FEDERAL UNIVERSITY OF PARÁ

INSTITUTE OF NATURAL AND EXACT SCIENCES

FACULTY OF COMPUTING

A NOVEL UNSUPERVISED APPROACH BASED ON A GENETIC ALGORITHM APPLIED TO STRUCTURAL DAMAGE DETECTION IN BRIDGES

MOISÉS FELIPE MELLO DA SILVA

UFPA / ICEN / FACOMP

Guamá University Campus

Belém-Pará-Brazil

2015

FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF NATURAL AND EXACT SCIENCES
FACULTY OF COMPUTING

MOISÉS FELIPE MELLO DA SILVA

A NOVEL UNSUPERVISED APPROACH BASED ON A GENETIC ALGORITHM APPLIED TO STRUCTURAL DAMAGE DETECTION IN BRIDGES

UFPA / ICEN / FACOMP
Guamá University Campus
Belém-Pará-Brazil
2015

FEDERAL UNIVERSITY OF PARÁ

INSTITUTE OF NATURAL AND EXACT SCIENCES

FACULTY OF COMPUTING

MOISÉS FELIPE MELLO DA SILVA

A NOVEL UNSUPERVISED APPROACH BASED ON A GENETIC ALGORITHM APPLIED TO STRUCTURAL DAMAGE DETECTION IN BRIDGES

This work was submitted to the Examining Board of the Faculty of Computing of the UFPA to obtain the Bachelor's Degree in Computer Science.

Advisor: Ph.D. Claudomiro de Souza de Sales Junior

Co-advisor: MSc Adam Dreyton Ferreira dos Santos

UFPA / ICEN / FACOMP

Guamá University Campus

Belém-Pará-Brazil

2015

Acknowledgements

First, I thank to God for my life and all achievements obtained until this moment.

I thank to my parents: José Carlos Lima da Silva and Cátia Regina Mello da Silva for the love, care and unconditional support provided to me since my choice for Computer Science course until its end.

To the professor Ph.D. Claudomiro de Souza de Sales Junior that accompany and drive my career since I got in this university and was my door entrance to Applied Electromagnetism Laboratory (LEA)[1] in which I reached valorous experiences and friendships along these almost three years of many researches and publications.

In special to MSc Adam Dreyton Ferreira dos Santos that is not just an attentive tutor and working partner but also a reliable friend and model of professional.

Thanks to professors Ph.D. Marco José de Sousa and Ph.D. João Crisóstomo Weyl Albuquerque Costa for friendship, companionship, experience, guided and knowledge provided in LEA.

Thanks to my best friend, Thiago Araújo, who since childhood has shared with me uncountable stories and funny moments. And Luena Canavieira, which more than my best friend is my sister, not of blood but of union. For you two a thank you.

To my other friends (not less important) with who I always share smiles and histories at the weekends: Alessandra Silva, Vivian Castro, Thiago Gama, Lívia Vitória, Rafael Reis, Israel Hounsou, Paula Castro and so many others.

To my graduation colleagues that despite the jokes always were solicitous to me: Marcos Senna, Lana Priscila, Bleno Vale, Dan Jonathan, Daniel Gomes, Ariane Garcia, Arthur Moraes, Bruno Fukushima, Diego Abreu e Leonardo Afonso.

To my others working friends and LEA's members: Kárytha Nascimento, Waldeir Monteiro, Daniel Levy, Max Freitas, Daynara Dias, Alexandre Van Der Ven, Brenda Penedo, Gustavo Ikeda, André Fernandes, Wederson Medeiros, Roberto Almeida, Aline Ohashi, Seiji Fujihara, Reginaldo Filho, Dércio Mate, Roberto Menezes, Gilvan Borges, Lamartine Souza, Clenilson Silveira, and Fabrício Lobato. In special for the always good humoured Cindy Fernandes and my others working and course friends: Alexandre Freitas, Alessandra Araújo, Caio Rodrigues and Walisson Cardoso which are always with me.

[1] `http://www.lea.ufpa.br`

My acknowledgements for the financial support received from Vale[2], Fapespa[3], CNPq[4] and CAPES[5].

[2] http://www.vale.com/
[3] http://www.fapespa.pa.gov.br/
[4] http://www.cnpq.br/
[5] http://capes.gov.br/

"What man is a man who does not make the world better?"
Balian of Ibelin.

"The characteristic of a genuine heroism and geniality is its persistency. All men have wandering impulses, fits and starts of generosity and brilliance. But when you have resolved to be great, abide by yourself, and do not weakly try to reconcile yourself with the world. The heroic cannot be the common, nor the common the heroic."
Ralph Waldo Emerson (adapted).

Summary

List of Figures

List of Tables

List of abbreviations and acronym

AANN	Auto-Associative Neural Network
MSD	Mahalanobis-Squared Distance
GMM	Gaussian Mixture Models
NLPCA	Nonlinear Principal Component Analysis
GADBA	Genetic Algorithm for Decision Boundary Analysis
PCA	Principal Component Analysis
SHM	Structural Health Monitoring
DI	Damage Indicator
ROC	Receiver Operating Characteristics
CHM	Concentric Hypersphere Method
GA	Genetic Algorithm
SPR	Statistical Pattern Recognition
BMS	Bridge Management System
ML	Maximum Likelihood
EM	Expectation-Maximization
MCMC	Markov-chain Monte Carlo
DC	Damage Condition
BC	Baseline Condition
AIC	Akaike Information Criterion

Resumo

Este trabalho propõe um novo algoritmo genético não paramétrico de aprendizado não supervisionado para análise de fronteira de decisão (GADBA) a partir da identificação de agrupamentos, aplicando-o na detecção de danos estruturais em pontes, ainda que sob a influência de efeitos lineares e não-lineares causados por condições ambientais e operacionais, incluindo danos e falha estrutural. Além disso, são apresentados uma nova expressão para caracterização de componentes estruturais e um método para eliminar componentes redundantes baseado em análise de densidade de distribuição. O desempenho da abordagem proposta é avaliado através da comparação com três dos mais difundidos métodos da literatura utilizando dois conjuntos de dados baseados em séries temporais extraídos a partir de sistemas de monitoramento instalados em duas diferentes pontes: Ponte Z-24 e Ponte Tamar. Os resultados demonstram que o algoritmo proposto é mais competente na tarefa de contornar condições normais da estrutura e identificar componentes estruturais. Adicionalmente, o método revela um melhor desempenho de classificação frente as demais técnicas do estado-da-arte em termos de falsas-positivas e falsas-negativas indicações de dano, sugerindo a sua aplicabilidade em cenários reais de monitoramento.

Palavras-chave: Monitoramento de saúde estrutural, Algoritmo genético, Detecção de danos, Condições ambientais, Condições operacionais, Clusterização.

Abstract

This work proposes an unsupervised and non-parametric genetic algorithm to decision boundary analysis (GADBA) from the creation of clusters. This technique is applied to detect structural damage even in the presence of linear and non-linear effects caused by environmental and operational conditions including structural cracks and damage. Moreover, a new expression for characterize components and a method for eliminate redundant components based on the analysis of density distribution are presented. The performance of the proposed algorithm is demonstrated through comparison with three of the most widespread techniques in the literature using two time series data sets extracted from monitoring systems deployed in two different bridges: the Z-24 Bridge and Tamar Bridge. The results demonstrate that the proposed algorithm is more competent in the task of fitting normal conditions and identify structural components. This technique revealed to have better classification performance than the state-of-art algorithms in terms of false-positive and false-negative indications of damage, suggesting its applicability in real world monitoring systems.

