criteria for Energy Efficient Wireless Damage Detection. *Submitted to a journal*, 3rd Dec. 2012 .

Author's Contribution

Publication I: "Feature Extraction and Selection from Vibration Measurements for Structural Health Monitoring"

The present author defined the feature set and found a set of relevant feature extraction methods, given the resource limitations of wireless sensors. The idea of assessing the proposed feature set in terms of supervised classifier performance was formalized by the present author. Experiments on the data set were carried out by the present author. The report was written together with Dr. Hollmén and published in a reviewed conference.

Publication II: "Novelty Detection in Projected Spaces for Structural Health Monitoring"

As feature selection is not possible in supervised or "wrapper-based" way in a general online novelty detection setting, the option of projections to lower-dimensional feature space was identified by Dr. Hollmén. The set of experiments were designed jointly by the present author and Dr. Prada. The initial feature extraction remained as in the first publication, while the projections were made by Dr. Prada and novelty detection phase was carried out by the present author. The report was written jointly by the present author and Dr. Prada and published in a reviewed conference.

Publication III: "Three-way analysis of structural health monitoring data"

Dr. Prada conceived the idea of applying three-way analysis to SHM data. The present author contributed in the feature extraction and novelty detection part of the experiments and some figures. Also the aspect of feature selection, as opposed to dimensionality reduction, was proposed by the present author. The simulated beam data set was provided by Dr. Kullaa. Three-way analysis part of the experiments was performed by Dr. Prada. The report was written mainly by Dr. Prada with review from the other authors and published in a reviewed journal. An earlier version of the article was published in a reviewed conference [70].

Publication IV: "Collaborative Filtering for Coordinated Monitoring in Sensor Networks"

The present author identified the problem of coordinating feature extraction in resource-limited sensor network and formalized a solution in terms of collaborative filtering relying on a domain-specific rating heuristic. Differences between the traditional application of recommender systems and the proposed monitoring application were characterized by the present author. Experiments on the data were performed and reported in a reviewed workshop by the present author. Dr. Hollmén rephrased part of the report and named the concept "coordinated monitoring problem".

Publication V: "Evaluation Criteria for Energy Efficient Wireless Damage Detection"

Issues in evaluating wireless damage detection were identified jointly by the present author and Dr. Kullaa. The present author noticed the problem of using ROC curves in assessing change detection, and proposed the experimental setting of studying trade-offs in AUC values versus sensor output dimensionality in independent detections. Dr. Kullaa proposed visualizing the results by plotting TPR/FPR vs. amount of data per sensor. The wooden bridge measurements, beam simulations, and a half of the detection experiments were performed by Dr. Kullaa, while the present author performed the experiments on methods proposed in this work. The article was written jointly and submitted to a reviewed journal.

Acronyms

ALS	alternating least squares
AR	autoregressive model
ARX	AR with exogenous inputs
AUC	area under (ROC) curve
BN	Bayesian network
CCA	curvilinear component analysis
CF	collaborative filtering
CM	condition monitoring
CORCONDIA	core consistency diagnostic
CPU	central processing unit
CS	compressed sensing
DBN	dynamic Bayesian network
DCT	discrete cosine transform
DFT	discrete Fourier transform
DLAC	damage localization assurance criterion
DST	discrete sine transform
DTMF	dual-tone multi-frequency
EEG	electroencephalogram
EM	expectation maximization
EMD	empirical mode decomposition
FEM	finite element method
FFT	fast Fourier transform
FPR	false positive rate
FRF	frequency response function
IID	independent and identically distributed
IIR	infinite impulse response
ISMO	Intelligent Structural Health Monitoring System

LANL	Los Alamos National Laboratory
LOF	local outlier factor
MCU	microcontroller unit
ML	maximum likelihood
MoG	mixture of Gaussians
MSE	mean square error
NN	nearest neighbor
OCC	overlapping correlation clustering
PARAFAC	parallel factors
PCA	principal component analysis
PDF	probability density function
PSD	power spectral density
QAM	quadrature amplitude modulation
RAM	random access memory
RF	radio frequency
ROC	receiver operating characteristic
RP	random projection
SHM	structural health monitoring
SNR	signal-to-noise ratio
SOM	self-organizing map
TPR	true positive rate
VQ	vector quantization
WSN	wireless sensor network

Notation

c	a value of C, class label
C	a univariate random variable, class
$\hat{C}$	an estimate of C
d	feature index, feature space dimension
D_F	total number of potential features
D_C	number of features input to a classifier
D_S	number of features measured by a sensor
f	frequency (continuous)
f_S	sampling frequency (Hz)
FP	number of false positive detections
FPR	false positive rate
k	DFT index, $k \in [0, K-1]$
K	number of DFT samples
$Mean_t(\cdot)$	sample mean over time
n	discrete time index, $n \in [0, N-1]$
N	time window length
$\mathcal{N}(\mu, \sigma)$	Normal distribution, mean μ and SD σ
$\mathcal{N}(\mu, \mathbf{\Sigma})$	Multivariate normal distribution
$\mathcal{O}(\cdot)$	asymptotic upper bound
$p(X\|c)$	PDF of $X \in \mathbb{R}$, given class $C = c$
$P(c\|X)$	conditional probability of class c, given X
P_D	probability of detection
P_F	probability of false alarm
$r^s[k]$	rating of feature k at sensor s
r	component index
R	tensor rank, number of components
$\mathbb{R}^{S \times N}$	space of S-by-N arrays of real numbers
s	sensor index, $s \in [1, S]$

S	number of sensors
t	time window index, "epoch", $t \in [1, T]$
$1:T$	sequence of integers, $\{t \in \mathbb{Z} \vert t \in [1, T]\}$
T_{CF}	number of time windows used for CF
T_{tr}	number of time windows used for training
$T^{s_1,s_2}[k]$	transmissibility magnitude between sensors s_1 and s_2 at DFT sample k
TP	number of true positive detections
TPR	true positive rate
θ	a model parameter
$\hat{\theta}$	an estimate of a model parameter
$Var_t(\cdot)$	sample variance over time
$w^{s_1,s_2}[k]$	rating for feature k, based on sensors s_1 and s_2
x^2	x squared
x^i	input excitation to a structure
x^s	measurement at sensor s
x_t	data for tth time window
$x[n]$	discrete-time signal x at time $n \in [0 : N-1]$
$X[k]$	DFT of $\mathbf{x}$ at sample $k \in [0 : K-1]$
$\vert X \vert$	magnitude of $X \in \mathbb{C}$
$\vert \mathbf{X} \vert$	determinant of $\mathbf{X} \in \mathbb{R}^{D \times D}$
$\underline{/X}$	phase of complex number X
$\mathbf{x}$	a vector
$\Vert \mathbf{x} \Vert$	Euclidean/Frobenius norm of $\mathbf{x}$
$\mathbf{x} \circ \mathbf{y}$	outer product
$\mathbf{x} * \mathbf{y}$	element-wise product
$\mathbf{X}$	a matrix
$\mathbf{X}'$	a modified matrix, e.g., centered
$\mathbf{X}^T$	transpose of $\mathbf{X}$
$X_{i,j}$	element $(i, j) \in [1 : I] \times [1 : J]$ of matrix $\mathbf{X}$
$X_{i,J_1:J_2}$	elements $(i, j) \in [i] \times [J_1 : J_2]$ of matrix $\mathbf{X}$
$\mathbf{x}_i$	the ith (column) vector of $\mathbf{X}$
$\mathcal{X}_{ij}$	distance $\Vert \mathbf{x}_i - \mathbf{x}_j \Vert$
$\mathbf{X}_{1:T}$	sequence of matrices from $\mathbf{X}_1$ to $\mathbf{X}_T$
$\underline{\mathbf{X}}$	a tensor
$\mathbf{X}_{(t)}$	mode-t unfolding of tensor $\underline{\mathbf{X}}$

1. Introduction

1.1 Background

This piece of research was performed as a part of a multidisciplinary research project, called Intelligent Structural Health Monitoring System (ISMO). In short, the project was about applying wireless sensor network (WSN) technology and data analysis techniques to measure ambient vibrations in large structures and infer the current condition of the structure from the measurements.

Examples of such large structures include bridges, cranes, towers, masts, wind turbines, and airplanes, where the installation and maintenance of a post-hoc wired sensor system would be costly. Although some of the structures may appear static and mostly at rest, they are subject to various random mechanical loading such as earthquakes, wind, snow, or traffic. There are some consequences: a small, but measureable vibration traveling across the material, and damage to the structure via fatigue, corrosion etc. Eventually, damages progress to the scale where the material breaks under the load and the structure fails. Detecting the damage well before failure would aid in minimizing the costs by timing the maintenance and the use of structure appropriately. This work elaborates on the prospect of using wireless accelerometer sensors to monitor the vibrations in a structure and detect damages indirectly by their effect on *transmissibility*.

The focus is in the intersection of two established problem domains: structural health monitoring (SHM) [28] and WSNs [19]. On the SHM side, two major problems are the lack of sensors that would directly indicate damages in a structure and the lack of prior measurement samples from a damaged structure [83]. This highlights the need to infer the condition of the structure indirectly from some other measurements, and the

need for methods whose inference is based on comparing the new data to previous data from a healthy structure only. On the WSN side, the problems are related to the goal of minimizing the cost, energy consumption, and bandwidth of the wireless devices: lack of on-chip memory and the speed of computation and communication.

In the intersection of the two disciplines, lies the problem of coupling between communication in a network and the distributed data processing in a monitoring application. On one hand, routing and other operations in the network communication stack define how data are delivered from a set of sensor nodes to the user. On the other hand, the monitoring application defines which part of the information is worth acquiring and how the data can be fused on its way from sensor node to the user. The ultimate goal, even beyond this work, would be to organize both networked communication and intermediate data processing stages so that the *lifetime* [23] of the resulting monitoring system is maximized.

This leads to various choices of focus: is the operational efficiency optimized only at the networking layer, only at the application layer, independently on both layers, or somehow co-designed on both layers? If the approach was to design a monitoring system independently on both layers, communication and data processing separately, it would result in merely replacing wires with radio links. Despite the interest towards the co-design approach, this work considers WSNs mostly as sets of independent wireless sensors which are able to process their local measurements before transmitting the results to further centralized processing.

1.2 Objectives and Scope

The purpose of this work was to study how data from wireless accelerometer sensors can be processed into a binary decision about the existence of damage in a monitored structure. *Damage* is defined as a state of suboptimal performance, exceeding a defined threshold for insignificant *defects* present in all materials, but smaller than an obvious *fault* or *failure* [28, 83]. The long term objective is a monitoring system which can be easily deployed to a structure, can automatically determine its operating parameters from its initial measurements, and finally provides alerts if there are significant changes to the properties of the structure.

The scope could be broadly described as *wireless damage detection*. This term encompasses three essential constraints of the scope:

- wireless sensors set certain limitations for distributed data fusion,
- the target being detected is a property of the monitored structure, not a property of the environment, and
- the interest is not in determining the continuous extent or location of the damage, but its existence: either the monitored structure has a damage or it doesn't.

1.2.1 Wireless Sensors

This work concentrates on data measured with accelerometers. Besides accelerometers, vibrations in a structure can be monitored with, for example, terrestrial laser scanners and ground-based radar interferometers [67]. Large damages on the surface of a structure could also be observed by cameras, avoiding analysis of vibration altogether. However, the scope of this work includes only accelerometers as they can be embedded into WSN nodes on a single chip. Other sensors possible to integrate, like temperature and humidity sensors, are not covered in this work.

In addition to sensing, another fundamental feature of WSN nodes is their capability of executing some local computation before transmitting the results to the centralized user. This is expected to be useful, if the communicated results are represented by significantly smaller amount of data than the original raw data measured by the sensor, and thus saves part of the energy and bandwidth used for communication. The major limitations on local computation are the available random access memory (RAM) and central processing unit (CPU) time on the WSN nodes. This work considers relatively inexpensive microcontroller units (MCUs) and therefore puts strict constraints on what is plausible to compute locally on the nodes.

There is a "chicken or the egg" dilemma with co-designing WSN together with SHM: if wireless sensors have to discard some piece of information from the raw measurements, how can those discarding algorithms be developed in the first place. This work avoids the problem by considering complete accelerometer data sets measured with *wired* sensors in laboratory conditions and then experimenting what would happen, if the communication was constrained.

1.2.2 Monitoring a Structure, Not Its Environment

One of the major problems considered in this work is that monitoring is performed passively: the monitoring system has no control over what kind of excitation is *input* to the structure. The objective is to measure (*output*) vibration that appears naturally in the structures in their operational environment.

It's important to note that the measured vibrations themselves do not indicate damage [83]. Especially, the amplitude of vibration varies according to the environment. For example, vibrations in a bridge may change with traffic or wind. Thus, one of the challenges is to eliminate environmental variability in the data [55, 22, 74]. In the scope of this work, the variability of the excitation amplitude is considered.

In addition to external forces, there are also other environmental variables, like temperature and humidity, that influence the behavior of structures. This work considers only accelerometer measurements in approximately constant conditions. A research environment with variable, and preferably controlled, temperature and humidity conditions would have been required for including those other variations in the analysis.

1.2.3 Detection

Given a set of data measured and delivered by a WSN, the purpose of an SHM system is to provide information about the current state of the monitored structure. In the scope of this work, the state of the structure needs to be inferred indirectly from the sensor data according to some mathematical model [83].

Besides environmental variability, there are properties of the structure which a set of accelerometers cannot measure and may be difficult to provide for the monitoring system, or in the modeling approach in general. Two examples of such properties are the exact geometry and material of the structure. The above mentioned "easy deployment" of a sensor network partly implies that the monitoring system cannot rely on models that require accurate information about the geometry or the material of the structure. Approximate geometry could potentially be inferred from the observed WSN radio link parameters, assuming the physical distance between the sensor nodes is reflected in the realized radio links of the network.

The (un)availability of geometrical information is an example of the

differences between (physical) model-driven and data-driven approaches mentioned in [84]. If the geometry and some material properties of the structure are known, a detailed physics-based model can be used for inferring relatively accurate estimates about the extent and location of the damage. Such approaches include finite element method (FEM) [84] and damage localization assurance criterion (DLAC) [36]. This work focuses on data-driven approaches in identifying the existence of damage in unsupervised setting as defined in [83], so FEM and DLAC are out of the scope. Thus, the terms *detection* and *classification* are used.

Yet another fork in the road is the choice between a *supervised* and an *unsupervised* setting: in the supervised setting, training data are available from both healthy and damaged condition, while the unsupervised setting attempts to identify unforeseen damages [83]. In some cases, the monitored structure may be simple enough or manufactured in large numbers, so that there are data about the damaged conditions and modes of failure available. However, this work assumes the monitored structure is unique and no prior data from any damage are available for tuning the model parameters. Thus, this work considers *novelty detection* [7, 78], where the monitoring samples of data are compared to a model of normal behavior estimated from initial samples of data collected during healthy state of the structure.

The detection scope does not include the problem of damage localization, one of the SHM steps listed in [28]. The methods discussed in this work attempt to detect damages in any part of the structure, although the issue of analyzing the exact *coverage* of the sensor network is left for future work. Such analysis would have required more extensive data sets.

There is certain amount of granularity in WSNs: if a network is divided into two subnetworks and the other half detects a damage, but the other half doesn't, the damage may be considered roughly localized. Thus, the localization process might not be completely independent from detection.

1.3 Research Process and Dissertation Structure

The structure of this thesis can be described as a "bottom-up" approach: starting with the accelerometer measurements and proceeding through various stages of data processing towards the final detection and its assessment. The stages of the proposed SHM system are:

- feature extraction (Chapter 2),
- dimensionality reduction (Chapter 3),
- novelty detection (Chapter 4), and
- performance assessment (Chapter 5).

The particular problems identified and studied in this work, as opposed to current literature, are denoted with **Problem** and general level reasons leading to them are **Requirements**.

At first, it was studied what the raw accelerometer measurements are like and how they could be processed locally on a WSN node. Then, the work progressed in studying what kind of combinations of features are required on the global level to achieve invariance to certain environmental variability. These are covered in Chapter 2.

As the combinations of features grew prohibitively large, the next processing stage is dimensionality reduction. Chapter 3 discusses the proposed approaches how the extracted features can be mapped into another feature space with fewer dimensions. There are two objectives for this: reduce the number of communicated features and lower the input space dimensionality of the subsequent classifier stage.

Third processing stage is proposed to consist of a general purpose novelty detector: a model of normality learned from initial non-damaged structure and means for comparing new data to the established reference. The considered methods are discussed in Chapter 4.

The final stage is about assessing the performance of damage detection systems on an experimental basis, as described in Chapter 5. The detections from the previous stage are compared to the "gold standard" labels of an experimental data set to provide empirical estimates of the detection accuracy. Energy efficiency is considered in terms of the amount of data transmitted by each sensor node.

The publications do not follow this bottom-up order of data processing, but a more iterative approach. Publication I does focus on feature extraction, but presents also the further proposed processing stages and the idea of assessing the system in terms of the final output. Then, each aspect of the prototype system are iteratively developed towards better functionality. It could also be called a holistic approach: considering an SHM system as a whole and assessing only the final output, instead of attempting to optimize single intermediate stages of a system. However, what is considered a system here could be considered a subsystem elsewhere.

The work began with older and simple methods first and then proceeded towards studying more complex methods. This approach is supposed to behave like the Occam's razor: more complex methods are justified only if they provide significantly better performance or functionality. Combined with limited time, this had also the side effect of pruning out the most advanced methods in the classifier stage and only basic novelty detector models were included. Also, separating feature extraction and dimensionality reduction from novelty detection left the latter with less emphasis: if the earlier stages perform well, the detection should be simple.

This work can be seen as an evolution of a demonstration system, instead of developing optimizations to a cost function defined in advance. This shows especially in the fact that assessment criteria were studied last. The experimental results are explained in Chapter 6 and discussed in Chapter 7.

2. Feature Extraction Methods for Wireless SHM

2.1 Problem Setting

2.1.1 Accelerometer Measurements

The current work began with a data set measured earlier for [55]. It provided an example of what kind of data are considered in vibration-based SHM. The measurements can be represented as a coarse scale time series of matrices:

$$\mathbf{X}_{1:T} = [\mathbf{X}_1, \ldots, \mathbf{X}_T]. \tag{2.1}$$

In a broader scope, monitoring systems deal with *data streams* [32], which means that there is no fixed end T, but an endless sequence of measurements made by the sensor network. On the other hand, the experiments in this work have been performed on finite sequences of data, so that a fixed number of measurements, $t \leq T_{tr}$, has been used for estimating model parameters and all the subsequent data, $T_{tr} < t \leq T$, was used for testing the actual online monitoring (described later in Chapter 4).