Keywords: Structural health monitoring, Genetic algorithm, Damage detection, Environmental conditions, Operational conditions, Clustering

1 Introduction

Improved and more continuous condition assessment of bridges (FIGUEIREDO; MOLDOVAN; MARQUES, 2013) has been demanded by our society to better face the challenges presented by aging civil infrastructure. In the last two decades, bridge condition assessment approaches have been developed independently based on two concepts: bridge management systems (BMSs) and structural health monitoring (SHM). The BMS is a visual inspection-based decision-support tool developed to analyse engineering and economic factors and to assist the authorities in determining how and when to make decisions regarding maintenance, repair and rehabilitation of structures (LEE et al., 2008; WENZEL, 2009). On the other hand, the SHM traditionally refers to the process of implementing monitoring systems to measure in real time the structural responses, in order to detect anomalies and/or damage at early, current and future stages (FARRAR; WORDEN, 2007).

While the BMS has already been accepted by the bridge owners around the world (MIYAMOTO; KAWAMURA; NAKAMURA, 2001; ESTES; FRANGOPOL, 2003; GATTULLI; CHIARAMONTE, 2005), even though with inherent limitations imposed by the visual inspections, the SHM is becoming increasingly attractive due to its potential ability to detect damage at varying stages and near real-time, with the consequent life-safety and economical benefits (WORDEN et al., 2007). The author believe that all approaches to SHM, as well as all traditional non-destructive evaluation techniques, can be posed in the context of a statistical pattern recognition (SPR) problem. Thus, the SPR paradigm for the development of SHM solutions can be described as a four-phase process (FARRAR; DOEBLING; NIX, 2001): (1) operational evaluation, (2) data acquisition, (3) feature extraction, and (4) statistical modeling for feature classification.

This work focus in the statistical modeling for feature classification purposes concerned with the implementation of algorithms that analyse and learn the distributions of the extracted features in an effort to determine the structural health condition (WORDEN; MANSON, 2007). Inherent in the data acquisition, feature extraction, and statistical modeling is data normalization, which is the process of separating changes in damage-sensitive features caused by damage from those caused by varying operational and environmental conditions (SOHN; FARRAR, 2001). Actually, these influences on the structural response have been cited as one of the major challenges to the complete transit of SHM technology from research to practice (SOHN, 2007; XIA et al., 2012). Considering the fourth phase of the SHM process numerous studies have established the robust

concept of automatically discovered the number of normal and stable states conditions of bridges, even when it is affected by extreme operational and environmental conditions (FIGUEIREDO; CROSS, 2013; FIGUEIREDO et al., 2014a). In these works, the damage detection is carried out on the basis of an outlier detection strategy using distance metrics (SOHN, 2007) which permits to track the outlier formation in time in relation to the chosen main groups of states.

In this work, a novel unsupervised approach using genetic algorithm (GA) to detect structural damage in bridges is proposed, namely genetic algorithm for decision boundary analysis (GADBA). Combined with the strong search capability inherent in GAs, this work presents a new term to characterize the main clusters/components of features/observations that correspond to the normal state conditions of a bridge and a new algorithm to regularize these number of clusters, namely concentric hyperspheres method (CHM), mitigating the cluster redundancy. In that regard, a first step is proposed in order to automatically discover the main state conditions of bridges by clustering the training observations according to the closest centroids, which are targets of the optimization performed with genetic operators of the GA. This optimization aims to find feasible state conditions of the bridge, defining boundary regions between the clusters, as well as reducing the number of discovered state conditions. A second step is related to the damage detection method, where is computed the Euclidean distances between the test observations and the centroids optimized in the first step. The minimum distance represents the damage indicator (DI) of observation and indirectly the cluster which it belongs.

The proposed approach is performed on standard data sets from the Z-24 Bridge, in Switzerland, and the Tamar Bridge in England. The classification performance is evaluated through receiver operating characteristics (ROC) curves, which are a means of determining performance on the basis of Type I/Type II error trade-offs. In SHM, in the context of damage identification, a Type I error is a false-positive indication of damage and a Type II error is a false-negative indication of damage.

The overall organization of this work is as follows. Section 2 describes all the new constraints and mechanisms developed to cluster the normal state conditions of bridges, by using the GADBA-based approach, and to detect damage based on those identified clusters. Section 3 highlights a structural description of both bridges as well as a summary of the data sets from those bridges that encompass a wide spectrum of challenges associated with practical damage detection problems. Section 4 presents the applicability of the proposed approach on those real-world data sets and compares its performance with two approaches. Finally, Section 5 summarizes and discusses the implementation and analysis carried out in this study.

1.1 Related work

Damage detection based on machine learning algorithms applied to bridges has received a lot of attention in the past few years due to its highly improved performance to model and separate changes in damage-sensitive features caused by damage from those caused by varying operational and environmental conditions. Compared to approaches that consist of measuring the parameters related to operational and environmental variations (e.g. live loads and temperature), these algorithms pave the way for data-based models applicable to structural systems of arbitrary complexity, intends to avoid the measure of operational and environmental variations and physics-based model approaches such as finite element analysis. Note that for most civil engineering infrastructure where SHM systems are applied, the unsupervised learning algorithms are often required because only data from the undamaged condition are available (FARRAR; WORDEN, 2013). Some of the traditional unsupervised machine learning algorithms and their adaptations for damage detection in bridges are discussed below.

The mahalanobis-squared distance (MSD) algorithm is one of the most traditional methods for damage detection, having widespread use in real scenarios due to its characteristic to identify outliers (WORDEN; MANSON, 2007; WORDEN et al., 2007; NGUYEN; CHAN; THAMBIRATNAM, 2014). Those abnormal observations appear inconsistent with the rest of the data and therefore is believed to be generated by an alternative mechanism that is not related to the normal conditions established with a mean vector and a covariance matrix derived from baseline data sets. However, when nonlinearities are present in the observations, the MSD fails in modeling the normal conditions of a bridge because it assumes the baseline data sets as multivariate Gaussian distributed (FIGUEIREDO; CROSS, 2013).

Neural networks are also intelligent strategies to filter operational and environmental factors in the damage-sensitive features, detecting structural damage based on some distance metric between the unfiltered and filtered features by the network trained with datasets from normal conditions. The major unsupervised approach in this scenario is the nonlinear principal component analysis (NLPCA) (SOHN; WORDEN; FARRAR, 2002; HSU; LOH, 2010; HAKIM; RAZAK, 2014), which is based on the auto-associative neural network. This subclass of neural networks consists of three hidden layers: the mapping layer, the bottleneck layer, and de-mapping layer (KRAMER, 1991). The number of nodes in the bottleneck layer is related to the number of factors that should be filtered. The main challenge of this approach in real SHM systems is to know in advance the number of factors that can mask the damage-sensitive features, since the criteria defined in (KRAMER, 1991) only estimate the number of nodes in the mapping and de-mapping layers.

In (FIGUEIREDO et al., 2011), Figueiredo performed a comparison study of sev-

eral unsupervised machine learning algorithms on standard data sets. This study was performed upon experimental vibration monitoring tests in the laboratory using a three-story frame structure with different configurations. The operational and environmental effects were simulated by stiffness or mass changes, while damage was simulated with a bumper mechanism causing a nonlinear effect due to collisions. The four models chosen were based on a NLPCA, factor analysis, the MSD, and singular value decomposition. The overall analysis provided by this article has demonstrated that the NLPCA and MSD algorithms had the best classification performance when one wants to minimize false-negative indications of damage and when life-safety issues are the primary motive for deploying the SHM system.

New concepts are developed in (FIGUEIREDO; CROSS, 2013; FIGUEIREDO et al., 2014a) based on a two-steps damage detection strategy. In their first step, those works apply the Gaussian mixture models (GMMs) algorithm in order to model the main clusters that correspond to the normal and stable state conditions of a bridge, even when it is affected by unknown operational and environmental conditions. In (FIGUEIREDO; CROSS, 2013), the parameters of the multivariate finite mixture models are estimated from the training data using the classical maximum likelihood (ML) estimation based on the expectation-maximization (EM) algorithm. On the other hand, in (FIGUEIREDO et al., 2014a), the parameters are estimated using a Bayesian approach based on a Markov-chain Monte Carlo (MCMC) method. The Bayesian approach stands as an improvement over the classical ML estimation based on the EM algorithm. For the second step, a simpler and well-know approach is proposed, the MSD algorithm, which permits to track the outlier formation in relation to the chosen main groups of states. These cluster-based approaches have surpassed the traditional unsupervised damage detection methods, e.g., based on the MSD and NLPCA algorithms (FIGUEIREDO; CROSS, 2013). However, when the structure response do not follow a multivariate Gaussian distribution, the algorithm fails in model structural normal conditions.

1.2 Clustering genetic algorithms

Genetic algorithms (CHAMBERS, 2000; GOLDBERG, 1989) are powerful techniques of stochastic optimization for search and objectives evaluation guided by evolutionary principles and natural genetics performing solutions of multimodal complex problems leading with different restrictions at the same time (DEB et al., 2002). There are many tasks in pattern recognition area which GAs are used in order to identify complex parameters, therefore its application in other problems in this same area appears as natural, specially in the clustering data field. Hence, some of the most widespread approaches for GA-based clustering are discussed in more detail below.

The GA-clustering (MAULIK; BANDYOPADHYAY, 2000) is a well-known unsupervised GA for solve clustering problems in m-dimensional Euclidean space $\mathbb{R}^m$. Its approach is very similar to the K-means (MACQUEEN, 1967) algorithm where a given set of points is divided into a number **K** of subsets by applying genetic operators (fitness, selection, crossover and mutation) in each individual generated randomly and selecting **K** points as the initial centers. Note that genetic operators are applied directly on the features of the individual resulting in most incisive changes of value. The final result is the best position of the centers in each subset. The principal challenge in this technique is estimate the correct number of data subsets **K** which represents a cluster configuration and needs to be indicated previously. By analogy, in SHM applications, the number of clusters indicate the number of main structural components. Since estimate the parameter **K** remains challenging, so the applicability of GA-clustering is not indicated for real SHM scenarios.

In (COWGILL; HARVEY; WATSON, 1999) is proposed a clustering algorithm called COWCLUS that uses GA as a global searching technique. In COWCLUS the function used to evaluate a single solution is the variance ratio criteria (VRC) which defines the cluster isolation and internal cluster homogeneity analysing the degree of isolation between different clusters. The principal goal of this technique is to determine the best partition of the data in subsets in such way that maximizes the VRC. The practical results demonstrated that approach has a superior performance than K-means and Ward's in terms of maximization of the VCR function. However, the COWCLUS limitation is related to the previous knowledge of the parameter **K** and its problem of stuck in local optima.

In the context of techniques inspired by hybrid concepts in (HALL; OZYURT; BEZDEK, 1999) is developed a genetically guided algorithm (GGA) for clustering applied to brain tissue quantization that uses objective functions from other two well known algorithms: fuzzy c-means (FCM) (WEN; CELEBI, 2011) and hard c-means (HCM) (RUNKLER; KELLER, 2012). This approach consists of the minimization of an adapted function from originals objectives used in FCM and HCM that rewrites the fuzzy partition matrix by other matrix that represents a measure of distance from each feature vector to all centroids contained in a single solution. The analysis provided by this article when GGA is compared with FCM and HCM has demonstrated that the GGA provides equivalent results in terms of a "good" clustering and is indicated only if its time cost can be reduced. The limitation of this algorithm consists of know the number of clusters to be modeling. This occurs due GGA to be based on HCM and FCM techniques, thereby it first step consists in run HCM and FCM procedures that require this prior knowledge of the partitions number.

2 Genetic algorithm for decision boundary analysis (GADBA)

In general, the GADBA capabilities for searching and optimizing are presented in this study with the purpose of grouping data into logical structural components given a maximum number of clusters, K_{max}, resulting in suitable geometric centers (centroids) for each cluster in the Euclidean space, $\mathbb{R}^m$. In particular, the task of the proposed CH algorithm is to support the automatic identification of the number of clusters, K, and to choose the appropriate centers $C = c_1, c_2, ..., c_K$ of each cluster, through the maximization of the objective function proposed for the GADBA, which contributes to use the lowest number of clusters as possible. Essentially, the GADBA-based approach performs the CH algorithm in the set of solutions in each generation, aiming to estimate the correct number of components through an agglomerative clustering process.

For general purposes in SHM, the training matrix $x \in \mathbb{R}^{n \times m}$ is composed of n observations under operational and environmental variability when the structure is undamaged, where m is the number of features per observation obtained during the feature extraction phase. The test matrix $\mathbf{Z} \in \mathbb{R}^{t \times m}$ is defined as a set of t observations collected during the undamaged/damaged conditions of the structure. Note that an observation represents a feature vector encoding the structural condition at a given time.

2.1 Individual representation

Each individual is a sequence of K_{max} points defined in $\mathbb{R}^m$ representing the centroids of a candidate solution as shown in Figure 1.

F(1, 1)	*F(1, ...)*	*F(1, m)*	*F(2, 1)*	*F(2, ...)*	*F(2, m)*	...	*F(K, 1)*	*F(K, ...)*	*F(K, m)*

Figure 1: Representation scheme of a single individual.

C_i is a centroid index of the i-th centroid and $F(i, j)$ is the value of the j-th dimension of the i-th centroid. The number of alleles stored in the same individual is variable and its size K can be ranging between $1, \cdots, K_{max}$, this means that an individual consists of a list of real values with maximum size equal to $m \times K_{max}$. Each gene is composed of a centroid having an integer index identifying the allele position on the

individual being possible disable some of genes applying CHM algorithm discussed in section 2.5. As an illustration, consider the following example.

Example 1. Defining $m = 2$, $K_{max} = 6$ and $K = 3$ there is an individual in the euclidean space of order two composed of three centroids. Thereby, the chromosome

$$(1, 51.6, 72.3) \longrightarrow (2, 18.3, 51.7) \longrightarrow (3, 34.0, 21.34)$$

is a representation of a clustering solution containing three centroids.