Each matrix $\mathbf{X}_t$ contains a (fine scale) time series of N accelerometer samples simultaneously measured by S sensors:

$$\mathbf{X}_t = x_t^{1:S}[0 : N-1] = [\mathbf{x}_t^1, \ldots, \mathbf{x}_t^S] \in \mathbb{R}^{N \times S}. \tag{2.2}$$

In this representation, each sensor $s \in \{1, \ldots, S\}$ produces a (column) vector of data for each time window or "epoch" t:

$$\mathbf{x}_t^s = x_t^s[0 : N-1] = [x_t^s[0], \ldots, x_t^s[N-1]]^T. \tag{2.3}$$

If the context of the sensor network and the coarse time scale is ignored, a single sensor node measures N consecutive samples of data at a con-

stant sampling frequency f_S. The sequence of measurements is represented as a vector:

$$\mathbf{x} = x[0 : N-1] = [x[0], \ldots, x[N-1]]^T \in \mathbb{R}^N. \tag{2.4}$$

In the above formulation, the difference in indexing the sensors with superscripts, the coarse time scale with subscripts, and the fine time scale with square brackets emphasizes the idea that the data are assumed to be accessed differently across the distributed sensors s from the way being accessed across the time indices n and t. Local computation performed in the sensor nodes is assumed to access only data measured by the same device.

Requirement 2.1.1. *Local computation: a data processing algorithm distributed to a sensor node can only access the raw data produced on the sensor itself.*

Compared to the coarse time scale, the discrete-time signal $x[n]$ is sampled:

- at a relatively high frequency,
- with regular intervals, and
- for a constant amount of samples N.

On the other hand, the time windows $\mathbf{X}_t$ experience the above mentioned indefinite streaming and potentially irregular time intervals. One of the main reasons for this "burstiness" is sensor node duty cycling: the sensor nodes perform other tasks, such as the radio communication or writing data to mass storage, between the sensing cycles. This kind of periodical measurements are made even with wired accelerometers.

Example 2.1.1. *Raw accelerometer data measured with wired sensors: The data set used in [55] consisted of $T = 2509$ measurement periods, each of which contained simultaneous data from $S = 15$ accelerometers. Each of the sensors had measured $N = 8192$ acceleration samples at $f_S = 256$ Hz. Thus the time window length was $8192/256 = 32$ s. The time interval between each of the time windows is unknown, but the measurements were made over a period of several days.*

The above considerations lead to a requirement that narrows down the options considered in this work:

Requirement 2.1.2. *Intermittent sampling: raw accelerometer samples cannot be measured and streamed continuously, so monitoring needs to be based on separate time windows of acceleration time series.*

Thus, the subsequent feature extraction methods are applied to each of the time windows $\mathbf{X}_t$ separately.

The example of wired accelerometers controlled by a centralized measurement system is ignoring an important aspect of WSNs and *distributed systems* [18] in general: the time synchronization issue. Making measurements simultaneously (enough) in a WSN is a topic beyond the scope of this work, but has been researched by others [17, 88].

2.1.2 Features

Besides the structure of raw data, it is useful to consider what kind of properties of the data convey useful information, or which properties should be considered useless or unwanted. In pattern recognition literature [81, 7], measurements or intermediate statistics computed from raw data are called *features* and the process of computing them is referred to as *feature extraction*. The rationale is to transform the original data so that the actual classification task becomes easier. Typically, this means emphasizing the variability between data samples assigned to different classes or minimizing unnecessary variability in samples belonging to the same class.

The statistical pattern classification problem and methods are discussed later in Chapter 4, but at this stage, the problem from SHM perspective is the lack of direct indication. Accelerometer measurements $\mathbf{X}_t$ do not directly indicate damages in the monitored structure, but convey information on how the structure vibrates at multiple locations in response to ambient excitation. Vibration itself is not interesting, but the structure which serves as a medium for the vibrations.

Now, the first step towards successful identification of damages becomes a problem of computing suitable transformation of the data:

Requirement 2.1.3. *compute intermediate features that are sensitive to changes in the structure, but insensitive to changes in the environment.*

Then, the subsequent classification is performed using the resulting features as input. The pattern recognition problem is supposed to become easier in that knowledge about what is relevant in SHM is incorporated in the feature extraction (and shifted away from the classifier), while also

attenuating environmental variability.

Combined with the WSN objective of minimizing the amount of data transmitted from sensors, feature extraction performed locally on the sensor nodes can help in concentrating the useful part of the information in fewer bytes, in a potentially lossy transformation. This idea is elaborated further in Chapter 3, but at this point the focus is on transforming accelerometer data into a *feature space* that would have desirable properties. The goal could be stated as:

Requirement 2.1.4. *Parsimonious feature space: the features computed on sensor nodes should enable efficient representation of relevant information.*

Efficient representation of time series data is a widely researched topic. For example, proposed solutions have included piecewise linear models of the time domain data [52], symbolic representations of the time series [59], and utilizing the coefficients of autoregressive models (ARs) and ARs with exogenous inputs (ARXs) [61, 31]. However, this work is focused on frequency domain features extracted via discrete Fourier transform (DFT) [64], though the power spectral density (PSD) of the signal could be estimated also from AR coefficients [40]. Even an approach of using AR in frequency domain for modeling the DFT has been reported [29].

In the context of condition monitoring (CM) of rotating machinery, fast Fourier transform (FFT) has been used for feature extraction [41, 78]. The notion of *tracked orders* [41] has been used for feature selection (further discussed in Chapter 3), since the vibrations in such machines are expected to concentrate on the same frequency as the rotation, and the harmonic multiples of that frequency. Structures considered in SHM do not generally have rotating parts, speed of which could be measured and utilized in selecting subsets of features, but do exhibit *resonance*: tendency to vibrate on certain frequencies with higher amplitude.

Example 2.1.2. *Accelerometer signal and its* magnitude spectrum*: Figure 2.1 shows an example from the aforementioned data set. The upper panel contains a two-second acceleration signal excerpt from one of the sensors,* $x_1^8[0:511]$*, and the lower panel presents the first half of the corresponding magnitude spectrum* $|X_1^8[0:255]| = |DFT(\mathbf{x}_1^8)[0:255]|$.

Thus, the approach developed in this work is to use DFT: magnitude spectra of the accelerometer signals seem to have sparse patterns, characteristic to the monitored structure.

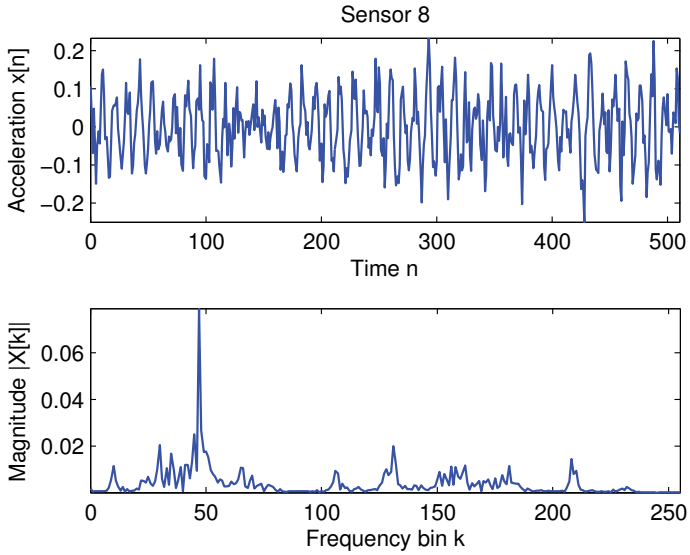


Figure 2.1. An accelerometer signal and its magnitude spectrum.

When shifting the attention from SHM back to WSN, there are still some issues to cover. DFT cannot be computed with FFT algorithms, since it was deemed too complex an operation for the actual processing units of the sensor nodes used in ISMO. More specifically, there was only half a kilobyte of RAM available for data processing, which subsequently limits N, if the raw data need to be buffered before processing.

Example 2.1.3. *Computing an N-point FFT of signal $x[0:N-1]$ on a sensor node requires access to all of the N samples [64]. If the samples are represented with 16 bits (two bytes) each, sampling requires $2N$ bytes of memory that can be written in real-time (at the speed the values are sampled by the sensor).*

Thus, there is a problem with limited local memory: N has to be small because of the memory access on the wireless sensor nodes. Finally, the approach chosen by the current author is that the methods should not only stream data over the time index t, but also consider online data processing in the smaller time scale n: to utilize some local feature extraction algorithm that runs online (over time index n), i.e., is able to discard the raw sensor sample $x[n]$ right after being measured, and has a smaller than $\mathcal{O}(N)$ memory requirement.

In short, the initial problems identified in feature extraction are:

Problem 2.1. *Which signal processing algorithms could be used locally on a WSN node for monitoring parts of vibration spectra?*

Problem 2.2. *Can it be achieved online, without buffering accelerometer data?*

Problem 2.3. *How to achieve invariance to environmental effects?*

2.2 Quadrature Amplitude (de-)Modulation

The first feature extraction method, proposed in Publication I, is influenced by the author's background with radio technology. Quadrature amplitude modulation (QAM) is a method originally developed for encoding digital data in a communication channel [71], such as radio. In short, the transmitter can be described as modulating the amplitude and phase of a sinusoidal radio frequency (RF) carrier wave, so that the parameters of the resulting signal follow discrete points of the so called *constellation diagram*. At the other end of a radio link, the radio receiver demodulates the signal by measuring and identifying the changes in the amplitude and phase.

Example 2.2.1. *Constellation diagram: a signal parameter space of* 16 *separate points can be used for encoding* 4 *bits of data. An example of this is shown in Figure 2.2.*

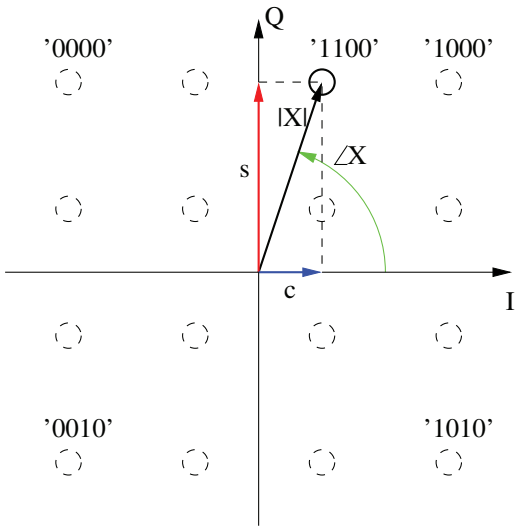


Figure 2.2. A constellation diagram for rectangular 16-QAM and phase vector corresponding to transmitting the symbol "1100".

Now, instead of identifying discrete parameters of an RF signal, the same demodulation approach can be used for monitoring a narrow frequency band of an accelerometer signal. The mechanical vibrations in

a structure are not expected to follow discrete patterns of conventional (digital) QAM, but the principle of demodulation is used for feature extraction.

Monitoring a narrow frequency band around a given frequency f is achieved by first multiplying the input signal $x^s[n]$ with sine waves of orthogonal phase. The intermediate results are:

$$c^s(f) = \frac{1}{N}\sum_{n=0}^{N-1} x^s[n]\cos(2\pi(f/f_S)n + \phi^s), \tag{2.5}$$

and

$$s^s(f) = \frac{1}{N}\sum_{n=0}^{N-1} x^s[n]\sin(2\pi(f/f_S)n + \phi^s), \tag{2.6}$$

where f_S is the sampling frequency and (f/f_S) normalizes the frequency scale relative to it. The sensor-specific phase constant ϕ^s denotes the idea that individual WSN nodes might not have perfectly synchronized clocks for generating coherent sine waves. Since there are methods for accurate enough synchronization [17, 88], the following formulation ignores ϕ^s and focuses on feature extraction on a single sensor node s.

The intermediate values c and s can be computed recursively as follows:

$$c(f,-1) = s(f,-1) = 0, \tag{2.7}$$

$$c(f,n) = c(f,n-1) + \frac{1}{N}x[n]\cos(2\pi(f/f_S)n), \tag{2.8}$$

$$s(f,n) = s(f,n-1) + \frac{1}{N}x[n]\sin(2\pi(f/f_S)n). \tag{2.9}$$

After computing $c(f,N-1)$ and $s(f,N-1)$ over the time window $n \in [0,N-1]$, the magnitude can be computed as illustrated geometrically in Figure 2.2:

$$|X(f)| = \sqrt{c(f)^2 + s(f)^2}. \tag{2.10}$$

The values $c(f)$ and $s(f)$ also contain phase information.

The continuous frequency scale f is discretized into frequency bins $k \in [0, \frac{N}{2}-1]$ by selecting $f[k] = \frac{k}{N}f_S$. An alternative approach of formulating the feature extraction in terms of discrete cosine transform (DCT) and discrete sine transform (DST) is also acknowledged in Publication I.

Example 2.2.2. *Toy data in frequency domain: an artificial signal was composed as a sum of three components, two sine waves with different parameters ($A_1 = 1.0, A_2 = 0.6, f_1 = 10$ Hz, $f_2 = 10.55$ Hz, $\phi_1 = 5\pi/4$, and $\phi_2 = \pi$) and normally distributed noise ($\mathcal{N}(0,1)$). The discrete time scale n was $N = 512$ samples at $f_S = 256Hz$. Figure 2.3 shows the relevant samples $k \in [18,23]$ of FFT magnitude spectra computed from each of the*

signal components separately. The frequency of the second sine wave does not coincide with the frequency bin and shows a non-ideal property of DFT: neighboring DFT samples are not ideally insensitive to the signal, aka. leakage.

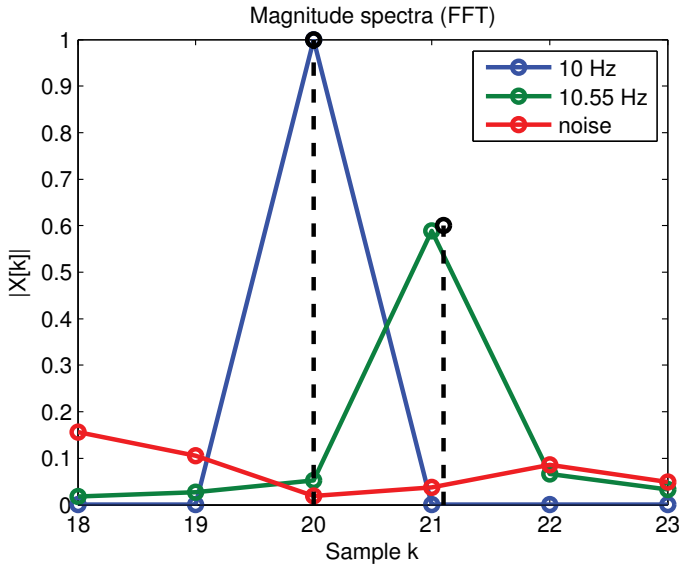


Figure 2.3. Magnitude spectra of the three components of an artificial signal. Dashed lines indicate the actual parameters.

Example 2.2.3. *QAM over toy data: the sum of the three signal components was analyzed with QAM using the same monitoring frequency bins as FFT. Trajectories of* (c, s)*-points (I / Q-components) during the computation and phase vectors of the final results are shown in Figure 2.4.*

2.3 Goertzel Algorithm

The QAM approach requires synthesizing the sinusoidal reference signals, but there is another approach with even less computational requirements, if the phase information is not required. The *Goertzel algorithm* [33] is a method for computing individual terms of DFT magnitude online and it has been used for demodulating dual-tone multi-frequency (DTMF) signals [64], which encode symbols as combinations of tones at certain frequencies.

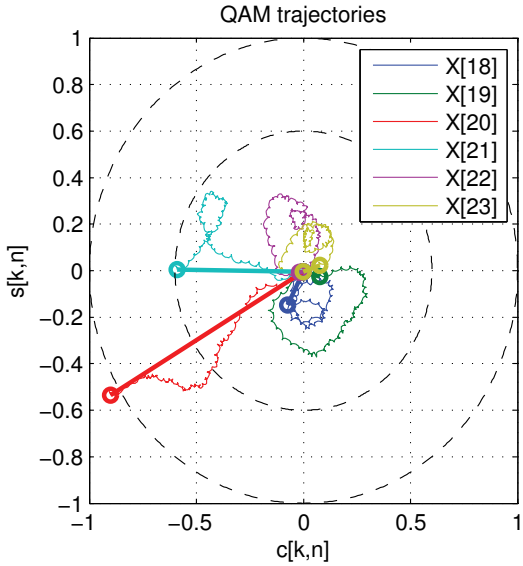


Figure 2.4. An example of how QAM is used for monitoring neighboring frequency bins. Straight lines indicate the final results, dashed circles indicate the actual signal amplitudes.

The iteration steps of the algorithm can be written as [64]:

$$c[k] = 2\cos(2\pi k/N) \tag{2.11}$$

$$v[k,-1] = v[k,-2] = 0, \tag{2.12}$$

$$v[k,n] = x[n] + c[k]v[k,n-1] - v[k,n-2], \forall n \in [0, N-1], \tag{2.13}$$

where $v[k,n-1]$, and $v[k,n-2]$ are the intermediate results needed for computing the square magnitude of DFT sample k after the above iterations:

$$|X[k]|^2 = v[k,N-1]^2 - c[k]v[k,N-1]v[k,N-2] + v[k,N-2]^2. \tag{2.14}$$

The corresponding point on a continuous frequency scale is $f \approx \frac{k}{N} f_S$.

The iteration steps show that the algorithm is effectively a second order infinite impulse response (IIR) filter [64] with specific parameters suitable for monitoring a narrow frequency band and minimizing the required memory.