The process of create an initial population occurs choosing randomly a number K on the closed interval $1, \cdots, K_{max}$. The K centroids to be coded on the individual are also chosen randomly selecting K points from training set. The process is repeated for all $\mathbf{p}$ individuals to be generated and inserted in $\mathbf{P}$.

2.2 Genetic operators

Aiming to perform several tasks of mutation, parent selection and survival selection, herein three well-known methods are highlighted and adopted to support the GADBA-based approach. The mutation process controls the exploration of the solution space by means of performing changes in the individuals. In this study, this process is composed of two steps:

(i) the number of centroids is changed via a stochastic variation method. An increment rate is previously determined by computing the inverse of the number of clusters, $T_x = K_{\max}^{-1}$. A random real value T_r defined in the range $[0, 1]$ is used to determine the number of centroids to be enabled in the offspring individual by applying $K_{new} = \left\lceil \frac{T_r}{T_x} \right\rceil$. In the case of $K \leq K_{new} \leq K_{max}$, then the miss positions are completed by selecting $K_{new} - K$ observations at random from matrix x, otherwise the last centroids are eliminated;

(ii) the mutation occurs in each centroid position in a stochastic manner. A mutation probability p_{mut} is associated to every position, which is subjected to the Gaussian mutation,

$$F_{i,j} = F_{i,j} + \mathrm{N}(0, 1), \tag{2.1}$$

where $\mathrm{N}(0, 1)$ is a random number from a Gaussian distribution with zero mean and unitary standard deviation, and $F_{i,j}$ the real value of the i-th centroid in the j-th dimension.

The selection operator drives the searching towards a promising region in the feature space. The parent selection method is based on the well-known tournament with

reposition. This method creates a r subset by selecting $|r|$ individuals randomly from the population. Afterwards, only the best individual is selected from r and submitted to the crossover process with another individual selected in the same manner. Besides, the survival selection is based on the elitism concept (DEB et al., 2002), in which two sets of parents $\mathbf{I}_p$ and offspring $\mathbf{I}_c$ are joint, creating a set $\mathbf{I}_u = \mathbf{I}_p \bigcup \mathbf{I}_c$. Then, a new fitness value is calculated based on the Pareto Front and crowding distance. The solutions that compose the new set $\mathbf{I}_u$ are sorted in order to select the $|\mathbf{P}|$ better individuals as the new population set $\mathbf{P}$.

The stopping criteria are: when the maximum number of generations is reached and/or the difference of the fitness between the two best individuals, of the last two generations, is less than a given threshold ϵ (e.g., $\epsilon = 5$).

2.3 Recombination

Recombination performs the exploration towards the known solution space aiming to refine the prior knowledges. Although a lot of different recombination operators are suggested in the literature (HRUSCHKA et al., 2009; MITCHELL, 1998), in this study is developed a strategy that combines not only useful segments of different parents, but also the centroid positions. The recombination method operates in three steps regarding two parameters previously defined, p_{rec} and p_{pos}:

(i) for each pair of parents $\mathbf{P}_i$ and $\mathbf{P}_j$, if a random number $r \leq p_{rec}$, then two cut points π_1 and π_2 are randomly generated, corresponding to a range within centroid positions of both parents, such that $1 \leq \pi_1 < \pi_2 \leq \min(K_i, K_j)$. The centroids in the range are switched to form two offspring individuals. In the case of p_{rec} is not satisfied, then both parents become the new offspring individuals;

(ii) each centroid position receives a random number $r \in [0, 1]$, in such a way if $r \leq p_{pos}$ then, for each pair of parent genes, an arithmetic recombination is performed according with

$$F_{x,t}^{(i)} = F_{x,t}^{(i)} + (F_{y,t}^{(j)} - F_{x,t}^{(i)}) * T, \tag{2.2}$$

$$F_{x,t}^{(j)} = F_{x,t}^{(j)} + (F_{y,t}^{(i)} - F_{x,t}^{(j)}) * T, \tag{2.3}$$

where T is a random value defined in $[0, 1]$, and $F_{x,t}^{(i)}$ and $F_{x,t}^{(j)}$ are the t-th positions of the x-th centroid from the i-th and j-th parents, respectively;

(iii) finally, a length ratio, λ, defines the number of centroids enabled in each offspring individual. Note that the parents already have λ_i and λ_j length ratios associated with themselves,

$$\lambda = \frac{K}{K_{\max}}. \tag{2.4}$$

Hence, λ maps the number of clusters, K, to the interval $(0, 1]$. Hereafter, another arithmetic recombination is performed on the parents' length ratio to generate λ_i' and λ_j' for the offspring individuals. Thus, the number of clusters (K_i and K_j) enabled in the final offsprings are

$$K_i = \frac{\max(\lambda_i, \lambda_j)}{\lambda_i'}, \tag{2.5}$$

$$K_j = \frac{\max(\lambda_i, \lambda_j)}{\lambda_j'}. \tag{2.6}$$

2.4 Term for characterizing components

Based on the approaches to create clusters from circular distributions (MACQUEEN, 1967), a nonlinear metric to characterize different clusters is proposed. This metric is used as the objective function, which intends to evaluate different set of clustering solutions by taking into account the observation dispersion in relation to the centroids and their proximity between centroids. The objective function assumes that each component (representing a structural behavior) from the training matrix introduces a quasi-circular cluster of observations, allowing the damage detection in the presence of operational and environmental variability and when damage introduces new orthogonal components. In addition, to evaluate the data dispersion around each centroid, the density of the observations in the clusters is also considered.

Therefore, the first term of the objective function takes the summation of each distance among the centroids (C_i and C_j),

$$\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \mathbf{G}_1(\|C_i - C_j\|), \tag{2.7}$$

where $\mathbf{G}_1$ is a nonlinear penalization function defined as

$$\mathbf{G}_1(d_1) = \frac{1 - e^{-d_1}}{e^{-d_1}}. \tag{2.8}$$

As Equation 2.8 increases positively for all $d_1 > 0$, one easily concludes that when $\mathbf{G}_1$ increases, the distances between centroids also increase. The second term is defined as

$$\sum_{k=1}^{K} \mathbf{G}_2 \left(\sum_{\forall x \in C_k} \|C_k - x\| \right), \tag{2.9}$$

where

$$\mathbf{G}_2(d_2) = \frac{1}{e^{2d_2}}. \tag{2.10}$$

In this case, Equation 2.10 decreases as the summation of the norms increases for all $d_2 > 0$; this function aims to achieve a balance between maximization (Equation 2.7) and minimization (Equation 2.9), as shown in 2. Therefore, the objective function is defined by the combination of those two terms, regularized by the number of components and the standard deviation of all distances between centroids,

$$\mathcal{F}\left(x, \mathbf{C}K, \sigma\right) = \frac{1}{\sigma K} \left(\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \mathbf{G}_1(\|C_i - C_j\|) + \sum_{k=1}^{K} \mathbf{G}_2 \left(\sum_{\forall x \in C_k} \|C_k - x\| \right) \right), \tag{2.11}$$