Example 2.3.1. *Goertzel algorithm over toy data: the above-mentioned artificial data were sampled over $T = 16$ time windows and the Goertzel algorithm was used for monitoring the same six DFT samples $k = [18, 23]$, which correspond to frequencies of about $f = [9, 11.5]$ Hz. The results are shown in Figure 2.5.*

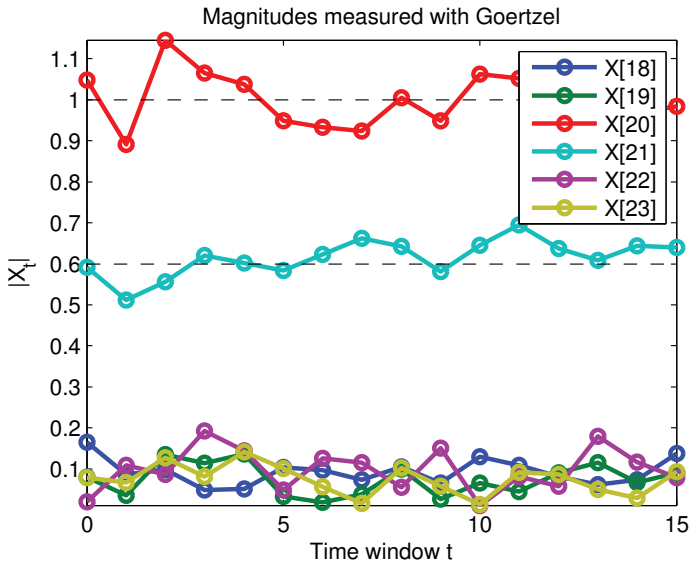


Figure 2.5. An example of monitoring neighboring frequency bins with the Goertzel algorithm. Dashed lines indicate the actual parameters.

From computational complexity point-of-view, it takes linear number of operations for each DFT sample, $2N + 4$ multiplications and $4N + 4$ additions. Multiplied with the number of DFT samples N, the time complexity becomes quadratic, $\mathcal{O}(N^2)$, and is a waste of resources compared to $\mathcal{O}(N \log N)$ complexity achieved by FFT [64], *if* all N samples are required. On the other hand, if at most $D_S \approx \log_2 N$ frequency bins are monitored, the Goertzel algorithm is preferable in terms of computational operations.

In terms of memory requirements, this algorithm needs to store the intermediate values $v[k, -1]$ and $v[k, -2]$ for each bin k, instead of needing the access to the whole input sequence $x[n]$ or use N values for in-place computation of FFT. Thus, it is preferable to use the Goertzel algorithm, if $D_S < N$.

This algorithm was actually implemented and run on the considered WSN hardware [8], and can thus be considered relevant for local feature extraction as originally proposed in Publication I.

2.4 Transmissibility

Being able to monitor magnitudes of an accelerometer signal is not enough for SHM, because of possible magnitude changes with the original source

of vibration. The goal of being able to monitor structures in their operational environment brings the problem of environmental variability [22, 74]. In realistic situations, the environment may have multiple sources of variability, like changes in temperature and humidity, and their (normal) effects on the structure. In the scope of this work, experimental data with other than accelerometer measurements were not available. Thus, this work addresses only one type of variability: changes in the amplitude of input excitation. At least, it demonstrates the need for fusing information from several sensors.

Considering frequency domain analysis of signals, a frequency response function (FRF) [64, 50] refers to the steady-state relation $H^s(\omega)$ between a sinusoidal input signal on a given frequency ω and the output s of a system, which is a monitored point in the structure in this case. On the other hand, considering an arbitrary input signal X^i as a superposition of sinusoidal signals of different frequencies, FRF also characterizes the measured response X^s of the system. In frequency domain:

$$X^s(\omega) = H^s(\omega)X^i(\omega). \tag{2.15}$$

In short, H^s would convey relevant information about the structure between the point of excitation and a sensor, but the sensor s can only observe X^s, and the input X^i remains unknown. One approach to avoiding this problem is to estimate *transmissibility* [50, 85, 84], which refers to the ratio of two responses:

$$T^{s_1,s_2} = \frac{X^{s_1}}{X^{s_2}} = \frac{H^{s_1}X^i}{H^{s_2}X^i} = \frac{H^{s_1}}{H^{s_2}}. \tag{2.16}$$

Thus, it attempts to measure a "dimensionless" quantity describing how well vibration travels between the two sensors. The formulation does rely on constant point of input, as H^s is assumed relative to it, but at least the potential variability in the amplitude of X^i gets eliminated.

Transmissibility is described as a *local* feature [50] in terms of sensitivity to damage. If a damage occurs at a point that is not "structurally" between the sensors s_1 and s_2, then T^{s_1,s_2} would ideally remain unaffected. On the other hand, transmissibility is described as a *global* feature in Publication IV, because it is computed from observations made by two separate sensor nodes: it is not local in terms of computation in a WSN. This is the reason why transmissibility features are considered to be computed centrally, not in the sensor nodes.

At first, there was uncertainty whether a WSN can provide accurate synchronization between measurements at two separate sensors and co-

herent phase information for transmissibility. Like in [85], this work ignores phase information and proceeds by considering *transmissibility magnitude* computed from the DFT magnitude estimates:

$$T^{s_1,s_2}[k] = \frac{|X^{s_1}[k]|}{|X^{s_2}[k]|}, \tag{2.17}$$

thus comparing the magnitude sample k from two sensors s_1 and s_2.

So called "damage indicator" features, which rely on computing ratios between a baseline transmissibility and newly measured ones, like the ones in [49, 50, 8, 62], are avoided in this work, since the responsibility of modeling differences between the non-damaged baseline and the current states of the monitored structure is left to the novelty detector stage.

Transmissibility considerations provide an example of the inability of a single sensor to produce useful results, information of interest being conveyed by a bulk of measurements, and the need for considering collaboration between sensor nodes.

3. Dimensionality Reduction

3.1 Problem Setting

In broader terms, this work considers structures equipped with WSNs as heterogeneous, feature-rich sources of data. Different kinds of sensor elements can be integrated into the wireless nodes and also feature extraction considered in the previous chapter provides a large set of potential measurements to choose from. This work faces the problem arising specifically from transmissibility magnitude features $T^{s_1,s_2}[k]$: a high-dimensional feature space.

Example 3.1.1. *High-dimensional feature space: if a sensor network has* $S = 15$ *accelerometers and each of them monitor* $D_S = K = 256$ ***DFT** samples, there are* $105 = \frac{S(S-1)}{2} \in \mathcal{O}(S^2)$ *unordered pairs of sensors* $\{s_1, s_2\}$ *and* $26880 = \frac{S(S-1)}{2}K \in \mathcal{O}(S^2K)$ *potential transmissibility magnitude features* $T^{s_1,s_2}[k]$ *to monitor.*

This can be considered a problem in two ways. First, it provides a high-dimensional input feature space for the subsequent pattern classifier stage (considered in Chapter 4), which then suffers from *the curse of dimensionality* [81, 7], the need for exponentially large training data sample, and numerical issues in estimating model parameters.

Secondly, the reason to monitor signals in DFT domain is to reduce the number of measurements made and transmitted by each sensor node. It would be useful to *select* which parts of the spectrum are monitored by which sensors.

Feature selection can be based on expert analysis, as in [85], where regions of transmissibility spectrum were selected manually. However, the relevance of features cannot be established on the basis of observing variability related to damages in a pure novelty detection setting, since the

damages are assumed to occur in the unforeseeable future: training data comes only from a non-damaged structure.

Requirement 3.1.1. *Extreme bias / imbalance of data: dimensionality reduction has to be based only on data from a healthy structure. This is referred to as "unsupervised" [83], although it does not imply completely unlabeled data (like unsupervised in [7]), just that all labels happen to be equal for training data. Also the term "semi-supervised" has been used to describe the situation [43].*

Feature space can also be *projected* into a smaller dimensional space, for example, by computing a sum over all frequencies, like the damage index DI in [49]. However, a projection helps only with the dimensionality problem, not the issue of reducing measurements in general. Therefore, the following sections are divided into projection/decomposition methods and a selection approach.

Instead of observing variability caused by damages, the feature selection can be based on some application specific heuristic that indicates preferences based on healthy state observations. One of the main objectives is to replace or shift expert effort from deploying and configuring an SHM system: thus the term "intelligent" SHM system.

In short, the problems are:

Problem 3.1. *Is it enough to monitor only few DFT samples?*

Problem 3.2. *How to reduce the feature space dimensionality from all possible transmissibility magnitudes?*

Problem 3.3. *How to choose the monitored parts of the vibration spectrum for each wireless sensor?*

3.2 Projections

3.2.1 Random Selection

From historical perspective of this work, damage detection was first considered as a *supervised* pattern classification problem: a classifier was trained with data from both non-damaged and damaged states of the structure. In that case, the issue of choosing what to measure could be solved with wrapper-based *feature selection* [34]: the classifier could be

used for evaluating which subsets of features lead to accurate classifications.

Now, if the classifier is "unsupervised" in the sense that only data from the non-damaged class are available for training (aka. *one-class classifier* [79]), features cannot be selected for detecting the future test set anomalies, based on the data alone. Because of this inability to make informed decisions, the research in Publication I is directed towards evaluating classification accuracy with (uniformly) *random* subsets of transmissibility magnitude features.

As noted above, S accelerometer sensors and K-point frequency resolution[1] results in $D_F = \frac{S(S-1)}{2}K$ different parameter combinations for feature $T^{s_1,s_2}[k]$, when ignoring symmetric sensor pairs ($T^{s_2,s_1}[k]$). The possible points in the parameter space are:

$$\mathbb{D} = \{(s_1, s_2, k) \in \mathbb{N}^3 | s_1 \in [1, S-1], s_2 \in [s_1+1, S], k \in [0, K-1]\}. \quad (3.1)$$

A random subset of features is then selected by taking a pseudorandom sample of $D_c << D_F$ points from the parameter space without replacement. The probability of selecting a particular feature to classifier input is uniform:

$$P(T^{s_1,s_2}[k] \text{ selected}) = \frac{2D_c}{S(S-1)K}. \quad (3.2)$$

The above random selection can also be considered as a *projection* from a high-dimensional feature space onto a lower-dimensional one. If all the possible transmissibility magnitude features are addressed sequentially with $d \in [1, D_F]$ instead of $(s_1, s_2, k) \in \mathbb{D}$, for example,

$$d = ((s_2 - s_1) + (s_1 - 1)(S - \frac{s_1}{2}) - 1)K + k + 1, \quad (3.3)$$

and collected into a vector[2] $\mathbf{a} = [T^{1,2}[0], \ldots, T^d, \ldots, T^{S-1,S}[K-1]]$, then feature selection can be represented as a linear transformation:

$$\mathbf{b} = \mathbf{aR}, \quad (3.4)$$

where the $D_F \times D_C$ matrix $\mathbf{R}$ has mostly zero elements and a single 1 on each column indicating the selected feature for the output dimension:

$$R_{d,i} = \begin{cases} 1 & \text{, if } T^d \text{ selected as i:th feature} \\ 0 & \text{, else.} \end{cases} \quad (3.5)$$

Due to this connection between feature selection and projections, Publication II continued the research with studying various transformations

[1] for N-point DFT, $K = N/2$ due to symmetry

[2] a vectorization of an $S \times S$ upper triangle matrix and frequency domain

from high-dimensional feature space to a space with fewer dimensions: the focus was on random projection (RP) [1], principal component analysis (PCA) [69, 44, 7], and curvilinear component analysis (CCA) [20].

3.2.2 Random Projection

The above sparse matrix multiplication resembles a certain kind of RP, which was formulated in [1].

A more general mathematical result, called Johnson-Lindenstrauss lemma after [51], promises that a set of points in a high-dimensional space can be mapped onto a lower-dimensional space with small distortion in the Euclidean distances between the points. After adapting the notation and concepts in terms of this work, the lemma states:

Lemma 3.2.1. *Given small distortion $\epsilon > 0$, a set of T_{tr} points (feature values) in $\mathbb{R}^{D_F}$, and sufficient output dimensionality $D_C > 8\ln(T_{tr})/\epsilon^2$, there exists mapping $f : \mathbb{R}^{D_F} \mapsto \mathbb{R}^{D_C}$, that preserves the distances between each pair of points $\mathbf{u}, \mathbf{v}$ in the set well enough:*

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leq \|f(\mathbf{u})-f(\mathbf{v})\|^2 \leq (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2. \tag{3.6}$$

This would provide means for embedding the training data from the original high-dimensional feature space to moderate dimensional space, in which the used classifier then operates.

It has been shown in [1], that such a mapping can be achieved with a high probability by projecting the data with a random sparse matrix, which satisfies certain requirements:

Theorem 3.2.1. *Given small distortion $\epsilon > 0$, probability parameter $\beta > 0$, a set of T_{tr} feature vectors in a $T_{tr} \times D_F$ matrix $\mathbf{A}$, and a sufficient output dimensionality*

$$D_C \geq \frac{4+2\beta}{\epsilon^2/2-\epsilon^3/3}\log T_{tr}, \tag{3.7}$$

a $D_F \times D_C$ random matrix $\mathbf{R}$, whose elements are distributed as

$$R_{d,i} = \sqrt{3} \times \begin{cases} 1 & ,\textit{with probability } 1/6 \\ 0 & ,\textit{with probability } 2/3 \\ -1 & ,\textit{with probability } 1/6 \end{cases} \tag{3.8}$$

maps the set of points in $\mathbf{A}$,

$$\mathbf{B} = \frac{1}{\sqrt{D_C}}\mathbf{A}\mathbf{R} \tag{3.9}$$

so that with probability of at least $1 - \frac{1}{T_{tr}^{\beta}}$*, the distances between each pair of points* i, j *are preserved*

$$(1-\epsilon)\|\mathbf{a}_i - \mathbf{a}_j\|^2 \leq \|\mathbf{b}_i - \mathbf{b}_j\|^2 \leq (1+\epsilon)\|\mathbf{a}_i - \mathbf{a}_j\|^2. \tag{3.10}$$

Compared to the random selection, this projection provides an "insurance against axis-alignment" [1]: if a relevant difference between data points was manifested in a single feature (dimension), the feature will not be catastrophically omitted in the projection, but likely to affect some of the elements of the output vector. This kind of random projection has been evaluated to perform well in other applications for text and image data [6].

Publication II explores what happens to diagnostic accuracy when D_C is varied over a whole range of values, not just sufficiently[3] large dimensions, and what happens when a fixed projection is used for new measurements. The above random projection results apply to a (training) set of points, with T_{tr} elements, but the subsequent monitoring is performed over a stream of data.

Example 3.2.1. *RP matrix: Figure 3.1 shows a small example of a sparse random projection matrix from* $D_F = 100$ *dimensions to* $D_C = 10$ *dimensions.*

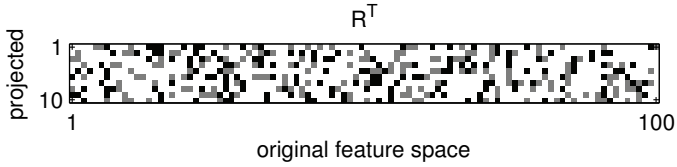


Figure 3.1. An example of sparse random projection matrix. Zero elements are shown in white, negative elements in gray, and positive elements in black.

3.2.3 Principal Component Analysis

PCA, also known as Karhunen-Loève transform, is a classic dimensionality reduction method [69, 44, 81, 7]. Given $T_{tr} \times D_F$ matrix of mean centered training data $\mathbf{A}' = \mathbf{A} - Mean_t(\mathbf{A})$, PCA results in a linear projection $\mathbf{B} = \mathbf{A}'\mathbf{R}$, where $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_{D_C}]$, transforming from the original D_F dimensions to chosen number of dimensions D_C, can be considered to

maximize variance of the projected data, for each output dimension i:

$$argmax_{\mathbf{r}_i} Var_t(\mathbf{A}'\mathbf{r}_i), \tag{3.11}$$

[3] for fixed amount of data T_{tr}

in descending order of variance $Var_t(\mathbf{b}_i) \geq Var_t(\mathbf{b}_j), i < j$, or

minimize projection error in terms of mean square error (MSE):

$$argmin_{\mathbf{r}_i} \sum_t \|\mathbf{a}'_t - \mathbf{b}_t\|^2, \tag{3.12}$$

while assuming $B_{t,D_C+1:D_F} = 0, \forall t$ for the missing projection dimensions,

with the constraint of orthonormal basis: $\mathbf{r}_i \cdot \mathbf{r}_j = 0, \forall i \neq j$ and $\|\mathbf{r}_i\| = 1, \forall i$. The orthogonality requirement results in mutually uncorrelated, different, output features and normalization avoids unbounded solutions with arbitrarily large scaling.

Publication II studied the effect of output dimensionality D_C on diagnostic accuracy. Mean centering is not expected to matter in this application, since the interest is in the potential deviations from the constant non-damaged baseline. On one hand, the transmissibility magnitude feature space is likely to have correlated features, so the new basis may prove useful for some detectors. On the other hand, variance in the training data might not characterize future anomalies, so the projection learned during training phase might not be optimal for certain damage.

Example 3.2.2. *PCA of 2D toy data: an artificial data set* $\mathbf{A}$ *with* $T_{tr} = 200$ *points was sampled from a bivariate normal distribution* $\mathbf{a}_t \sim \mathcal{N}(\mu, \mathbf{\Sigma})$*, where the mean was* $\mu = [4.0, 3.0]$ *and the covariance matrix was*

$$\mathbf{\Sigma} = \begin{bmatrix} 1.5 & 1.0 \\ 1.0 & 1.0 \end{bmatrix} \tag{3.13}$$

PCA was then used for determining the corresponding principal components (or eigenvectors of the sample covariance matrix [81]), $[\mathbf{r}_1; \mathbf{r}_2] = [0.7723, 0.6353; -0.6353, 0.7723]$*, and the variances/eigenvalues,* $Var_t(\mathbf{b}_1) = 2.2708$ *and* $Var_t(\mathbf{b}_2) = 0.2416$*.*

Another sample of $T_{test} = 5$ *points was drawn from slightly shifted distribution* $\mathcal{N}(\mu + \delta, \mathbf{\Sigma})$*, where* $\delta = [1.5, 1.0]$*, and projected to the first principal component axis of the training data. Figure 3.2 shows the shapes of the distributions with dashed ellipses, the sampled training data points in blue, the (re-centered and scaled) principal component vectors as green lines, and the way original test data points (red cross) end up projected to the first principal component axis (red circles). From novelty detection point of view, PCA makes sense if the deviation* δ *follows the selected basis.*

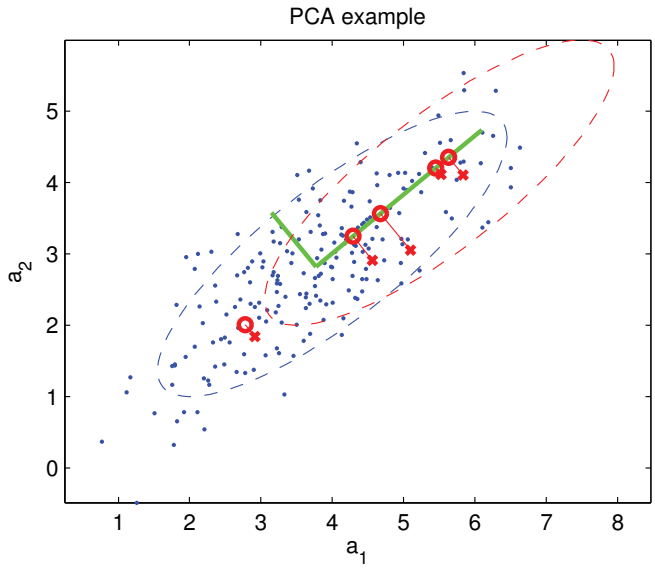


Figure 3.2. An example of projecting bivariate data from a test set (red) to the first principal component (longer green line) learned from a training set (blue).

3.2.4 Curvilinear Component Analysis

The above projections are linear, but there are also nonlinear transformations, which could be used to "disentangle" more complex structure in the data. Transmissibility measurements come from a dynamical system, so the usual variability in the data may take a form of a certain trajectory across the feature space, finally resulting in a set of points that span a some kind of *submanifold* in the space.

One approach to modeling complex structure in data is to use an artificial neural network that fits a set of *model vectors*, like vector quantization (VQ) or self-organizing map (SOM) [53]. Here, the focus is on projecting data from one continuous space to another, so the interest was directed towards CCA [20, 47], which has that ability.

CCA associates a set of model vectors in the input space to corresponding ones in the lower dimensional output space. Learning is based on minimizing a cost function that considers the distances between vectors in the input space, collected in the matrix $\mathbf{A}$, and associated output vectors, $\mathbf{B}$. The cost function used in this work and in Publication II is the kind proposed by [47]:

$$E = \sum_{ij} \begin{cases} (\mathcal{A}_{ij} - \mathcal{B}_{ij})^2 F_\lambda(\mathcal{B}_{ij}) & \text{if } \mathcal{B}_{ij} > \mathcal{A}_{ij} \quad \text{(Unfolding)} \\ (\mathcal{A}_{ij}^2 - \mathcal{B}_{ij}^2)^2 F_\lambda(\mathcal{B}_{ij})/4\mathcal{A}_{ij}^2 & \text{if } \mathcal{B}_{ij} <= \mathcal{A}_{ij} \quad \text{(Projection)}, \end{cases} \tag{3.14}$$

where the inter-point distances in the input space are denoted by $\mathcal{A}_{ij} = \|\mathbf{a}_i - \mathbf{a}_j\|$ and in output space respectively $\mathcal{B}_{ij} = \|\mathbf{b}_i - \mathbf{b}_j\|$. "Unfolding" refers to what happens at larger scale when the manifold is cut open and spread onto the lower dimensional space, and some distances need to change. "Projection" refers to local behavior of the points that maintain their distances during the transformation.

The monotonically decreasing weighting function F_λ governs the importance of output distances, making the preservation of local structure more important than the one manifested by distant points. This work uses an exponential function (unlike the step function used in [47]):

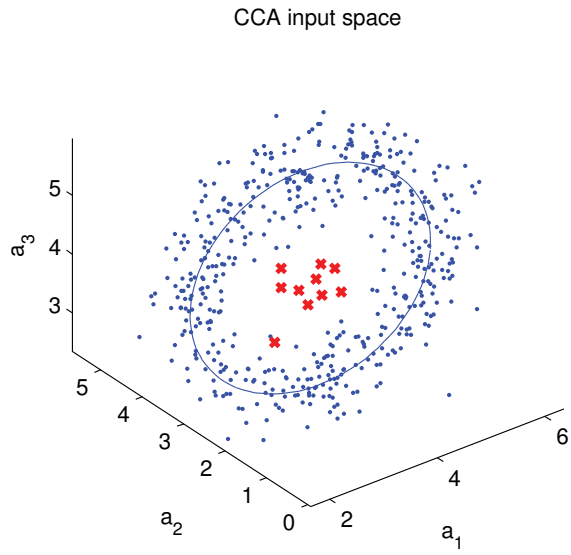
$$F_\lambda(y) = e^{-y/\lambda}, \tag{3.15}$$

where the parameter λ defines the range of what is being considered "local" in the output space.

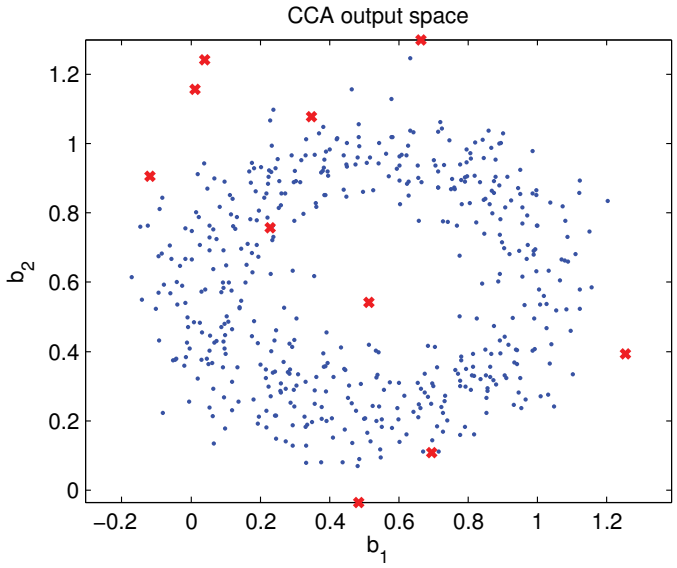
Out-of-sample projections for a new input space vector $\mathbf{a}_t$ are achieved by iteratively fitting a point $\mathbf{b}_t$ in the output space, relative to the model vectors learned from the training set and the above cost function. This can be regarded as interpolation when $\mathbf{a}_t$ belongs to the same region as training set (healthy structure), or extrapolation when the point $\mathbf{a}_t$ is outside of the learned region.

Example 3.2.3. *CCA of 3D toy data: an artificial data set of $T_{tr} = 500$ points from a submanifold, a circle on a tilted plane, was sampled with some additional normally distributed noise ($\mathcal{N}(0; 0.1)$). Also a noisy test set of $T_{test} = 10$ points from the center of the circle was sampled. This is a pathological example from feature selection or Gaussian novelty detector point of view, since test data looks normal (close to mean $\mu = [4, 3, 4]$) for each subset of features despite being clearly different from training data.*

A CCA model was learned from the training data and used for mapping both the training set and test set points to a two dimensional output space. Figure 3.3 shows both sets of points in the input and output spaces. This demonstrates the ability of CCA to preserve the local structure found in training data, and how the novel test data may become extrapolated differently from what would have been achieved by using PCA.



(a) Input space



(b) Output space

Figure 3.3. An example of CCA input and output spaces: blue points represent the training data and novel test data is denoted by red crosses.

3.3 Three-way Analysis

Besides projecting sets of data points from one feature space to another, the problem can be approached in terms of factorizations. Tensor decompositions [54, 65] offer a variety of approaches in approximating a multiway data set as a product of several factors and a sum of several components. One specific tensor decomposition is parallel factors (PARAFAC) [39], also known as *canonical (polyadic) decomposition*, which has been previously used for example in analyzing electroencephalogram (EEG) data [65].

EEG measurements resemble the vibration-based SHM setting in that there are several channels (sensors) monitored over time over several epochs, so it could be argued that the data has three "natural" *ways* or *modes*. This is also manifested in the indexing introduced in Chapter 2:

- accelerometer data are sampled over separate time windows (experiments), time, and sensors ($T \times N \times S$),
- DFT magnitude occurs over time windows, frequency bins, and sensors ($T \times K \times S$), and
- transmissibility magnitudes are measured over time windows, frequency bins, and sensor pairs ($T \times K \times S^2$).

The first case has been studied on strain sensor data in [66] in terms of unfolded (two-way) data and PCA. The three-way model used in this work focuses on the third case, simultaneously explaining the three modes of transmissibility magnitude data: time window (experiment) $\times$ DFT bin $\times$ sensor pair. Thus, a data set with T time windows is considered a three-way tensor:

$$\underline{\mathbf{X}} = \{X_{t,k,s} = T_t^{s_1,s_2}[k] \mid (s_1, s_2, k) \in \mathbb{D}; t \in [1, T]\}, \tag{3.16}$$

where the set of available sensor pairs (s_1, s_2) is addressed with a single index s: related with the *spatial* region of the structure between the two sensors.

The idea of PARAFAC is to decompose the data into a sum of R multilinear *components*, i.e., (rank-one) tensors that are outer products of vectors. If the time factor is represented with vector $\mathbf{a} \in \mathbb{R}^T$, and correspondingly the spectral factor $\mathbf{b} \in \mathbb{R}^K$ and spatial factor $\mathbf{c} \in \mathbb{R}^S$, then the PARAFAC

model can be represented as:

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r, \tag{3.17}$$

with constraints $\|\mathbf{a}_r\| = \|\mathbf{b}_r\| = \|\mathbf{c}_r\|$ to promote uniqueness. Alternatively, explicit normalization constraints [54] can be included on the loadings ($\|\mathbf{a}_r\| = 1$ etc.) by using separate scale parameters λ_r, and residual error denoted by tensor $\underline{\mathbf{E}}$:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r + \underline{\mathbf{E}}. \tag{3.18}$$

The factor vectors can be grouped into *factor matrices* or *loading matrices*, one for each mode:

$$\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_R] \in \mathbb{R}^{T \times R}, \tag{3.19}$$

for temporal mode and similarly for spectral and spatial modes: $\mathbf{B} \in \mathbb{R}^{K \times R}$ and $\mathbf{C} \in \mathbb{R}^{S \times R}$. As opposed to the projections in the previous section, three-way analysis keeps sensor pairs and vibration spectrum separate, thus avoiding the vectorization of Equation 3.3. Also, the decomposition includes temporal mode in the analysis: the data are not just independent and identically distributed (IID) points in $S \times K$ space, when $\mathbf{B}$ and $\mathbf{C}$ are determined.

The PARAFAC model can be written yet another way with the help of *matricization*, i.e. keeping one mode intact and unfolding the remaining modes into one ($\mathbf{X}_{(t)} \in \mathbb{R}^{T \times SK}$), and the Khatri-Rao product (denoted by $\odot$) [54]:

$$\mathbf{X}_{(t)} = \mathbf{A}\mathbf{\Lambda}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}_{(t)}, \tag{3.20}$$

where the component scales are included in a diagonal matrix: $\Lambda_{r,r} = \lambda_r$.

Determining the decomposition for a given data set has two phases: first the number of components R has to be selected, as the components depend on each other and typically cannot be sequentially identified, and then the actual factorization is computed for the particular R. One method for deciding appropriate number of components is a heuristic called core consistency diagnostic (CORCONDIA) [13], which is based on fitting a sequence of less restricted models (Tucker3 [82]) with increasing complexity, and evaluating diagonality of the resulting $R \times R \times R$ *core tensor*. After all, PARAFAC can be viewed as a special case of Tucker3 model, where the core tensor is exactly diagonal.

A popular approach for solving the decomposition itself is the alternating least squares (ALS) [39, 54] method, which is based on fitting each of

the loading matrices at a time, while keeping the others fixed. For example, new temporal factors and scales ($\mathbf{A}' = \mathbf{A}\Lambda$) are solved from the linear least squares problem:

$$\min_{\mathbf{A}'} \|\mathbf{E}_{(t)}\| = \min_{\mathbf{A}'} \|\mathbf{X}_{(t)} - \mathbf{A}'(\mathbf{C} \odot \mathbf{B})^T\|, \tag{3.21}$$

and the same is iterated for other modes respectively. The ALS process is iterated until convergence or certain maximum number of steps. It is not guaranteed to find a global minimum, but in practice repeated with another initial conditions when needed.

Finding a decomposition for a large batch of data is not interesting in itself, from online monitoring point of view. On the other hand, PARAFAC was used for offline analysis of SHM data in [70] and Publication III, where temporal factors $\mathbf{A}$ revealed the variability over time t.

Also an online approach was presented in Publication III for monitoring and novelty detection. The idea is to estimate spectral ($\mathbf{B}$) and spatial loadings ($\mathbf{C}$) from training data, and subsequently use them for finding temporal factors $\mathbf{a}_t \in \mathbb{R}^R$ for each new sample of transmissibility data $\mathbf{X}_t \in \mathbb{R}^{K \times S}$, via linear regression:

$$\min_{\mathbf{a}_t} \|\mathbf{X}_t - \mathbf{B}diag(\mathbf{a}_t)\mathbf{C}^T\|, \tag{3.22}$$

i.e., letting time explain damages (or other changes). Using $\mathbf{a}_t$ as the input to the subsequent classifier means the feature space dimensionality is reduced to $D_C = R$.

It may also prove interesting to find patterns in the R different components of $\mathbf{a}_t$ and the fixed loadings $\mathbf{B}$ and $\mathbf{C}$. For example, some components may specialize in explaining parts of spectrum at certain subsets of sensor pairs. In fact, Publication III also tried *selecting* features, based on training set PARAFAC, by picking the frequency bins and sensor pairs with highest loadings, for each R components. Then, monitoring phase can be performed with a model trained on the selected subset of data, and new monitoring samples with reduced amount of measurements.

3.4 Feature Selection with Additional Heuristics

3.4.1 Idea from Collaborative Filtering

Even a "database-friendly" random projection [1] of the centrally computed (*global*) features does not guarantee sparsity at the local feature

level. All the DFT features would have to be measured on each WSN node despite efficient centralized dimensionality reduction: a WSN with multiple layers of information fusion is not a centralized database. The resource limitations have to be considered from individual sensor node point of view. In this work, this refers to the limitation of monitoring only few DFT samples on each sensor.

On the other hand, the relevance of measurements made by individual sensors is evaluated according to the information fused from the measurements. Especially in the present application, it is not useful to monitor one part of vibration spectrum on one sensor and a disjoint part of the spectrum on another sensor, as this would not admit computation of any transmissibility features or detection of damages. "Coordinated monitoring" is required, as concluded in Publication IV.

Third complication is the lack of principled selection criterion in the general novelty detection setting: potentially any feature could change independently, so there is "no free lunch". In practice, some application-specific preferences exist, but they need to be defined both from the individual sensor and the collective utility/cost point of view.

One application that has the above problems – sparsity, a gap between individual and collective point of view, and a prediction setting – is that of *recommender systems* [72, 75]. Unrelated to SHM, recommender systems consider the problem of predicting a small subset of interesting *items* for a set of *users*. Collaborative filtering (CF) is a class of approaches in solving the problem by assuming few earlier ratings from the users and assuming some structure in the resulting sparse users-by-items rating matrix: if two users rate some of the items similarly, they probably would rate also the other items similarly.

Inspired by CF, Publication IV presents the idea of applying similar methods to selecting suitable features for monitoring in WSN setting: instead of users-by-items rating matrix, the problem revolves around a sensors-by-features rating matrix. Wireless sensor nodes are seen as a community of (lazy or resource limited) voters that attempt to collaboratively figure out what features are worth monitoring.

In the simplest scenario, certain amount of initial measurements ($t \in [1 : T_{CF}]$) are used for the collaborative filtering process, and then the process continues with ordinary novelty detector training phase ($t \in [T_{CF}+1 : T_{tr}]$) and online monitoring ($T_{tr} < t$), using the filtered features.

Similarities and Differences to Recommendation Systems

As stated above, features could be selected by using application-specific principles. The monitoring application may have simple rules which make some features more desirable than others, or some features obviously useless in relation to others. This kind of rules can be used in producing ratings about the available features and measurements, like users of recommendation systems produce ratings for available items.

On the other hand, limitations of sensor network technology can lead to *sparsity* of the rating data. For example, it may be infeasible to make simultaneous measurements about all modalities available to a sensor node, or it can be impractical to make all possible measurements and transmit them to a centralized controller for further pruning. Thus, a monitoring system may suffer sparsity problems similar to recommendation systems, where users typically give ratings only for few items.

In order to extract further features, to add more layers of information fusion, suitable combinations of measurements need to be made. So the question, which measurements are appropriate for a single sensor, depends on what the other sensors are measuring. This resembles the situation in recommendation systems, where users are assumed to follow more general patterns in their ratings. Just as some customers of an online bookstore are interested in reading books of a certain genre, some sensors of a monitoring system might specialize in certain measurements.

Collaborative filtering literature [75] has identified also other related problems in recommendation systems. One of them is the *cold start* problem: how does the system start producing recommendations when new users or items are added and none of the rating data is available yet. In the sensor network applications, this can be solved by running a separate initialization phase: assuming the sensor network is in centralized control, the administrator can command any (new) sensors to rate any (new) features. Conventional recommendation systems don't have such additional control over its participants.

Scaling is another problem familiar from recommendation systems [75]. Such a system may have millions of items, like news articles, and millions of users. Future applications in environmental monitoring may also have increasing amounts of sensors and possible measurements to make, but this work concentrates on an SHM example with a limited set of sensors and available feature space, leaving thorough scaling analysis for the individual implementations.

Third CF-related issue, *synonymy* [75], might become a problem in some monitoring applications. This work assumes that sensors, and measurements or feature values acquired by them, have unique identifiers, so exact synonyms would not appear. A more elaborate synonymy would arise in the case where two closely located sensors are measuring the same phenomenon and thus produce nearly identical results. In a novelty detection setting, training data acquired from such sensors might not discriminate two synonymous sensors from a pair of sensors that would later develop some differences in their measurements (test data) as an anomaly. So, the possibility of solving the issue completely by means of unsupervised data analysis seems low: redundancy of the sensors needs to be managed by some other means, like careful positioning of the sensors during deployment.

Recommendation system related issue of *privacy* is not considered to be an issue in this work. Sensor nodes do not have a privacy to protect, if the sensor network (and the monitored subject) are governed by the same administrator. Privacy of the people interacting with the monitored environment is a different issue, not in the scope of this work.

To assess usefulness of CF in a monitoring application, a representative set of monitored events in the test set might be required. For example, it may be pointless to create a global allocation of monitored features (across a structure) and assess the accuracy in detecting a single damage. This corresponds to the problem of *coverage* discussed in [42].

Monitoring applications have also issues not present in common recommendation systems. One significant difference is that recommendation systems assume static items, like editions of books, or movies: they don't change once published. An SHM system potentially deals with dynamically changing environment, where the preferences in different available measurements may change and make older ratings outdated. This application has at least two options to overcome the problem: either resorting to periodical updates in the collaborative filtering (making gradual drift considered normal), or dealing with a change detection problem, where the properties of the reference state are assumed constant until a (large, abnormal) change is detected and the system has fulfilled it's purpose. This work considers the latter case with either constant environmental factors (temperature etc.), or elimination of the variability (transmissibility).