rewriting equation in vectorized form is obtained

$$\mathcal{F}\left(x, \mathbf{C}K, \sigma\right) = \left[(\sigma K)^{-1} \left(\sum_{i=1}^{K-1} \mathbf{L}_1(C_i) + \sum_{j=1}^{K} \mathbf{L}_2(C_j) \right) \right] \; ; \; C \in \mathbb{R}^K, \tag{2.12}$$

where $\mathbf{L}_1$ and $\mathbf{L}_2$ are defined as

$$\mathbf{L}_1(C_i) = \left[\frac{1 - e^{-\|C_i - C_{i+k}\|}}{e^{-\|C_i - C_{i+k}\|}} \right] . \; \hat{x}_i \; ; \; 1 \leq i < k \leq K, \tag{2.13}$$

$$\mathbf{L}_2(C_j) = \left[e^{-2\|C_j - x_s\|} \right]^T . \; \hat{x}_j \; ; \; 1 \leq s \leq n \; , \tag{2.14}$$

where T is the transpose resulting column vector, $\hat{x}_i = [1, ..., 1]_{1 \times K-i-1}$ and $\hat{x}_j = [1, ..., 1]_{p \times 1}$ are used in the inner product, $C_{i+k} = [a_{ij}]_{K \times m}$ is a matrix composed of all feature vectors from centroids with index greater than i, $x_j = \forall x \in C_j$ is a set of observations assigned to the component j. The maximization of $\mathcal{F}\left(\cdot\right)$ provides the optimal clustering solution.

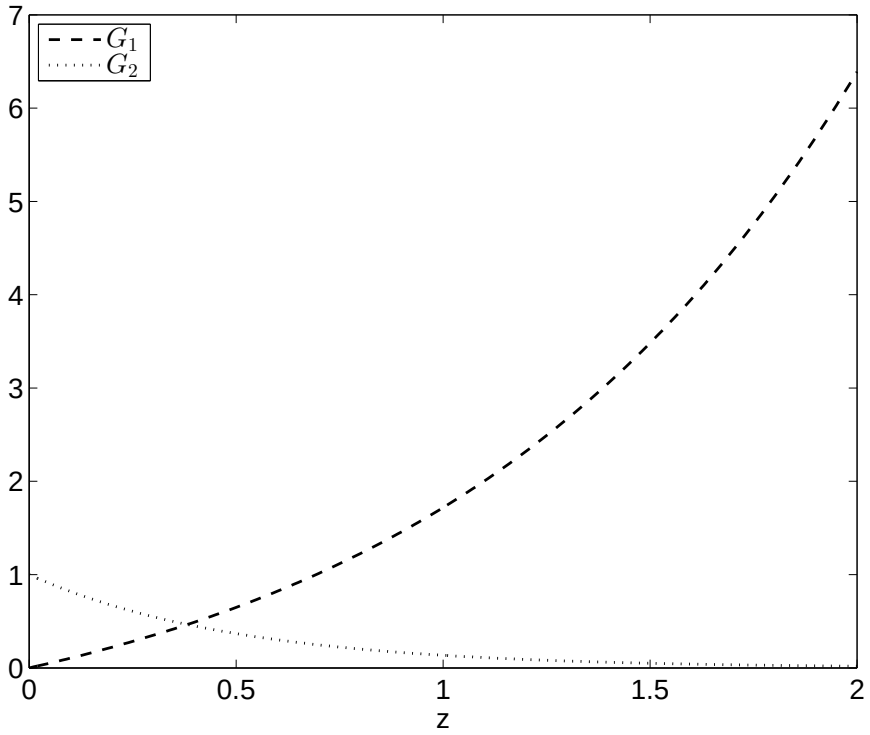


Figure 2: Bonus functions from spacial disposition and data dispersion term.

2.5 Modeling of structural components

In this work the capabilities of the GAs are combined with the expression of characterization of components to discriminate different configurations of clustering for a given number fixed of centroids K in a $\mathbb{R}^m$ dimensional space. However, in order to infer automatically the number of components in the interval $1 \leq K \leq K_{max}$ it is necessary a new method to analyse and model the real data groups through detection of redundant components aiming to agglutinate centroids positioned in a same set of observations that correspond a real cluster. First, the key ideas which support the developed strategy is described, following an asymptotic analysis of computation that shows the good computational efficiency of the method.

2.5.1 Concentric hypersphere method (CHM)

The CHM algorithm consists in analyse sequentially a list of centroids evaluating the boundary regions that limit each cluster. Divided in three steps the algorithm starts dislocating all centroids to the center of its respective component indicated by the means of samples assigned to the centroid. Following, each component is evaluated singly giving start to the *linear inflation* phase where multiple hyperspheres sharing the same center are positioned on the centroid being evaluated. In the last step is carried out the *components reduction* that agglutinate all centroids located in the same component. Each step is described in details below.

Step 1: Displacement in the feature space. The idea consists in dislocate all centroids to the locals with greater sample density from each component, in other words, its respective geometric centers. From the average data of each cluster the centroids are dislocated to these positions that pass to be the new centers of each distribution.

Step 2: Linear inflating of concentric hyperspheres. The second step uses an approach based on multiple overlapping of hyperspheres that shares the same center corresponding to the centroid being analysed. Passing through each centroid of a candidate solution a hypersphere is positioned on the centroid with an initial radius given by

$$r_0 = \log_{10}\left(\|C_i - x\| + 1\right) \mid x = \max(x) \in C_i \ , \tag{2.15}$$

where C_i is the centroid of the cluster i and x is the farthest point of C_i such that $\|C_i - x\|$ is the radius of the component centered in C_i. From this step, multiple hyperspheres are overlaid on each other. However, the radius length grows in the form of an arithmetic progression with common difference equal to r_0. The criteria for create new hyperspheres is the positive variation of the sample distribution density between each inflation and it is given by σ^{-2}, otherwise the process is stopped. It means that for every new hypersphere the density and correlation of the data inside the hypersphere are evaluated. While the density variation of the actual and the last hypersphere is positive the linear inflation continue, otherwise this step is interrupted.

Step 3: Components agglutination. In the last step if there is more than one centroid inside the last hypersphere all of them are assimilated and used to create an unique centroid of greater representation being on the average of these centroids. Otherwise, only the pivot centroid is within the last hypersphere which indicates that it is on the geometric center of a real component, thereby the agglutination of the centroids is not performed.

The process is summarized by Figure 3 that presents an example of the method applied to a scenario of three components with a candidate solution having 5 centroids. Initially, in the Figure 3a the centroids are dislocated to its clusters as indicated in the step 1. Note that in the figures 3b and 3c four centroids are agglutinated and form two new components, once they are under a same cluster. On other hand in Figure 3d only one centroid is located under a component, so none procedure is accomplished after the second step stops.