Another significant difference, related to the dynamically changing envi-

ronment, is that recommendation systems eliminate items already rated by the user. For example, there is no point in recommending a news article that was apparently read already. Meanwhile, an SHM system would continue sampling the best rated features.

If collaborative filtering is completely distributed to the sensor network nodes, energy spent in broadcasting CF messages across the network would need to be significantly smaller than simply transmitting all measurements to the sink node. The other option would be transmitting ratings from the individual nodes to a centralized controller (which has ample resources) for performing the CF with centralized methods.

This work considers a centralized approach and the possible energy consumption related to distributed CF is left for future work. Thus, the method is divided into producing ratings for features that are local to sensor nodes (*local ratings*) and, after transmitting the ratings to a centralized computer, the *global assignment* of which features are monitored by which sensors.

3.4.2 Local Ratings

One of the key properties of CF is that (subjective) evaluation of which items are considered interesting or useful is outsourced to the users, as opposed to *content-based* recommenders, where the system has to explicitly model the relationship between content of the items and their preference to the users [75]. However, this work considers a set of wireless devices that arguably don't have unknown subjective preferences to be used for CF. What CF approach offers here is the separation of sensor-wise preference heuristic from the following centralized selection procedure.

The rating heuristic, proposed as an SHM-specific example in Publication IV, operates by voting for few ($D_S << K$) DFT bins that have been observed to have higher magnitude than others, in the $t \in [1, T_{CF}]$ initial time windows used for the selection process. Other bins are assumed to have *default vote* of zero ($r_0^s[k] = 0$, $\forall s, k$). This results in energy efficient transmission of the local ratings, but also sparsity. For sensor $s \in [1, S]$ and DFT sample $k \in [0, K-1]$, the ratings are accumulated over the T_{CF} samples:

$$r_t^s[k] = r_{t-1}^s[k] + \begin{cases} 1 & \text{if } |X_t^s[k]| \in top(D_S, |X_t^s[j]|) \\ 0 & \text{else,} \end{cases} \tag{3.23}$$

where the ranking of DFT bins, $top(D_S, |X_t^s[j]|)$, is performed over sensed

bins j: either for all $j \in [0, K-1]$ or some random subset for the voting round t.

Local ratings could be considered as an attempt to maximize signal-to-noise ratio (SNR), since variability in low magnitude DFT sample disturbs relatively more than the same amount of variability in samples with high magnitude. Low values are problematic especially when considering the denominator $|X_t^{s_2}[k]|$ in transmissibility magnitude (Equation 2.17).

It has been commented to be useless to monitor frequencies where the response is zero [50]. On the other hand, there could be a situation where $|X_t^{s_1}[k]| \approx 0$ for the transmissibility numerator when the structure is healthy, but turns positive due to damage. Then, the above rating would have eliminated the measurements in vain.

3.4.3 Global Assignment of Features

The actual collaborative filtering stage is based on centralized analysis of the accumulated rating matrix:

$$\mathbf{R} = \{R_{s,k} = r^s_{T_{CF}}[k-1]\}. \tag{3.24}$$

There are various methods for implementing CF and they can be categorized into *memory-based*, *model-based*, and *hybrid* solutions [75]. Memory-based CF refers to computing user similarities and item rankings directly from available rating matrix elements, while a model-based approach would use some explicit model (e.g. Bayesian network (BN)) to account for missing data.

The approach in Publication IV can be classified as memory-based and aims at producing user-based top-N recommendations [75], or sensor-based top-D_S recommendations in the current terminology, without the elimination of already rated items.

The implementation used in Publication IV is based on computing *combined ratings*, $\underline{\mathbf{W}} = \{W_{s_1,s_2,k} = w^{s_1,s_2}[k-1]\}$, that are symmetric in terms of sensors and prefer combinations with high DFT magnitude at both sensors:

$$\mathbf{w}_{s1,s2} = corr(\mathbf{r}_{s1}, \mathbf{r}_{s2})\mathbf{r}_{s1} * \mathbf{r}_{s2}. \tag{3.25}$$

Local features are then selected by taking top-D_S bins k for each sensor. Effectively, $corr(\mathbf{r}_{s1}, \mathbf{r}_{s2})$ computes correlation between each pair of sensors over the frequency domain, and is used as a similarity measure between the sensors. This selection process also defines a subset of transmissibility magnitude features and guarantees that the features can be

computed at the centralized computer from the distributed DFT results, due to the symmetry. Naturally, the same parameter space $\mathbb{D}$ in Equation 3.1 applies also here. The final number of monitored transmissibility magnitudes, D_C (input to the subsequent classifier), depends linearly on both the limit D_S and number of sensors S: $D_C \in \mathcal{O}(D_S S)$.

The weakness of memory-based approach arises in extreme sparsity [75]. If the few votes were given completely randomly, it would be likely that any two sensors rate disjoint sets of features, thus preventing reliable similarity estimates between sensors. On the other hand, the local ratings described in previous section concentrate ratings for the parts of vibration spectrum with highest magnitude and vibration at adjacent sensors is likely to be correlated, so some amount of correlation in rating activity ($R_{s,k} > 0$) is expected in practice.

It was also noticed that plain clustering of sensors is not useful, as there can be clusters with lone sensors, isolated from collaboration with the other sensors. Also, crisp clustering as such prevents utilizing transmissibility magnitudes between the groups of sensors. Whether some other clustering scheme is better was left for future work.

4. Novelty Detection

"2. If a message can be interpreted in several ways, it will be interpreted in a manner that maximizes the damage." – Osmo A. Wiio

4.1 Problem Setting

After deciding which features are being monitored and continuing with measuring the selected features, an SHM system is left with the general level problem of detecting differences between a normal feature vector and an abnormal feature vector. In statistical pattern recognition terms, the system needs a binary *classifier* [81, 7]: the two class labels being $c^0 =$ "$healthy$" and $c^1 =$ "$damaged$". A third option, called "$reject$" [7], could be used for indicating that the classifier doesn't have enough separation between the two classes for some measurements, but that is left beyond the scope of this work.

Classifiers are typically based on a given model, whose parameters θ are estimated from initial measurements, a *training set*, denoted here by (temporal) subscripts $t \in [1, T_{tr}]$. Then, the model is used to *decide* appropriate class C_t for a new sample, given the learned model parameters $\hat{\theta}$ and the new feature vector $\mathrm{x}_t \in \mathbb{R}^{D_C}$ (output from the feature extraction and dimensionality reduction stages). From classifier performance assessment point of view, the new samples ($t > T_{tr}$) are said to belong to a *test set*, since the goal is to *generalize* beyond the training set [7].

As stated in previous chapters, no training data are available from a damaged structure, only from the healthy one. Thus, similarly to Requirement 3.1.1, also the classifier needs to consider this:

Requirement 4.1.1. *Extreme bias/imbalance of data: classifier parameters have to be estimated from healthy samples only.*

Additionally, the goal is to support streaming over the coarse time scale,

which brings more requirements, stated here for clarity:

Requirement 4.1.2. *Causality across time: data is handled sequentially and the output of the monitoring system at time t can only depend on data observed so far. Thus, the classifier output is a function of previous data $\hat{C}_t = f(\mathbf{x}_{1:t})$.*

Requirement 4.1.3. *Online monitoring: as time t increases without limit, the system cannot store all of the data received so far, so computational complexity (and size of θ) has to be bounded by constant with respect to t.*

The latter requirement explains the need for summarizing training data with some kind of a model, although the model can include some or all of the (fixed amount of) training data as such. Estimation of model parameters ($\hat{\theta} = g(\mathbf{x}_{1:T_{tr}})$) divides the data set into the two parts mentioned above: training set $\{\mathbf{x}_t | t \leq T_{tr}\}$ and test set $\{\mathbf{x}_t | t > T_{tr}\}$. Then, the classifier can be considered as a function of test measurements and the parameters: $\hat{C}_t = f(\mathbf{x}_{T_{tr}+1:t}, \hat{\theta})$.

Classifiers that meet these requirements, especially 4.1.1, are said to perform *novelty detection* [7, 63, 78], also known as *anomaly detection* [14], *outlier detection* [43, 89], or *one-class classification* [79].

It could be argued that despite the efforts of measuring "clean" training data from a healthy structure, the data set can have *outliers* due to measurement noise alone and that true novelty is different: clean measurements from a different (damaged) state. In general, a classifier cannot make a difference between a faulty observation and an observation of a fault. Thus, this work does not make a difference between an "outlier" and a "damage": erroneous measurements misclassified as damage are considered unavoidable, but possible to analyze in a controlled test environment, as discussed in Chapter 5.

The focus of this work is in "semi-supervised recognition" (Type 3) as defined in [43]: because of the online monitoring requirement, outliers are not identified retrospectively, and lack of data from damages prevents supervised setting. The kind of semi-supervised case, where only part of training data has labels and the rest is from an unknown condition, and the purely unsupervised case where the training set is completely from unknown condition(s), are left beyond the scope of this work. It is considered unlikely, that a statistical pattern recognition system could identify some set of observations as "healthy" without additional information (i.e., unsupervised).

Many novelty detection methods for high-dimensional data combine a dimensionality reduction method and the actual detection method [89], but this work considers the methods as separate steps in data processing. Then, dimensionality reduction is possible to be customized for SHM data, while the novelty detection phase remains universal.

Yet another focus in the problem setting is made by assuming IID features for each time step t (in the healthy state). This is called *static* novelty detection [41, 78]: computing the discriminant function, $f(\mathbf{x}_t, \hat{\theta})$, for each new measurement separately. This is different from novelty detection with *dynamic* models [41, 78], such as dynamic Bayesian network (DBN) models, where the model summarizes past test measurements. Also, the approach is different from *change detection* [4], which assumes sudden and permanent change and is based on accumulating evidence of the change. In the scope of this work, it is required to detect also slow drift, so models that would somehow adapt to drift in test data are out of the question.

Large amount of different static and semi-supervised novelty detection methods have been published [79, 63, 43, 14]. The research questions here are simply to explore the following aspects:

Problem 4.1. *Which (existing) novelty detection model would be relevant in the context of feature spaces described above?*

Problem 4.2. *Detection methods are known to suffer from high-dimensional data (in terms of accuracy and execution time), but how D_C affects each combination of dimensionality reduction and novelty detection methods?*

The following sections explain each of the detection models used in this work: one variant of distance-based detection and three common variants of density models.

4.2 Nearest Neighbor Model

Maybe the most straightforward classifier is the nearest neighbor (NN) rule [81]: compute distances between the new feature vector and training data vectors, and select the class label associated with the closest training vector. The NN approach can be described as a *nonparametric* and *distance-based* method, that doesn't assume a specific form for the class distributions. On the other hand, the used distance measure is assumed to be suitable for indicating difference between feature vectors, which ob-

viously depends on the scaling and type of features received from the previous processing phase.

Novelty detection setting is missing training data from the damaged state, so all the nearest training vectors will have $C = c^0$. Thus, the NN classifier has to be modified for novelty detection. There are several different modified variants [79, 80] which belong roughly into two approaches: compute distance to (kth) nearest neighbor, and estimate local density based on the (kth) nearest neighbor distances.

The simplest ($k = 1$) nearest neighbor novelty detector just computes the minimum distance to training data:

$$f_{NN}(\mathbf{x}_t, \mathbf{x}_{1:T_{tr}}) = min_i\{\|\mathbf{x}_t - \mathbf{x}_i\| | i \in [1, T_{tr}]\}, \tag{4.1}$$

and subsequently uses some suitable threshold to decide which test vectors $\mathbf{x}_t$ are too distant to be normal. The training vectors could be seen as the set of parameters of the model, so $\hat{\theta}$ contains $D_C \times T_{tr}$ values, although some of the training vectors are likely to lie close to others in the middle of the set and are potentially redundant (in low dimensional feature space).

The other approach is to include (local) density estimation into the distance computations [79]: comparing the distance to (kth) nearest neighbor and the distance between the nearest neighbor and its nearest neighbor. For $k = 1$, this could be expressed as:

$$\begin{aligned} i &= argmin_i\{\|\mathbf{x}_t - \mathbf{x}_i\| | i \in [1, T_{tr}]\}, && (4.2)\\ j &= argmin_j\{\|\mathbf{x}_i - \mathbf{x}_j\| | i \neq j \in [1, T_{tr}]\}, && (4.3)\\ f_{NNd}(\mathbf{x}_t, \mathbf{x}_{1:T_{tr}}) &= \frac{\|\mathbf{x}_t - \mathbf{x}_i\|}{\|\mathbf{x}_i - \mathbf{x}_j\|}, && (4.4) \end{aligned}$$

where the nearest training vector $\mathbf{x}_i$ is identified first and then compared to its (constant) neighbor $\mathbf{x}_j$. This is similar to the even more complex method called local outlier factor (LOF) in [12], but without "smoothing" caused by the use of *reachability distance* and averaging over $MinPts$-neighborhood in *local reachability density*.

As explained about the research process of this work, Publication II and subsequent papers in this work concentrated on the simpler models, leaving LOF models for future assessment.

4.3 Gaussian Density Estimate

On one hand, nonparametric models make few assumptions about the shape of the proper decision region, but on the other hand, simple para-

metric density models offer fewer number of model parameters to fit in exchange for the modeling assumptions.

Gaussian or normal density function is a popular and simple parametric density model [81]. One of the common motivations for using it is its analytical tractability and *central limit theorem* [7]: if the variability in data is caused by a sum of many IID random variables, the distribution of the sum is approximately Gaussian.

The probability density function (PDF) for normally distributed vector, $\mathbf{x}_t \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, is

$$p(\mathbf{x}_t|\mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^{D_C}|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}_t-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_t-\mu)}, \tag{4.5}$$

where the parameters are mean $\mu \in \mathbb{R}^{D_C}$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{D_C \times D_C}$. The parameters can be estimated from sufficient amount of training data in terms of maximum likelihood (ML):

$$\hat{\mu} = \frac{1}{T_{tr}} \sum_{t=1}^{T_{tr}} \mathbf{x}_t, \tag{4.6}$$

for mean and

$$\hat{\mathbf{\Sigma}} = \frac{1}{T_{tr}-1} \sum_{t=1}^{T_{tr}} (\mathbf{x}_t - \hat{\mu})(\mathbf{x}_t - \hat{\mu})^T, \tag{4.7}$$

for unbiased covariance. There exist also sequential methods for estimating the parameters online [7], but they are beyond the current scope.

From classification point of view, the normalization constant and monotonic exponential function of Equation 4.5 can be ignored, since they don't change the shape of the final decision region (although, change the appropriate decision threshold). Once the parameters are estimated, and the inverse $\hat{\mathbf{\Sigma}}^{-1}$ computed, classification can be done in terms of squared *Mahalanobis distance* [81, 79]:

$$f_G(\mathbf{x}_t, \hat{\mu}, \hat{\mathbf{\Sigma}}) = (\mathbf{x}_t - \hat{\mu})^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_t - \hat{\mu}), \tag{4.8}$$

where values above a given threshold are considered abnormal.

Compared to the NN method, this offers compensation of scale differences and covariance between individual features (whitening), but on the other hand, distances are computed with respect to the global mean, not the closest training samples. If used together with PCA, data from healthy structure is already axis-aligned and a diagonal covariance matrix is enough, so this detection method overlaps with dimensionality reduction performed with PCA.

4.4 Mixture of Gaussians Density Estimate

In case the variation in the (healthy state) data has some more complex structure than the above unimodal multivariate Gaussian, a more complex density model may be required. One approach in increasing the flexibility of the model is to consider *mixture models* [79, 7]. A natural step from the basic Gaussian is a mixture of Gaussians (MoG) model, which assumes the distribution is a weighted sum of multivariate normal components:

$$p(\mathbf{x}_t|\mathbf{w},\mathbf{M},\underline{\mathbf{\Sigma}}) = \sum_{i=1}^{I} \frac{w_i}{\sqrt{(2\pi)^{D_C}|\mathbf{\Sigma}_i|}} e^{-\frac{1}{2}(\mathbf{x}_t-\mu_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}_t-\mu_i)}, \tag{4.9}$$

where the parameters consist of the normalized weight vector $\mathbf{w} \in \mathbb{R}^I$, $\sum_i w_i = 1$, mean vectors $\mathbf{M} = [\mu_1,\ldots,\mu_I] \in \mathbb{R}^{D_C\times I}$, and corresponding covariance matrices $\underline{\mathbf{\Sigma}} \in \mathbb{R}^{D_C\times D_C\times I}$.

The parameter estimation problem does not have a closed-form analytical solution due to the included summation over a latent variable (component identity), but the parameters can be estimated in terms of ML with the expectation maximization (EM) algorithm [21, 7]. In short, EM starts from a random (but valid) guess for the parameters, updates them iteratively while alternately estimating component posteriors and component parameters, and converges to a locally optimum solution.

Given the parameters $\hat{\theta} = \{\hat{\mathbf{w}},\hat{\mathbf{M}},\hat{\underline{\mathbf{\Sigma}}}\}$ estimated from training data, classification is performed in terms of the density function:

$$f_{MoG}(\mathbf{x}_t,\hat{\mathbf{w}},\hat{\mathbf{M}},\hat{\underline{\mathbf{\Sigma}}}) = p(\mathbf{x}_t|\hat{\mathbf{w}},\hat{\mathbf{M}},\hat{\underline{\mathbf{\Sigma}}}), \tag{4.10}$$

and the area(s) of feature space with density values above a given threshold is classified as normal ($\hat{C}_t = c^0$).

In principle, the classification could be based on *posterior probability* [7] computed according to the Bayes rule and class prior probability: $p(C_t = c^0|\mathbf{x}_t\hat{\theta}) \propto p(\mathbf{x}_t|C_t = c^0,\hat{\theta})p(C_t = c^0)$, but the prior probability is considered as an unknown parameter equivalent with choosing another decision threshold. Also, prior distributions for the model parameters θ could be used, especially when only a short sequence of training data is available, but that is beyond the experiments of this work.

4.5 Parzen Density Estimates

A kind of hybrid approach between the NN and MoG models, or an extreme version of MoG, is called *Parzen* or *kernel density estimation* [68, 79, 7]. Instead of fitting a sum of few parametric densities to the training vectors, the idea is to sum a kernel function (Parzen window) positioned at each training vector. One common choice for the kernel is symmetric Gaussian [7], which results in the following density function:

$$p(\mathbf{x}_t|\mathbf{x}_{1:T_{tr}}, h) = \frac{1}{(\sqrt{2\pi}h)^{D_C} T_{tr}} \sum_{i=1}^{T_{tr}} e^{-\frac{1}{2h^2}(\mathbf{x}_t-\mathbf{x}_i)^T(\mathbf{x}_t-\mathbf{x}_i)}, \tag{4.11}$$

where the single parameter h (in addition to training data) governs the window width. This model has also been used in [78] for novelty detection in jet engine vibration data.

An appropriate value for h is estimated by using a gradient ascent algorithm [7] to find ML solution for the training set, as described in [79, 26]. Trivial estimate ($h = 0$) is avoided with a leave-one-out scheme.

Again, classification in performed in terms of the density estimate:

$$f_P(\mathbf{x}_t, \mathbf{x}_{1:T_{tr}}, \hat{h}) = p(\mathbf{x}_t|\mathbf{x}_{1:T_{tr}}, \hat{h}), \tag{4.12}$$

and areas of feature space with sufficient density are considered normal.

4.6 Decision Thresholds

The above descriptions of novelty detection methods output a single continuous value, $f(\mathbf{x}_t, \hat{\theta}) \in \mathbb{R}$, which measures either normality or outlierness of $\mathbf{x}_t$. The actual decisions are made by setting a constant threshold ϕ which divides the feature space into two *decision regions* [7].

Suitable threshold value ϕ ultimately depends on various *decision theoretic* factors [73, 7], like expected misclassification cost, expected loss, or expected utility of related consequences in general. A cost or utility function depends on what the user of the SHM system is trying to achieve. Large part of this work avoids setting a particular threshold value, but is based on receiver operating characteristic (ROC) analysis instead, as described in the following chapter.

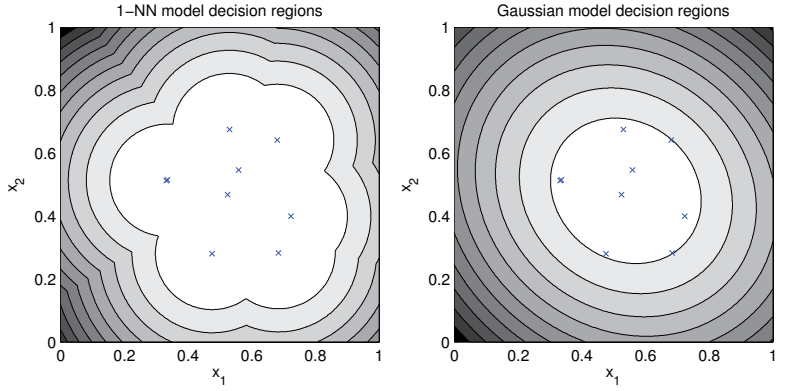
NN distance f_{NN} and Mahalanobis distance f_G measure normality so that smaller values correspond to normal data, and densities f_{MoG} and f_P indicate normality with high values. Thus, the decision rules for indicating healthy structure are:

Distances: $f_{NN}(\mathbf{x}_t, \hat{\theta}) < \phi$, and $f_G(\mathbf{x}_t, \hat{\theta}) < \phi$,

Densities: $f_{MoG}(\mathbf{x}_t, \hat{\theta}) > \phi$, and $f_P(\mathbf{x}_t, \hat{\theta}) > \phi$.

Example 4.6.1. *A small toy data set of $T_{tr} = 9$ bivariate vectors was sampled and each of the four detector models were fitted to visualize the shape of the corresponding decision boundaries. The data is from a uniform distribution, so the point is not to establish the "correctness" of the learned model: just to visualize differences between the models.*

Figure 4.1 shows each of the resulting decision regions: the training data is marked with blue crosses (the two leftmost overlap) and shades of gray indicate decision regions with various decision threshold values. Areas of the regions appear different between the detectors, since the thresholds are not normalized.



(a) Nearest neighbor **(b)** Gaussian density

(c) Mixture of Gaussians **(d)** Parzen with spherical kernel

Figure 4.1. Examples of decision boundaries for each detector model.

5. Performance Assessment

"One man's system is another man's subsystem."

5.1 Problem Setting

Each of the data processing methods described in previous chapters have been originally developed for certain purposes, make certain modeling assumptions, and provide optimal solutions with respect to particular criteria. For example, transmissibility magnitude features cover insensitivity towards changes in excitation amplitude, PCA ignores feature components with small variance (in training data), but doesn't eliminate measurements per each sensor, and NN detector assumes that Euclidean distances in the (reduced) feature space are appropriate for assessing abnormality. At the final decision step, the detection threshold can be varied and the resulting system can have different sensitivity to damages.

The approach, chosen already in the beginning of this work, is to assess a proposed SHM system as a whole. Similar idea has been used for wireless SHM systems before in [36] and further in [35]. While it may prove difficult to accurately determine appropriate boundaries in what is included in an SHM system and what is considered external to it, this work concentrates on monitoring systems that include a set of wireless accelerometers as the source of data and binary novelty detection results as the output. Thus, the potential assessment criteria are related to the amount of measurements made by the wireless sensors and counting confusions at the detector output.

This work relies on empirical assessment of the proposed combinations of methods, according to criteria that are possible to compute from the given material. The first assessment criterion is *detection accuracy*: in the presence of noise and uncertainty, the system is bound to produce

false alarms and miss some damages. Another assessment criterion, formulated later in the research, is the (simplified) *energy consumption* of wireless sensors, while measuring a given number of local features.

Notably, this work does not consider misclassification costs or risk minimization from decision theory point of view [7], or realized WSN lifetime [23], as the analysis would require more information about a specific deployment. In order to set a particular decision threshold and reach a final decision under uncertainty, one would need to consider misclassification costs or their ratio [7, 38], but such parameters are unavailable in the scope of this work. Likewise, to analyze how long a WSN operates for a given amount of battery capacity [23], one would need an actual wireless network with certain topology and realistic communication conditions, but that kind of experimentation would exceed the resources available for this work.

Thus, the problems in this work are in finding general enough assessment criteria:

Problem 5.1. *The detection accuracy of a proposed wireless damage detection system needs to be assessed without fixing particular misclassification costs.*

Problem 5.2. *The energy consumption of wireless sensors, or cost of acquiring data in general, needs to be assessed without knowledge of the particular communication network parameters.*

5.2 Detection Accuracy

Given experimental test data from both the healthy structure and damage cases (with true labels $C_t \in \{c^0, c^1\}$), and a decision threshold, the detector output $\hat{C}_t$ can belong to one of four combinations:

true negative when $\hat{C}_t = C_t = c^0$,

true positive when $\hat{C}_t = C_t = c^1$,

false positive when $\hat{C}_t = c^1$, but $C_t = c^0$,

false negative when $\hat{C}_t = c^0$, but $C_t = c^1$,

where the latter two are also known as type I and type II errors in statistical hypothesis testing [81].

The number of occurrences of each of the cases in test data ($T_{tr} < t \leq T$) can be counted and are denoted here with TN, TP, FP, and FN, respectively. Naturally, the total number of samples is $TN + TP + FP + FN = T - T_{tr}$, $TN + FP$ equals the number of healthy samples, and $TP + FN$ equals the number of damaged samples in the test data.

From the user point of view, it may be more interesting to know how likely a monitoring system is to produce true positive or false positive detections. If the detections are assumed IID, the probability of detection P_D, also known as statistical power or *sensitivity*, can be estimated from test data as true positive rate (TPR):

$$P_D \approx TPR \equiv \frac{TP}{TP + FN}, \tag{5.1}$$

and similarly, the probability of false alarm P_F (or 1 - *specificity*) as false positive rate (FPR):

$$P_F \approx FPR \equiv \frac{FP}{TN + FP}. \tag{5.2}$$

If the detections were not IID, as with dynamic models or change detection, the assessment criteria would need to be different, as noted in Publication V. Then the monitoring task is not to determine *if* there is a damage, but *when* it appeared.

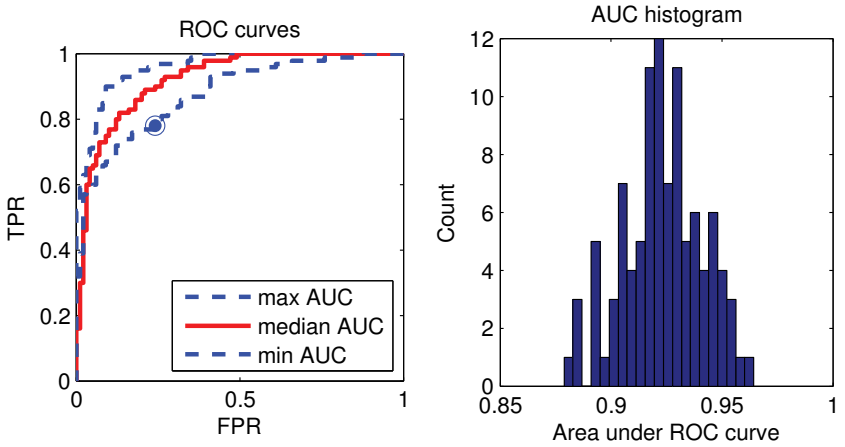
In the case of overlapping class distributions, such as noisy transmissibility measurements and small damages in the structure, there is a trade-off between sensitivity and specificity: a sensitive detector will have false alarms more often, and conversely, a detector that avoids false alarms will also be insensitive to damages. The situation can be analyzed in terms of ROC curves and area under (ROC) curve (AUC) [76, 11]. An ROC curve is a parametric curve of TPR versus FPR, while the decision threshold is varied over possible values. A random classifier, without any discrimination between the classes, produces (approximately) a diagonal curve ($TPR = FPR$), and the ideal classifier, with perfect discrimination, produces a step curve ($TPR = 1, \forall FPR$). Thus, ROC measures *class discrimination capability* [81] and, in this work, provides a summary of all the data processing stages prior to the decision step. If the feature extraction produces pure noise, no amount of subsequent processing will make the classes separate at the decision step.

An ROC curve can be summarized by integrating the area under the curve, AUC [11], which then provides a value that is near $AUC \approx 0.5$ for random or otherwise poor detection systems and $AUC = 1.0$ for ideal detection. If the system made consistently wrong classifications, the area

would be $0 \leq AUC < 0.5$, but then the results could simply be inverted. The weakness of the approach is that it averages the result over all possible FPR values, while the relevant part of ROC curve may be, for example, with threshold values that result in detectors with high specificity (low FPR).

There are also practical issues with estimating ROC curves, like limited number of test samples, which in turn results in few points for the parametric curve. This work assumes *data imbalance* [15] concerns only training data. However, there is an issue of *coverage* [23]: whether an SHM system is sensitive to damages in *all* parts of the structure, while the experimental data might have only few damage locations. This kind of coverage and localization considerations are left for future work.

Example 5.2.1. *ROC curves and AUC values of toy detectors: values from two overlapping univariate distributions,* $\mathcal{N}(0; 0.5)$ *and* $\mathcal{N}(1; 0.5)$*,* $T = 100$ *samples from each, were drawn to simulate detector output and compute an ROC curve. The experiment was iterated* $I = 100$ *times and the ROC curves were sorted in terms of their AUC values. Figure 5.1 visualizes the distribution of the* I *ROC curves in the experiments: the actual curves with minimum, median, and maximum AUC values are shown in Figure 5.1a, and a histogram of the AUC values is shown in Figure 5.1b.*



(a) ROC curves **(b)** AUC values

Figure 5.1. Examples of ROC curves and AUC values.

5.3 Cost of Acquiring Data

Traditional data mining is posed with a situation where large amount of data has already been collected for a specific primary purpose and the task is to find something new in it [37]. There, part of the costs of computation are related to accessing the existing database. As noted already in Chapter 2, this work aims at WSN monitoring applications with *data streams* [32], which sets various hard and soft limits on accessing the data from wireless sensors.

At a very general level (e.g., from *value of information* [45, 73] point of view), investing enough resources into the measurement hardware can make arbitrary amount of data available for monitoring. Ultimately, the accelerometers could have dedicated high-bandwidth media (cable) for communication and practically unlimited (wired) power supplies, *if* the gained insight about structural health was considered worth it. Continuing cost considerations at this general level leads to the option of limiting the cost of measurement hardware and considering WSN technology.

As a part of such larger context of deciding what is worth measuring and how, this work limits to considering techniques suitable for wireless monitoring of accelerometer spectra, as detailed in Chapter 2. The use of Goertzel algorithm as a feature extraction stage is dictated by a hard limitation of WSN nodes: features transmitted from the sensors need to be computed online with a small amount of memory. The limitations are caused by the CPU and RAM available on-board the wireless sensors and future technology may set different limits.

One of the remaining free parameters in the local feature extraction, and the resulting energy consumption of wireless sensors, is the number of sensed and transmitted DFT samples per sensor node (denoted with D_S, for dimensionality of sensor data). This leads to an experimental setting used in Publication V, where D_S is varied over a range of reasonable values, and the resulting detection accuracy of the system is analyzed to find potential trade-offs. After all, trade-offs between accuracy and energy consumption is a traditional WSN topic of its own [10].

It has to be noted, that considering D_S as "energy consumption" is a simplification. There are additional factors, like the CPU energy consumption during the local computation and the dynamics of the wireless communication network, that make the ultimate WSN lifetime more complex to model [23, 30].

6. Experiments and Results

6.1 Data

This work relies on empirical observations about the performance of proposed combinations of data processing stages. The evaluations were performed with three kinds of accelerometer data sets, gradually exposing the methods to new data. In chronological order, the data sets were from:

- a wooden model bridge, introduced in [55, 56, 58],
- Los Alamos National Laboratory (LANL) bookshelf, a metal structure introduced in [2, 29, 60], and
- numerical simulations performed by one of the coauthors.

The model bridge was a 4.2-meter, 36-kg wooden truss structure monitored in approximately constant laboratory conditions [55], shown in Publication III, Figure 2. Excitation x^i was pseudorandom noise input to the structure by an electrodynamic shaker and was not directly measured. Fifteen wired accelerometers were used for measuring the (output) vibration of the structure ($S = 15$), with sampling frequency of $f_S = 256$ Hz and time window of 32 s ($N = 8192$). In total, there were $T = 2509$ measurement periods, while some of the measurements had small added masses placed at various locations on the structure to simulate damage. Table 6.1 summarizes the data set: the class label $C = 0$ indicates healthy structure, while the other values are various damages. Publication I, Fig. 6 visualizes the labels as a bitmap image. This data set was used in all of the publications of this work, although Publication V uses only a subset

of the data (only one damage location, $C \in [5, 9]$) and a different indexing for time t.

Table 6.1. Wooden model bridge data

Label	Added mass (g)	Measurements t	Δt
$C = 0$	0	1–1515	1515
$C = 1$	23.5	1516–1549	34
$C = 2$	47.0	1550–1577	28
$C = 3$	70.5	1578–1599	22
$C = 4$	193.5	1600–1623	24
$C = 0$	0	1624–2080	457
$C = 5$	23.5	2081–2100	20
$C = 6$	47.0	2101–2123	23
$C = 7$	70.5	2124–2145	22
$C = 8$	123.2	2146–2165	20
$C = 9$	193.5	2166–2185	20
$C = 0$	0	2186–2208	23
$C = 10$	23.5	2209–2220	12
$C = 11$	47.0	2221–2232	12
$C = 12$	70.5	2233–2244	12
$C = 13$	123.2	2245–2252	8
$C = 14$	193.5	2253–2260	8
$C = 0$	0	2261–2357	97
$C = 15$	23.5	2358–2377	20
$C = 16$	47.0	2378–2400	23
$C = 17$	70.5	2401–2426	26
$C = 18$	123.2	2427–2452	26
$C = 19$	193.5	2453–2475	23
$C = 0$	0	2476–2509	34

Likewise, the LANL bookshelf data [60] was measured from a model

structure in laboratory conditions and excitation was input by a shaker. The structure consisted of three aluminum floor plates supported by four columns, while the damages were simulated by loosening and removing bolts from the joints at two locations in the structure (1st floor column C, and 3rd floor column A). Accelerometer sensors ($S = 24$) were attached to the edges of each floor, as shown in Publication V, Figure 3, and were sampled at $f_S = 1600$ Hz for 5.12 s at a time ($N = 8192$) for total of $T = 270$ measurements. The data set is summarized in Table 6.2. Apparently, another data set from a similar structure has been published [31], and they used fewer accelerometers and a different damage model, but that data set was not available at the time when the current data set was used in Publication III, Publication IV, and Publication V.

Table 6.2. Bookshelf data

Label	Damage	Measurements t	Δt
$C = 0$	none	1–150	150
$C = 1$	1C loosened to 0.6 Nm	151–160	10
$C = 2$	1C loosened to 1.1 Nm	161–170	10
$C = 3$	1C loosened to hand tight	171–180	10
$C = 4$	1C bolts removed	181–195	15
$C = 5$	1C bracket removed	196–210	15
$C = 6$	3A bolts removed	211–225	15
$C = 7$	3A bracket removed	226–240	15
$C = 8$	1C and 3A bolts removed	241–255	15
$C = 9$	1C and 3A bracket removed	256–270	15

A third data set was sampled from a numerical simulation of a 1.4-meter *cantilever beam* (fixed at one end, while the other end is free). The simulation included random excitation, with random load distribution (no fixed location of x^i), and measurement noise. A progressively increasing damage was modeled by decreasing stiffness of the beam at one location (between "sensors" s_3 and s_4). Transverse acceleration was sampled from $S = 20$ equidistant points along the structure at $f_S = 4000$ Hz, but the spectrum was used only up to 256 Hz in the analysis. Measurement periods lasted 20 seconds ($N = 80000$) and there were $T = 100$ measurements.

Table 6.3 summarizes the data set. It was used in [70] and the subsequent Publication III.

Table 6.3. Data from simulated cantilever beam

Label	Damage level	Measurements t	Δt
$C = 0$	0.00	1–50	50
$C = 1$	0.11	51–60	10
$C = 2$	0.27	61–70	10
$C = 3$	0.43	71–80	10
$C = 4$	0.59	81–90	10
$C = 5$	0.75	91–100	10

Likewise, a fourth data set was sampled from a numerical simulation, which modeled a 1.4-meter long 50 mm $\times$ 5 mm *simply supported beam* (fixed from both ends) additionally supported by a spring at 612.5 mm from the other end. Simulation included random excitation at various amplitudes at the midpoint and a constant amount of measurement noise. Damages were modeled as a decrease in the beam depth at the spring support: thickness varied from the original 5 mm down to 2.5 mm. Acceleration was monitored at $S = 47$ equidistant points along the beam at $f_S = 500$ Hz for 5 s at a time ($N = 2500$) for a total of $T = 300$ measurements. The data set is summarized in Table 6.4. This data set was used in Publication V.