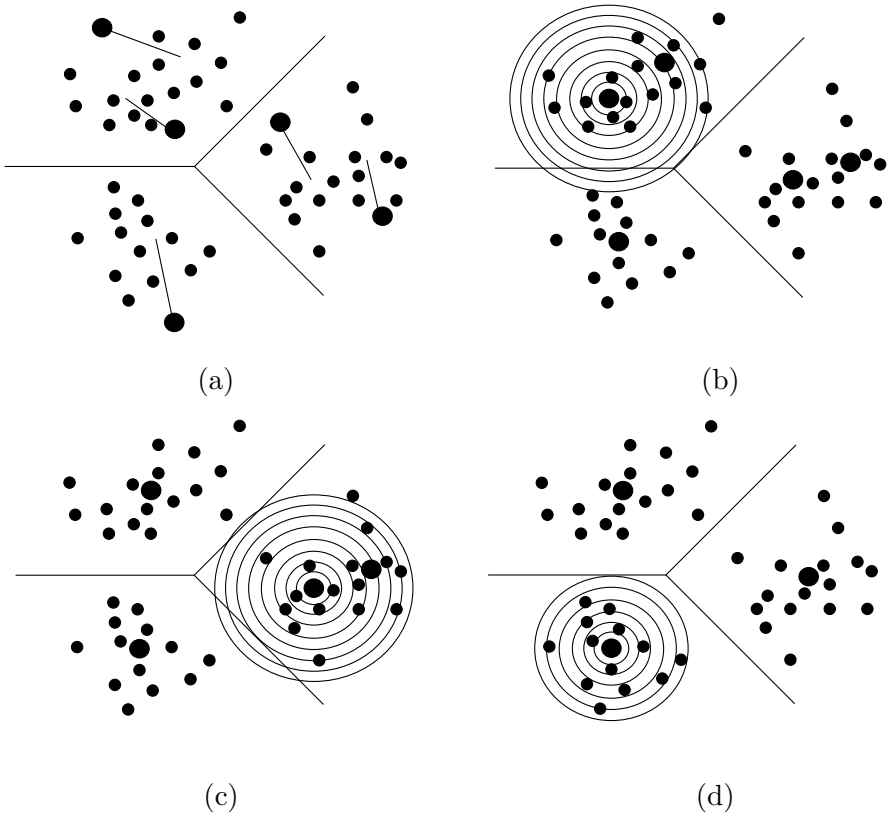


Figure 3: Concentric hyperspheres method using linear inflation. In (a) the algorithm dislocate all the centroids to the geometric center of the their group points, in (b) two centroids under a unique component are assimilated creating only one centroid as the most representative of the cluster as done in (c) for the second cluster. Finally, in (d) the algorithm is applied to a centroid positioned in the center of a real cluster, in this case no centroids are assimilated.

The steps of the CH algorithm are summarized in Algorithm 1. Initially, it identifies the cluster in which each observation belongs and moves the centroids to the mean of their observations. Then, a hypersphere is built on a pivot centroid, by inflation until the density between two consecutive hyperspheres decreases. Finally, the agglutination of all centroids is performed within the last hypersphere, by replacing these centroids by their mean. The process is repeated until convergence, i.e., the solution is composed by only one centroid or there is no centroid agglutination in any iteration.

Input : A centroids list $\mathbf{C}$ such that $\mathbf{C} \in \mathbb{R}^K$
A training data matrix x such that $x \in \mathbb{R}^{n \times m}$
Output: The matrix $\mathbf{C}$ after components reduction

if $|\mathbf{C}| \neq 1$ **then**
  $i \leftarrow 1$
  `createClusters`$(\mathbf{C}, x, n)$
  `dislocate`$(\mathbf{C}, x, n)$
  while $i \leq |\ \mathbf{C}\ | \ AND \ |\ \mathbf{C}\ | > 1$ **do**
    $radius_0 \leftarrow$ `calcRadius`(C_i, x, n)
    $radius, density_0, density_1, delta_1 \leftarrow 0$
    $delta_0 \leftarrow 1$
    while $delta_0 > delta_1$ **do**
      $radius \leftarrow radius + radius_0$
      $\mathbf{H} \leftarrow$ `calcHypersphere`$(\mathbf{C}, i, x, n, radius)$
      $density_0 \leftarrow density_1$
      $density_1 \leftarrow$ `calcDensity`$(\mathbf{H})$
      $delta_0 \leftarrow delta_1$
      $delta_1 \leftarrow |density_0 - density_1|$
    end while
    $j \leftarrow$ `reduce`$(\mathbf{C}, \mathbf{H})$
    if $j > 0$ **then**
      $i \leftarrow 1$
      `createClusters`$(\mathbf{C}, x, n)$
    else
      $i \leftarrow i + 1$
    end if
  end while
end if

Algorithm 1: Concentric hyperspheres algorithm.

2.5.2 Asymptotic properties

Since $x \in \mathbb{R}^{n \times m}$ is composed by training examples and $C \in \mathbb{R}^K$ is composed by K sets of values disposed in the feature space is possible to infer two asymptotic proprieties inherent to the model with greater computational cost. The first property concerns to the maximum number of necessary iterations to the most external loop before convergence occurs.

***Property* 1** *Assuming that C is composed of nonempty components $C_1, C_2, \cdots, C_K \in \mathbb{R}^m$ admitting the same operations such as $C_i \subseteq C$ and x has only one real component to be defined, then between its $K!$ possible permutations there is at least one which makes necessary $\frac{K^2+K-2}{2}$ iterations before algorithm converge.*

Since C admit any one of the $K!$ combinations of its elements, there will be one to keep the centroids distributed by feature space in such way that the algorithm passing

through each element in C always agglutinates only two components in one per iteration making necessary check all the components previously verified searching for new possible divergences. It causes $K+(K-1)+(K-2)+(K-3)+(K-4)+\cdots+2 = \frac{K^2+K-2}{2}$ loops before the algorithm determine which there is only one component defining the cluster form. Thereby, there are K centroids positioned on the K points with greater sample dispersion in relation to the average. The second propriety derives from the first and establish a limit of iterations to the most internal loop defining the number of hyperspheres in a same component.

***Property* 2** *Being the increment value of the hypersphere radius defined by Equation 2.15 and C_i is close to the geometric center of the component, then the maximum number H_y of hyperspheres built before algorithm convergence is given by*

$$H_y \leq \left\lceil \frac{\max(\|C_i - x\|)}{r_0} \right\rceil \tag{2.16}$$

The property 2 becomes trivial to imagine that if a centroid is positioned on the center of a real component (or in its neighbourhood) and the radius increase with a common difference equal to r_0, which is a portion of the initial radius admitted to a component C_i, so a conclude is that the hypersphere radius never is greater than the component radius, since the longer the distance from C_i more sparser are the observations which reduces the sample density compared with the last iteration.

2.5.3 Complexity proof

Based on previously properties, in this section is shown the asymptotic complexity proof. However before to start the analysis is required to introduce some cost informations. For estimate a superior limit, it is necessary associate a maximum value of cost in the execution of an unique instruction. In each simple line, that is of easy computation, is assumed that its maximum cost assume a constant value equal to one, on the other hand the lines in which occur function calls the cost is calculated from some analytical considerations.

Before reviewing lines 5 and 12 of the CHM algorithm, concerning to the two loops that maintain the most part of algorithm complexity, it is needed to analyse the lines with constant cost. The line 3 is responsible for classify each observation as belonging to a cluster that give us a cost equal to $K \times n$. In similar way the line 4 assume a cost equal to the product of the feature space dimension m by the number of training data points n added with the number of centroids K since for dislocate K centroids, it is necessary evaluate n samples in a m-dimensional space.

For compute a component radius, in the worst case, it is necessary evaluate $n-1$ observations which may be attributed to the cluster in analysis. In this case the line 6 result in a computational complexity of $\lceil n-1 \rceil$. The line 14 is responsible for indicate which points are inside of the hypersphere and it is necessary analyse all the n points in the training matrix X, that has a constant complexity equal to n. In a similar manner, for compute the sample density of a hypersphere, the line 16 needs a maximum of $n \times m$ iterations before convergence, since it is necessary analyse in the worst case n points in the m-dimensional space.

The function *reduce*, in the line 20, needs to analyse all the centroids in each iteration with the purpose to define which may be agglutinated, resulting in a complexity equal to $\mid \mathbf{C} \mid$ that corresponds to the cardinality of the set C. In the line 23 a new function call to $createClusters$ is made. However, due to the number of centroids may be changed in each iteration, it is needed to substitute the value K by cardinality of C.

For understand the maximum complexity estimated in the line 12 the Property 2 discussed in the earlier section is needed. The property assumes that the maximum number of hyperspheres built depends of the component radius and the increment value given by Equation 2.15. Therefore, the number of iterations in the line 12, in the worst case, is equal to $\left\lceil \frac{\max(\|C_i - x\|)}{r_0} \right\rceil$. The particularities stipulated in the worst case give the intuition that for a number of centroids K occurs an agglutination made two by two until there is more than one centroid defining the component, in this case after each agglutination is necessary revalidate all the components previously validated. Its complexity is equivalent to an arithmetic progression (AP) with common difference and initial term equal to two and one, respectively.

Being the cardinality of the set C reduced by one per iteration and the counter i positioned in the top of the list in each occurrence of a cardinality reduction of the set C, so a conclude is that the maximum number of iterations realized by the loop in the line 5 is $\left\lceil \frac{K^2+K-2}{2} \right\rceil$ which is the AP formula with common difference and initial term equal to one and two, respectively. From that, it is possible inductively to infer that the lines 20 and 23 obey the same AP stipulated to the most external loop since the both costs depends of the cardinality of C. For that reason, after algorithm converge the lines 20 and 23 were executed $\frac{K^2+K-2}{2}$ and $\frac{n(K^2+K-2)}{2}$ times, respectively.

From these knowledge, it is possible to calculate the algorithm complexity. Adding the cost of all terms and multiplying those inside loops by the maximum number of iterations give us:

$$\mathcal{T}(K,n,m) = \left(\frac{K^2+K-2}{2}\right)\left((n-1)+7+H_y(mn+n+4)+\frac{(n+1)(K^2+K-2)}{2} \right) + nK+nm+K$$

ordering and retiring components of less asymptotic order results in

$$\begin{aligned}\mathcal{T}(K,n,m) &= \left(\frac{nK^2-K^2+7K^2+nmK^2H_y+4K^2H_y+2nK+2nm+2K}{2}\right) \\ &\quad + \left(\frac{nK^4+K^4+2nK^3+2K^3+K^2+4}{4}\right) \\ &\quad - \left(\frac{3nK^2+4nK+4K+4n}{4}\right) \\ &< nmK^2H_y+nK^4+2nK^3+nK^2+K^4+4K^2H_y+2K^3 \\ &\quad + 6K^2+2nm+2nK+K^2+2K+4-3nK^2-4nK-4n-4K \\ &< nmK^2H_y+nK^4+2nK^3+nK^2+K^4+4K^2H_y+2K^3+2nm \\ &\quad + 2nK+6K^2+K^2+2K\end{aligned}$$

Initially, it is possible to imagine that the term of greater order is nK^4, however the increase asymptotic curve of the term nmK^2H_y is greater due to $K \ll m$. Substituting H_y and r_0 is obtained

$$\begin{aligned}\mathcal{T}(K,n,m) &= nmK^2H_y \\ &= nmK^2\left\lceil\frac{\max(\|C_i-x\|)}{r_0}\right\rceil \\ &= nmK^2\left\lceil\frac{\max(\|C_i-x\|)}{\log_{10}(\|C_i-x\|+1)}\right\rceil \\ &\simeq nmK^2\frac{\max(\|C_i-x\|)}{\log_{10}(\|C_i-x\|+1)}\end{aligned}$$

Assuming that $D = \max(\|C_i-x\|) = \max(|C_i-x|)$, then

$$\begin{aligned}\mathcal{T}(K,n,m) &= nmK^2\frac{D}{\log_{10}(D+1)} \\ &= nmK^2\log_{10}\left((D+1)^{D^{-1}}\right)^{-1}\end{aligned}$$

Thus, a conclude is that the algorithm has a computational complexity equal to

$$\mathcal{O}\left(nmK^2 \log_{10}\left((D+1)^{D^{-1}}\right)^{-1}\right),$$

which is a second order polynomial function. An empirical confirmation of the previous algebraic analysis is obtained varying the number of components, computing the complexity function and evaluating the average of the real computational time of the method in several executions. The Figure 4 presents the asymptotic complexity curve in number of instructions against the average of the real computational time curve in nanoseconds. When dealing with asymptotic analysis and observing the curves characteristics is possible to assent that they both follow the same increase term established previously. Furthermore, according to the Big-O concept for different constant values with different machines and resources the complexity function presented always appear as an asymptotic limit for every real time curves generated in other empirical experiments.

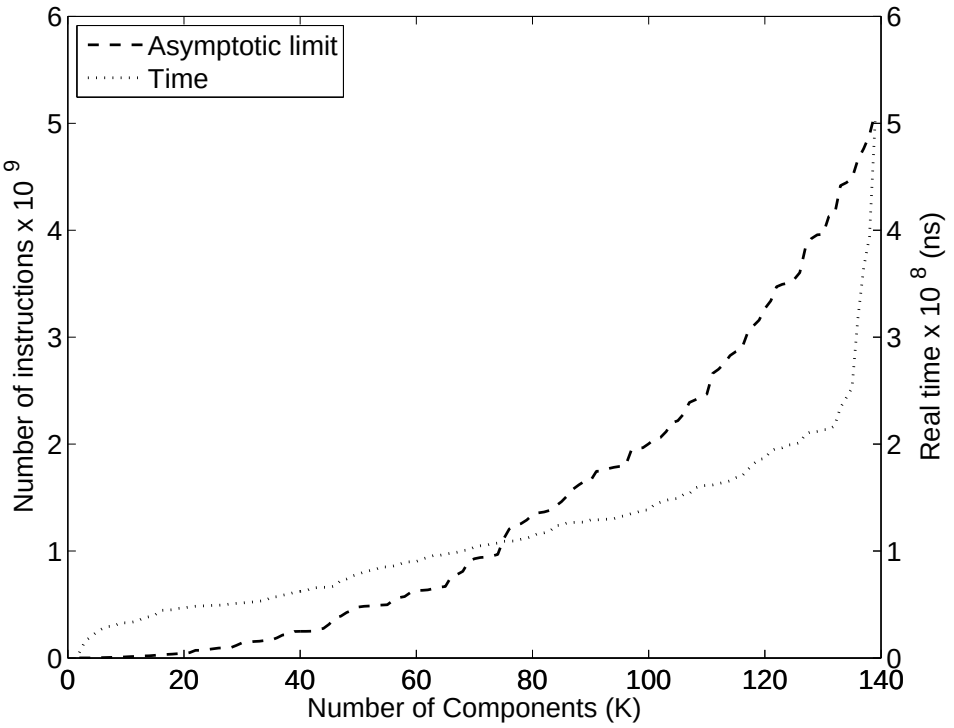


Figure 4: Asymptotic curve against real computational time obtained during several running of the CHM.