Table 6.4. Data from a simulated simply supported beam and a spring

Label	Remaining thickness (mm)	Measurements t	Δt
$C = 0$	5.0	1–150	150
$C = 1$	4.5	151–180	30
$C = 2$	4.0	181–210	30
$C = 3$	3.5	211–240	30
$C = 4$	3.0	241–270	30
$C = 5$	2.5	271–300	30

6.2 Feature Extraction

Publication I proposed using the following data processing stages for wireless damage detection:

- online frequency domain feature extraction on wireless accelerometers with the Goertzel algorithm or QAM,
- centralized transmissibility magnitude estimation,
- random selection of transmissibility features to reduce classifier input space dimensionality,
- naive Bayes model for statistical classification (i.e., Gaussian density estimation model with a diagonal covariance matrix), and
- use of ROC curves and AUC values as the evaluation criteria.

The emphasis of Publication I was on validating the feature extraction stage, while also considering what the eventual output of the detection system should be. This relies on the assumption that features relevant for a supervised classifier will be relevant also for unsupervised monitoring with true novelty detection methods. The Problems 2.1 and 2.2 were addressed by using the two online algorithms described in Chapter 2 to extract features separately for each sensor. Instead of actually implementing it on wireless sensors, this work relies on computations performed offline on the wired wooden bridge measurements. The environmental variability issue (Problem 2.3) was addressed by using centrally computed transmissibility magnitude features as described in Chapter 2.

From dimensionality reduction point of view, Publication I explored what happens to the detection accuracy if small random subsets of transmissibility magnitude features are measured: an attempt to answer Problem 3.1. From the complete space of all $D_F = 6300$ possible transmissibility magnitude features, random subsets of size D_C were selected and used as input to naive Bayes classifier. The feature space dimensionality was varied, $D_C \in \{8, 16, 32, 64, 128, 256, 512, 1024\}$, and a supervised classification experiment was performed for each dimensionality, so that half of the measurements (from both classes) were used for training and the

other half as the corresponding test set. Each of the random feature selection and classification experiments was iterated $I = 1000$ times to observe variability in the associated ROC curves and AUC values. The use of ROC curves and AUC addressed Problem 5.1.

To summarize the results in Publication I:

- the considered feature extraction methods worked in the sense that the resulting AUC values were significantly better than for random classifiers ($AUC \in [0.7, 0.85] > 0.5$), shown in Figure 6.1,

- the dependency of AUC on the number of measured features was explored and characterized: as D_C grows, the random feature subsets enable more accurate classifiers and with less variance in the accuracy (Publication I, Fig. 5), and

- already at $D_C \in \{16, 32\}$, the classifications were relatively accurate *with a certain feature subset*, so there is potential for feature space sparsity in the wooden bridge data set (Problem 3.1).

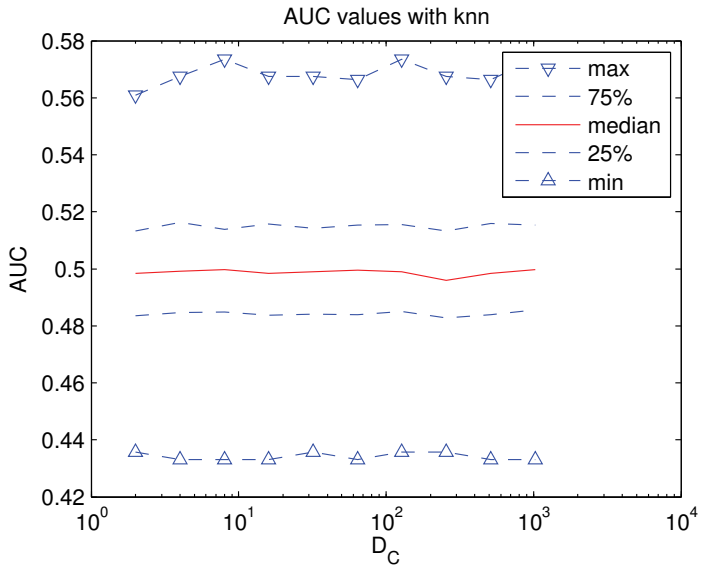


Figure 6.1. AUC values for corresponding experiments with random data.

6.3 Projections and Novelty Detection

Publication II continued the development by replacing random selection and the naive Bayes classifier with:

- random projections, PCA, or CCA as the (unsupervised) dimensionality reduction stage,

- nearest neighbor method, Gaussian, MoG, or Parzen density estimates for novelty detection,

while the feature extraction relied on the distributed Goertzel algorithm and centralized computation of transmissibility magnitudes, and AUC for evaluation, as in the previous publication. Thus, the emphasis of Publication II was on studying the dimensionality reduction stage (addressing Problem 3.2), while also exploring a true novelty detection setting and a set of various detection methods (Problem 4.1).

The experiments were again based on the wooden model bridge measurements and the classifier feature space dimensionality D_C was varied to study its effect on classification accuracy (Problem 4.2):

$$D_C \in \{2, 4, 8, 16, 32\} \cup \{3, 5, 10, 20, 50, 100, 200, 500\}. \tag{6.1}$$

An additional restriction in the feature space was used: only transmissibilities between physically adjacent and parallel sensors were considered[1], which reduced the number of sensor pairs to 19 and the total number of potential features to $D_F = 4541$.

This time, each of the projection and classification experiments had random subsets of healthy state data ($C = 0$, $T_{tr} \approx 0.75T$ samples) as the training data for determining the projection and detector parameters. Each learned projection-classifier combination was then used in detecting anomalies in the rest of the data ($0.25T$ samples, with and without damages). Separation into training set and test set, and the subsequent classifier experiments, were iterated for $I = 10$ times to observe variability of the detection accuracy.

The experiments resulted in sets of points that describe the AUC values for each detection system, where parameters are combinations of a

[1]This restriction can be considered as additional source of geometric information about the structure.

projection method, projection dimensionality, and a detection method:

$$AUC = f(projection, D_C, detector). \tag{6.2}$$

The results were presented as curves in AUC vs. D_C plane for each detection method in Publication II, Fig. 3.-Fig. 6.. Also results with unprojected data are presented for reference in NN and Gaussian detector experiments, while MoG and Parzen density estimation failed in high dimensional feature spaces ($D_C > 100$). Dimensionality of around $D_C = 50$ was found to be the "sufficient output dimensionality" for RP of the data set when using the NN detector. On the other hand, PCA and CCA enabled more accurate classification in the lower dimensional projections: particularly PCA-NN combination at $D_C \in [16, 20]$ and CCA consistently across $D_C > 3$.

6.4 Three-way Analysis

Publication III replaced the dimensionality reduction stage with three-way analysis described in Section 3.3: considering time, vibration frequency, and sensor pair (location) as separate factors, and outputting the R-dimensional temporal factors as the remaining explanation for anomalies in test data. Thus, the emphasis is on experimenting whether three-way analysis is suitable for dimensionality reduction in damage detection (Problem 3.2). Additionally, a feature selection approach was studied by eliminating test set measurements that have PARAFAC loadings closest to zero (a step towards addressing Problem 3.3).

In total, the experiments were made with three data sets: the wooden model bridge measurements (with a setting comparable to the one in Publication II), LANL bookshelf data, and the simulated cantilever beam. Instead of random feature selection or projection space dimensionality, the varied parameters in feature extraction and dimensionality reduction were the length of averaging window in estimating transmissibility magnitude (related to N), frequency resolution (K), number of PARAFAC components (R), and three different flavors of the model (regular, loadings constrained to non-negative values, or pre-centered PARAFAC).

At first, the effect of window length N and DFT frequency resolution K in feature extraction was studied. In the simplest case, N time domain samples allow the computation of $K = N/2$ (independent) DFT magnitude samples, but combined with averaging the results over several separate

time windows brings the option of lowering the rate of data. Increasing N (and lowering T) allows more frequency resolution (higher K and D_F) and/or averaged transmissibility features. Publication III, Figure 3 shows the effects of averaging and frequency resolution in the feature extraction on the final AUC values of the four different detection systems on the three different data sets. Generally, averaging transmissibility magnitude estimates helps in producing more accurate detections, while also frequency resolution affects the results.

The number of PARAFAC components R was determined by using CORCONDIA and compared with AUC values achieved with models of different size. Publication III, Figure 4 shows the agreement between the criteria: values of $R \in [2, 4]$ were found appropriate for the data sets and were used in the experiments.

The effect of PARAFAC models and detection methods was studied by running experiments on combinations of:

- appropriate features extracted from the three different data sets,
- the three different PARAFAC models of appropriate size, and
- the four different novelty detectors, like in Publication II.

The achieved AUC values are summarized in Publication III, Table 1. It shows an improvement over the ones reported for the wooden bridge data in Publication II and comparable values also for the other data sets.

Also a feature selection approach based on PARAFAC loadings was proposed. The idea is to fit a PARAFAC model to the training data and then perform continuous (test set) monitoring only for a reduced set of features that have the highest loadings. A brief experiment was performed with the standard PARAFAC model, so that for each component r:

- the sensor pair s with maximum value in the spatial factor $\mathbf{c}_r$, and
- five frequency bins k with highest value in the spectral factor $\mathbf{b}_r$

were selected and used for estimating novelty detector parameters, and were the only features used for the test set. This corresponds to reducing the number of DFT measurements performed in the supposed WSN during monitoring phase. Results with AUC values comparable to the ones with full data were achieved and reported in Publication III, Table 2.

6.5 Coordinated Monitoring

In parallel with Publication III, Publication IV takes the development in Publication II further by considering local feature dimensionality (D_S) for each sensor. Instead of just centralized feature space dimensionality reduction from a D_F-dimensional space to D_C-dimensional one, after observing all D_F attributes, the monitoring was proposed to be coordinated so that most measurements are avoided on a per sensor basis (addressing Problem 3.3).

Compared to Publication II, the projection stage was replaced with a collaborative filtering approach (Section 3.4). The considered novelty detection methods were limited to using NN and Gaussian density models, mainly because of the relative simplicity and robustness observed in Publication II. In addition to the wooden model bridge data, experiments were performed with the LANL bookshelf data set.

Instead of thresholding PARAFAC loadings as above, the CF approach assumes application dependent rating scheme as discussed in Section 3.4.2, including default votes. The ratings were also based on a short set ($T_{CF} = 10$) of initial measurements, so that the majority of training data set was assumed to be measured with the reduced feature set. In the case of wooden model bridge data, the training sets for detector parameter estimation were $T_{tr} - T_{CF} = 1752$ samples, and $T_{tr} - T_{CF} = 90$ samples for the bookshelf data. Publication IV, Table 1 shows also other experiment parameters, such as the corresponding test set compositions of 105 healthy samples and 105 damaged samples from the wooden bridge, and $50 + 50$ samples from the bookshelf data.

The experiments, including:

- training, and test set selection,
- CF stage for feature selection,
- detector model parameter estimation with the reduced feature set,
- detections with the detector model for the test set, and
- computation of AUC values for the detector output,

were repeated for $I = 10$ times, for each value of local feature dimensionality $D_S \in \{2, 4, 6, \ldots, 20\}$. These values of D_S map to final classifier feature spaces of $D_C \in \{30, \ldots, 300\}$ in 15-sensor bridge data and $D_C \in \{48, \ldots, 480\}$ in 24-sensor bookshelf data correspondingly, when us-

ing the CF approach. The resulting AUC values versus D_C are shown in Publication IV, Figure 8 for the bridge data and Publication IV, Figure 9 for the bookshelf data.

For reference, a modified random selection approach was used: transmissibility magnitude features were selected randomly as a long as the number of frequency bins monitored by any of the sensors remained at most D_S. The resulting classifier feature spaces had half the dimensionality compared to the CF approach, so the corresponding curves in Publication IV, Figures 8 and 9 are shorter. Also a simple selection approach based on majority vote was used ($D_S \in \{2, 4\}$), but the resulting feature spaces remained high-dimensional as voting reduces only the number of frequency bins, not the combinations of sensor pairs. The CF method seems preferable especially with small D_s and the wooden bridge data, where $AUC \approx 0.8$. The relative differences between CF and the reference methods are smaller on the bookshelf data, which also seems to have a better class discrimination overall, since $AUC \approx 0.9$ or better for all of the methods.

6.6 Evaluation Criteria and Trade-offs

Publication V focuses on the final, experimental step of the data processing stages: assessment of wireless damage detection systems. The original purpose was to just benchmark existing methods in a "fair" setting, but there were problems in selecting fair evaluation criteria. The experiments in this work started by considering ROC curves and AUC values against the total number of extracted features input to a classifier (D_C). Then the focus was shifted to how many features a single sensor is required to measure (D_S). The assumptions and ramifications of these evaluation criteria were studied, finally proposing evaluation in terms of

- probability of detection P_D,
- probability of false alarms P_F, and
- number of transmitted features per sensor D_S,

where the first two are related via a chosen decision threshold, as visualized by ROC curves (Problem 5.1), and the last one behaves as a simple quantification of the cost of wireless acquisition of data (Problem 5.2). A trade-off between the amount of transmitted data and classification accu-

racy was expected.

Compared to Publication IV, the experiments in Publication V changed from threshold-agnostic AUC evaluation to fixing the decision threshold with respect to allowed FPR in the training set: low P_F was assumed desirable. The publication also added a third data set, the numerical simulations of a beam and spring. Finally, the experiments covered observing TPR and test set FPR versus several factors:

- two feature selection methods: CF and a random baseline,
- two detector models: NN and Gaussian density,
- three data sets: wooden model bridge[2], LANL bookshelf, and the numerical beam and spring simulation
- increasing damage magnitude (indicated by labels C), and
- number of features per sensor $D_S \in [2, 12]$.

The results with random selection are shown in Publication V, Figures 5 and 6, for the two detectors correspondingly, one subfigure for each data set, and one curve for each test set (damage level). The detection system suffers from poor discrimination and generalization, as the FPR values are close to TPR values and high test set FPR despite the value of $FPR = 0.15$ calibrated during training set. The value of $FPR = 0.15$ was set at such a high value, since the problem with poor discrimination ($FPR \approx TPR$) was more pronounced when attempting low values, like $FPR = 0.01$, and the number of training samples T_{tr} needs to be high enough with respect to FPR to reliably set a decision threshold. The only reasonable classification result resulted for the bookshelf data when using the Gaussian density model with moderate amount of measured features, $D_S \in [3, 9]$ (Figure 5c).

Publication V, Figures 8 and 9 show the corresponding results when using the CF approach. Again, both detectors suffer from poor class discrimination for the simulated beam data, but reasonable results are achieved for wooden bridge data at $D_S \in [4, 6]$. The Gaussian density model performed well in detecting damages in the LANL bookshelf data also with

[2]including only one damage location

CF (Figure 8c).

Publication V included also experiments with two centralized statistical damage detection methods [56, 57, 58] which rely on raw accelerometer data. Thus, the amount of data required from each sensor is significantly larger: $D_S = N \in \{50, 100, 200, 500, 1000\}$ in the experiments. On the other hand, the probability of false alarm was possible to be set as low as $P_F = 0.001$ and it was also realized for most of the test data, with the exception of the bookshelf data with shorter time windows ($D_S < 500$), as shown in Publication V, Figures 11 and 12. The figures also reveal the increase of sensitivity when D_S increases, especially for the smallest damages. From this work point of view, the comparison shows that accurate classifications are possible for the data sets.

7. Summary and Discussion

7.1 Summary

This work proposed a set of data processing stages aimed at wireless damage detection in SHM applications. First, extraction of frequency domain features from wireless accelerometers was considered, while also considering their mutual normalization to eliminate environmental variability. The resulting feature space was found to be high-dimensional, but also conjectured to have high amount of redundancy. Thus, several different dimensionality reduction schemes were proposed as the second stage, considering both centralized dimensionality reduction and a partly distributed scheme of selecting monitored features in a coordinated manner. Several general purpose statistical novelty detection methods were benchmarked for their use as the third stage of learning and deploying a decision region in the selected feature space. As a final and experimental data processing stage, general level performance criteria were defined and used for empirically exploring the behavior of the proposed detection systems.

Publication I was described to propose an actual feature selection process in [30]. Searching through random sets of features is not useful in the wrapper-based selection with a supervised binary classifier, as there are no samples from the damaged structure at the time when the selection has to be made. What Publication I showed was that the proposed feature space did contain small and useful sets of features for damage detection: a kind of validation for extracting features with the Goertzel algorithm and transmissibility magnitude estimates. The Goertzel algorithm was later implemented as a part of wireless damage detection system in [8].

Publication II included the use of pure novelty detection methods and

was also used as an example to motivate practical usefulness of novelty detection framework in [86]. Publication II also benchmarked combinations of centralized projection methods and novelty detection methods, while observing the effect of projected feature space dimensionality on the detection accuracy. Random projections were found to perform well in case the output dimensionality is sufficient, but at the same time, higher dimensionalities became a problem for the classifier stage. PCA and curvilinear component analysis were found to perform better with lower output dimensionalities. The simpler classifiers relying on nearest neighbor rule and Gaussian density estimation were found more robust in high-dimensional feature spaces.

Publication III proposed modeling the transmissibility magnitude data in terms of three-way factorizations. This provided both means for projecting the data onto a low-dimensional *temporal* factor, and means for feature selection so that most prominent factors in *spectral* and *spatial* loadings are represented in the selected features.

Publication IV presented the idea of wireless accelerometers as a community that could produce local ratings to express preferences in which features to measure, and a collaborative filtering scheme to select a globally coordinated set of measurements for monitoring. The CF approach was observed to lead to better classification accuracy than a random selection with similar amounts of data transmitted per sensor.

Publication V emphasizes the fact, that a complete detection system includes a rule for selecting an appropriate decision threshold: evaluating only class separability of damages of given size is not enough for a practical implementation. Comparison to fully centralized statistical damage detection methods showed, that while the amount of features transmitted per sensor could be significantly reduced by the proposed detection system (distributed DFT, CF, and NN/Gaussian detector), the sensitivity and specificity of the resulting system was not competitive. The wireless damage detection system developed in this work makes more accurate detections than a random baseline, but did not achieve the kind of accuracy possible with complete accelerometer data measured with wired sensors.

7.2 Reliability and Validity

A few notes on the practical applicability are in order. The use of DFT, as discussed in Chapter 2, implies steady state or stationary signal. Thus,

the vibration should be “ambient”, i.e., measurement period should occur after any transient shocks etc. applied to the structure. This may limit the use with structures that are in continuous operational use.

Transmissibility addresses only some of the environmental variability, not changes in temperature or location of input excitation etc. The results apply to the data acquired in approximately constant laboratory conditions. One potential direction for future research is to include data with measured or even controlled temperature variations, and either formulate temperature compensated transmissibility features or leave the responsibility of temperature compensation to the classifier model.

Transmissibility magnitude is not symmetric with respect to the sensors, but it was considered as such in CF feature selection. Another option could have been *transmissibility power* [50], which utilizes absolute value of a logarithm to achieve invariance to order of sensors ($|log(T^{s_1,s_2})| = |log(T^{s_2,s_1})|$).

While the projection methods covered in Section 3.2 have certain guarantees, there are also potential issues. Random projections and Johnson-Lindenstrauss lemma apply to a set of points, while the monitoring happens over a continuous stream of samples. Thus, it has to be noted that the role of RP is to provide a feature space that represents training data well and it is only conjectured to be sufficient for detecting the differences in test data. Similarly, using PCA is a valid idea only if the (future) damages affect the maximum variance components of the healthy state signal.

In general, a *novelty* detector cannot be guaranteed to detect any *damage*. There are at least two reasons for this: environmental variability and insufficient coverage. On one hand, (unexpected) environmental variability leads to false alarms, which may subsequently lead to setting a decision threshold that results in an insensitive detector. On the other hand, insufficient coverage leads to false negative detections, i.e., if there are no sensors monitoring the part of the structure (or vibration spectrum) where a damage is introduced. Transmissibility is known to be insensitive to changes at boundary conditions and outside the sensor array [50], which may explain some of the false negatives with the wooden model bridge data in the experiments of this work.

Some of the potential pitfalls of this work lie in feature selection for novelty detection, “feature engineering”, and data snooping. A failed experimental approach would be such that a single constant set of data was used in benchmarking different systems over and over, finally leading to

a single seemingly best design or a set of features. This work started with the wooden model bridge data, but then exposed the proposed systems also to other data sets: the LANL bookshelf data, which turned out to be somewhat easier to classify, and simulated data, which turned out more challenging.

Choosing some mathematically convenient and general level evaluation criteria could be criticized for *ludic fallacy* [77], or "the misuse of games to model real-life situations". Applied to SHM, the argument is that complex structures could develop damages in unexpected ways and being totally prepared to detect any damage is hopeless. On the other hand, there is the idea of a *rational agent* [73], which in this case translates to performing measurements that are *expected* to be worth it, given the finite resources and knowledge. Quantifying the ultimate utility of information as a function of detection accuracy, or the eventual cost of measuring data with a particular WSN system, and estimating a single utility value, are beyond this work.

The rationality idea reveals a problem with the incentives in deploying SHM systems: the users and owners of a structure are interested in SHM only if they expect the structure to fail in the near future, maybe because it is old or already damaged. On the other hand, installing an SHM system to a new structure may be perceived as admitting that there is likely something wrong with the structure.

7.3 Future Directions

In feature extraction, instead of picking individual DFT bins for monitoring, one option could be Empirical mode decomposition (EMD) [46], but it was considered too complex to implement in the considered hardware. In the projection stage, using a *sketch* [16, 48] of some sort, instead of a static projection fitted to the training data, could prove more suitable for streaming applications.

The problem of finding and validating proper features for monitoring can also be formulated as a machine learning problem. There is an interesting topic of deep architectures and deep learning [5] which could potentially become utilized in WSNs. Deep architectures refer to neural network models that aim at maximizing expressive power for a given amount of computational entities (neurons), hence a possibility for relevance in some WSN setting.

Deep learning can also be considered as finding useful representations on the intermediate layers of the neural networks, or features. Then, the SHM system could be claimed to have even more "intelligence", not just relying on manually engineered intermediate feature representations. Applicability in the novelty detection setting, or in the presence of additional domain specific assumptions, is left for future work.

Three-way analysis was done with PARAFAC over transmissibility magnitude features, but maybe *individual differences in scaling* (INDSCAL) or *decomposition into directional components* (DEDICOM) [54] over the magnitude data would make more sense. The model could have sensors in two separate modes ($\mathbb{R}^{S\times S\times K\times T}$), with certain symmetry between the two spatial modes, instead of having sensor pairs as a single mode. This would correspond to including the transmissibility idea into the dimensionality reduction method in a suitable SHM-specific way.

The coordinated monitoring problem was inspired by collaborative filtering techniques and consequently named according to the concepts of that field. It may well be, that the solutions can also be formalized in terms of compressed sensing (CS) [24, 3, 25, 87], overlapping correlation clustering (OCC) [9], or subspace clustering [89].

This work has been an example of balancing between the amount of communication and the amount of local computation or memory on network nodes. This kind of trade-offs have been presented also in [27, 90].

Bibliography

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, June 2003.

[2] David W. Allen, Sergio Castillo, Amanda L. Cundy, Charles R. Farrar, and Robert E. McMurry. Damage detection in building joints by statistical analysis. In *Proceedings of SPIE, the International Society for Optical Engineering*, volume 4359, pages 955–961, 2001.

[3] Waheed Bajwa, Jarvis Haupt, Akbar Sayeed, and Robert Nowak. Compressive wireless sensing. In *Proceedings of the 5th international conference on Information processing in sensor networks*, IPSN '06, pages 134–142, New York, NY, USA, 2006. ACM.

[4] Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall Information and System Sciences Series. Prentice-Hall, 1993.

[5] Yoshua Bengio and Olivier Delalleau. On the expressive power of deep architectures. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *ALT 2011*, volume 6925 of *Lecture Notes in Artificial Intelligence*, pages 18–36. Springer, 2011.

[6] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, 2001.

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 2006.

[8] Maurizio Bocca, Janne Toivola, Lasse M. Eriksson, Jaakko Hollmén, and Heikki Koivo. Structural health monitoring in wireless sensor networks by the embedded Goertzel algorithm. In *IEEE/ACM Second International Conference on Cyber-Physical Systems (ICCPS 2011)*, pages 206–214. IEEE, April 2011.

[9] Francesco Bonchi, Aristides Gionis, and Antti Ukkonen. Overlapping correlation clustering. In Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu, editors, *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, pages 51–60, Washington, DC, USA, 2011. IEEE Computer Society.

[10] Athanassios Boulis, Saurabh Ganeriwal, and Mani B. Srivastava. Aggregation in sensor networks: an energy-accuracy trade-off. *Ad Hoc Networks*, 1(2–3):317–331, 2003.

[11] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[12] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 93–104, New York, NY, USA, 2000. ACM.

[13] Rasmus Bro and Henk A. L. Kiers. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, 17(5):274–286, May 2003.

[14] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58, July 2009.

[15] Nitesh V. Chawla. Data mining for imbalanced datasets: An overview. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 875–886. Springer, 2010.

[16] Graham Cormode and S. Muthukrishnan. An improved data stream summary: The count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

[17] Emre Ilke Cosar, Aamir Mahmood, and Mikael Björkbom. A-Stack: A TDMA framework for reliable, real-time and high data-rate wireless sensor networks. Technical report, Aalto University, Espoo, Finland, February 2012.

[18] George Coulouris, Jean Dollimore, and Tim Kindberg. *Distributed systems: concepts and design*. Pearson Education Limited, Harlow, UK, 4th edition, 2005.

[19] David Culler, Deborah Estrin, and Mani Srivastava. Guest editors' introduction: Overview of sensor networks. *Computer*, 37(8):41–49, August 2004.

[20] Pierre Demartines and Jeanny Hérault. Curvilinear component analysis: A self organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, 1997.

[21] A. P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[22] Arnaud Deraemaeker, André Preumont, and Jyrki Kullaa. Modeling and removal of environmental effects for vibration based SHM using spatial filtering and factor analysis. In *Proceedings of 24th Conference and Exposition on Structural Dynamics 2006 (IMAC - XXIV)*, pages 1803–1812, Red Hook, NY, USA, February 2006. Society for Experimental Mechanics, Curran Associates, Inc.

[23] Isabel Dietrich and Falko Dressler. On the lifetime of wireless sensor networks. *ACM Transactions on Sensor Networks*, 5(1):1–38, January 2009.

[24] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

[25] Marco F. Duarte, Michael B. Wakin, Dror Baron, and Richard G. Baraniuk. Universal distributed sensing via random projections. In *Proceedings of the 5th international conference on Information processing in sensor networks*, IPSN '06, pages 177–185, New York, NY, USA, 2006. ACM.

[26] Robert P.W. Duin. PRTools: A Matlab toolbox for pattern recognition, March 2011. version 4.1.6. from http://prtools.org/.

[27] Kevin Fall. A delay-tolerant network architecture for challenged internets. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '03, pages 27–34, New York, NY, USA, 2003. ACM.

[28] Charles R. Farrar and Keith Worden. An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society A*, 365:303–315, 2007.

[29] Timothy R. Fasel, Hoon Sohn, and Charles R. Farrar. Application of frequency domain ARX models and extreme value statistics to damage detection. In Shih-Chi Liu, editor, *Smart Structures and Materials 2003: Smart Systems and Nondestructive Evaluation for Civil Infrastructures*, volume 5057 of *Proceedings of SPIE*, pages 145–156. The International Society for Optical Engineering, March 2003.

[30] Fabio Federici, Fabio Graziosi, Marco Faccio, Andrea Colarieti, Vincenzo Gattulli, Marco Lepidi, and Francesco Potenza. An integrated approach to the design of wireless sensor networks for structural health monitoring. *International Journal of Distributed Sensor Networks*, 2012(594842):1–16, 2012.

[31] Elói Figueiredo, Gyuhae Park, Joaquim Figueiras, Charles Farrar, and Keith Worden. Structural health monitoring algorithm comparisons using standard data sets. Technical Report LA-14393, Los Alamos National Laboratory, Los Alamos, NM, USA, March 2009.

[32] João Gama and Mohamed Medhat Gaber, editors. *Learning from Data Streams: Processing Techniques in Sensor Networks*, volume 1. Springer, 2007.

[33] Gerald Goertzel. An algorithm for evaluation of finite trigonometric series. *American Mathematical Monthly*, 65:34–35, January 1958.

[34] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(1):1157–1182, 2003.

[35] Gregory Hackmann, Weijun Guo, Guirong Yan, Chenyang Lu, and Shirley Dyke. Cyber-physical codesign of distributed structural health monitoring with wireless sensor networks. In *Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems*, ICCPS '10, pages 119–128, New York, NY, USA, 2010. ACM.

[36] Gregory Hackmann, Fei Sun, Nestor Castaneda, Chenyang Lu, and Shirley Dyke. A holistic approach to decentralized structural damage localization

using wireless sensor networks. In *Proceedings of the 2008 Real-Time Systems Symposium*, RTSS '08, pages 35–46, Washington, DC, USA, 2008. IEEE.

[37] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, 2001.

[38] David J. Hand. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77:103–123, October 2009.

[39] Richard A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970. (University Microfilms, Ann Arbor, Michigan, No. 10,085).

[40] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc., USA, 1996.

[41] Paul Hayton, Simukai Utete, Dennis King, Steve King, Paul Anuzis, and Lionel Tarassenko. Static and dynamic novelty detection methods for jet engine health monitoring. *Philosophical Transactions of the Royal Society A*, 365(1851):493–514, 2007.

[42] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22:5–53, January 2004.

[43] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[44] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, September 1933.

[45] Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, August 1966.

[46] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A*, 454(1971):903–995, 1998.

[47] Jeanny Hérault, Claire Jausions-Picaud, and Anne Guérin-Dugué. Curvilinear Component Analysis for High-Dimensional Data Representation: I. Theoretical Aspects and Practical Use in the Presence of Noise. In *Engineering Applications of Bio-Inspired Artificial Neural Networks, IWANN '99*, volume 1607 of *Lecture Notes in Computer Science*, pages 625–634. Springer, 1999.

[48] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM*, 53(3):307–323, May 2006.

[49] Timothy J. Johnson and Douglas E. Adams. Transmissibility as a differential indicator of structural damage. *Journal of Vibration and Acoustics*, 124(4):634–641, 2002.

[50] Timothy James Johnson. Analysis of dynamic transmissibility as a feature for structural damage detection. Master's thesis, Purdue University, August 2002.

[51] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, Providence, RI, USA, 1984.

[52] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. An online algorithm for segmenting time series. In *Proceedings of the IEEE International Conference on Data Mining*, pages 289–296, November 2001.

[53] Teuvo Kohonen. *Self-organization and associative memory*. Springer, New York, NY, USA, 3rd edition, 1989.

[54] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[55] Jyrki Kullaa. Elimination of environmental influences from damage-sensitive features in a structural health monitoring system. In Daniel L. Balageas, editor, *Proceedings of the First European Workshop on Structural Health Monitoring 2002*, pages 742–749. Onera, DEStech Publications Inc, July 2002.

[56] Jyrki Kullaa. Eliminating environmental or operational influences in structural health monitoring using the missing data analysis. *Journal of Intelligent Material Systems and Structures*, 20(11):1381–1390, 2009.

[57] Jyrki Kullaa. Sensor validation using minimum mean square error estimation. *Mechanical Systems and Signal Processing*, 24(5):1444–1457, 2010.

[58] Jyrki Kullaa. Distinguishing between sensor fault, structural damage, and environmental or operational effects in structural health monitoring. *Mechanical Systems and Signal Processing*, 25(8):2976 – 2989, 2011.

[59] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, DMKD '03, pages 2–11, New York, NY, USA, 2003. ACM.

[60] Los Alamos National Laboratory, Engineering Institute. Experimental data for download. Internet, 2012. Accessed 15 Aug. 2012.

[61] Jerome Peter Lynch, Arvind Sundararajan, Kincho H. Law, Anne S. Kiremidjian, and Ed Carryer. Embedding damage detection algorithms in a wireless sensing unit for operational power efficiency. *Smart Materials and Structures*, 13(4):800–810, June 2004.

[62] Nuno M.M. Maia, Raquel A.B. Almeida, António P.V. Urgueira, and Rui P.C. Sampaio. Damage detection and quantification using transmissibility. *Mechanical Systems and Signal Processing*, 25(7):2475–2483, 2011.

[63] Markos Markou and Sameer Singh. Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[64] Sanjit K. Mitra. *Digital Signal Processing: A Computer-Based Approach*. McGraw-Hill, New York, NY, USA, second edition, 2002.

[65] Morten Mørup. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):24–40, January 2011.

[66] Luis E. Mujica, Josep Vehí, Magda Ruiz, Michel Verleysen, Wieslaw Staszewski, and Keith Worden. Multivariate statistics process control for dimensionality reduction in structural assessment. *Mechanical Systems and Signal Processing*, 22(1):155–171, 2008.

[67] Frank Neitzel, Sven Weisbrich, Martin Lehmann, and Wolfgang Niemeier. Investigation of low-cost accelerometer, terrestrial laser scanner and ground-based radar interferometer for vibration monitoring of bridges. In Christian Boller, editor, *Proceedings of the 6th European Workshop on Structural Health Monitoring (EWSHM 2012)*, pages 542–551. Deutsche Gesellschaft für Zerstörungsfreie Prüfung, July 2012.

[68] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, September 1962.

[69] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series*, 2:559–572, 1901.

[70] Miguel Á. Prada, Jaakko Hollmén, Janne Toivola, and Jyrki Kullaa. Three-way analysis of structural health monitoring data. In Samuel Kaski, David J. Miller, Erkki Oja, and Antti Honkela, editors, *Proc. of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010)*, pages 256–261. IEEE, August 2010.

[71] John G. Proakis and Masoud Salehi. *Digital communications*. McGraw-Hill, New York, NY, USA, 2008.

[72] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.

[73] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, Inc., 2nd edition, 2003.

[74] Hoon Sohn. Effects of environmental and operational variability on structural health monitoring. *Philosophical Transactions of the Royal Society A*, 365(1851):539–560, February 2007.

[75] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey on collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4:1–4:19, January 2009.

[76] John A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, June 1988.

[77] Nassim Nicholas Taleb. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. Penguin Books Ltd, May 2007.

[78] Lionel Tarassenko, David A. Clifton, Peter R. Bannister, Steve King, and Dennis King. *Encyclopedia of Structural Health Monitoring*, chapter 35 Novelty Detection, pages 1–23. Wiley, 2009.

[79] David M. J. Tax. *One-class classification; Concept-learning in the absence of counter-examples*. PhD thesis, Delft University of Technology, June 2001.

[80] David M. J. Tax. DDtools, the Data Description Toolbox for Matlab, March 2011. version 1.9.0 from http://prlab.tudelft.nl/david-tax/dd_tools.html.

[81] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Elsevier, San Diego, CA, USA, second edition, 2003.

[82] Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[83] Keith Worden, Charles R. Farrar, Graeme Manson, and Gyuhae Park. The fundamental axioms of structural health monitoring. *Proceedings of the Royal Society A*, 463(2082):1639–1664, 2007.

[84] Keith Worden and Graeme Manson. The application of machine learning to structural health monitoring. *Philosophical Transactions of the Royal Society A*, 365(1851):515–537, 2007.

[85] Keith Worden, Graeme Manson, and David Allman. Experimental validation of a structural health monitoring methodology: Part I. novelty detection on a laboratory structure. *Journal of Sound and Vibration*, 259(2):323–343, 2003.

[86] Yingchao Xiao, Huangang Wang, Wenli Xu, and Junwu Zhou. L1 norm based KPCA for novelty detection. *Pattern Recognition*, 46(1):389–396, 2013. Available online 4 July 2012.

[87] Allen Y. Yang, Michael Gastpar, Ruzena Bajcsy, and S. Shankar Sastry. Distributed sensor perception via sparse representation. *Proceedings of the IEEE*, 98(6):1077–1088, April 2010.

[88] Huseyin Yiğitler, Aamir Mahmood, Reino Virrankoski, and Riku Jäntti. Recursive clock skew estimation for wireless sensor networks using reference broadcasts. *IET Wireless Sensor Systems*, 2(4):338–350, December 2012.

[89] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387, 2012.

[90] Mujdat Çetin, Lei Chen, John W. Fisher III, Alexander T. Ihler, Randolph L. Moses, Martin J. Wainwright, and Alan S. Willsky. Distributed fusion in sensor networks. *IEEE Signal Processing Magazine*, 23(4):42–55, July 2006.

Errata

Publication I

The author of citation [2] is “Leon W. Couch, II”

Publication II

The 3rd author of citation [6] is “Guido De Roeck”

Publication IV

The published proceedings have some geometrical errors on page 992 (6th page of the article): Figures 7b and 7d have shifted edges, and the text below has extra line feeds at “...(larger with larger D), and shows how large feature sets the majority voting method leads to.” and “($AUC = 1$)”.

Aalto-DD 80/2014

ISBN 978-952-60-5712-5
ISBN 978-952-60-5713-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL
DISSERTATIONS**