2.6 Structural damage classification

In the end of the statistical modeling starts the classification phase. Using the model estimated and using all data base is generated a column vector composed of scalar

values denominated damage indicators (**DI**) that corresponds to the level of damage in all feature vectors. The classification is performed through creation of a global threshold based on a cut off in the **DI** vector using a certain level of significance on the data collected in undamaged conditions. It is expected that in the case of a satisfactory statistical modeling may be possible to identify anomalies in damaged observations even under environmental and operational influences.

To generate the **DI** vector in this work is used a method known as distributed damage indicators (FIGUEIREDO et al., 2014a). First, for a feature vector x_i the euclidean distance for all centroids is computed, then the $\mathbf{DI}(i)$ must have to be the less distance value computed. In this work the threshold is defined using a cut off based on 95% over the training data, in which is expected that the algorithm miss hit less than 5% of all undamaged data used in test phase. The process of generating a **DI** is summarized by applying the following equation

$$\mathbf{DI}(i) = \min\left(\|x_i - C_1\|, \|x_i - C_2\|, \cdots, \|x_i - C_K\|\right), \tag{2.17}$$

where $C_1, C_2, ..., C_K$ are the means of K different components.

2.7 Summary of the GADBA-based approach

Many variants of genetic operators are available in literature. However, the proposed approach aims to reach satisfactory results by keeping its structure as simple as possible. A general schematic of the GADBA-based approach is summarized in Algorithm 2.

As each individual in the population represents a candidate solution, the final result is the one with best fitness provided by the objective function. In the start of the process, the CHM algorithm is performed on all individuals in the population, and their associated parameters are updated at iteration $t = 0$. Then, the objective function is determined for each updated individual. Genetic operators are applied until convergence, i.e., when the value given by the objective function does not change significantly for ten generations, providing the best set of centroids for the clustering problem. Finally, the CHM algorithm is used to refine the best achieved model and the damage indicators are estimated by applying Equation 2.17. $\mathbf{P}(t)$ denotes a population set of size $|\mathbf{P}|$ at generation t and $\mathbf{P}'(t)$ is the resulting population after recombination. $\mathbf{P}''(t)$ is the resulting set of union $\mathbf{P}(t) \cup \mathbf{P}'(t)$ with size $2|\mathbf{P}|$. The initial population $\mathbf{P}(t = 0)$ is generated randomly.

1 $t = 0$
2 Generate population P(t)
3 **while** *convergence is not reach* **do**
4 Apply `CHM(`$P(t)$`)`
5 `Evaluate(`$P(t)$`)`
6 Perform $P'(t)$ = `Recombine(`$P(t)$`)`
7 Perform $P''(t)$ = `Mutate(`$P'(t)$`)`
8 `Evaluate(`$P''(t)$`)`
9 Perform $P(t)$ = `Select(`$P(t) \cup P''(t)$`)`
10 $t = t + 1$
11 **end while**
12 Select P_{max} = `max(`$P(t)$*.fitness*`)`
13 Apply `CHM(`P_{max}`)`

Algorithm 2: GADBA algorithm summary.

3 Test bed structures and data

3.1 Z-24 Bridge

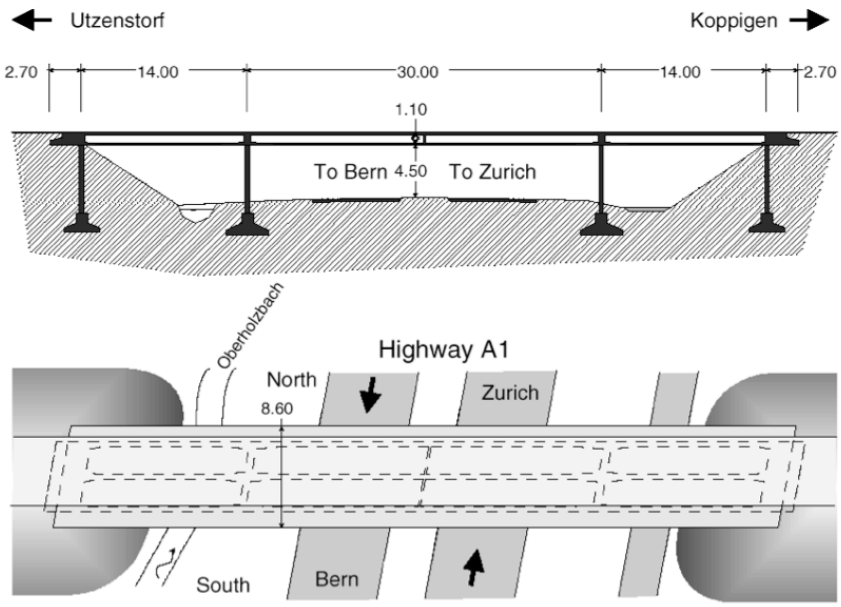
The Z-24 Bridge was a standard concrete box girder bridge connecting the cities of Bern and Zurich. This bridge was composed of a main span of 30m and two side spans of 14m each (see Figure 5a). The bridge demolition occurs from 4 August, 1997 to 10 September, 1998 and during this period was carried out a monitoring system that extracted vibration measurements under environmental and operational influences in order to provide a practicable tool for test and validate new schemes and monitoring solutions. Approximately in the last month of the observation period (from 4th August to 10th September, 1998) damage scenarios was artificially introduced in the system aiming to perform statistical modeling for damage detection. The monitoring system was composed of eight accelerometers that capture mechanical vibrations during eleven minutes by hour.

The damage scenarios introduced progressively during one month period before demolition of the bridge providing a real data sets of damaged features. Some scenarios introduced were concrete spalling, settlement, anchor head failure and tendon rupture. A summary of the tests realized in the structure is shown in Table 1.

3.1.1 Z-24 Bridge data sets

Aiming to extract the principal natural frequencies a method based on covariance-driven stochastic subspace identification was used (PEETERS; ROECK, 1999) on acceleration time series from observations collected hourly from 11th of November 1997 to 10th of September 1998 resulting in four natural frequencies and an amount of 3932 observations, where the first 3462 observations are correlated to undamaged conditions and the last 470 correspond to damaged conditions progressively introduced in the system. It is important to detach that the bridge was intensively influenced by thermal variations caused by freezing effects. Observations performed in the interval 11 of November 1997 to 3 of August 1998 (1-3462 observations) are referred as baseline or normal condition due to the existence only of undamaged data under environmental and operational variability. However, observations accomplished in the interval 4 of August to 10 of September 1998 are related to damaged conditions. Figure 6 shows the first four natural frequencies obtained and partitioning according to its respective structural state.

The use of only undamaged observations in the statistical modeling highlights the



(a)

(b)

(c)

Figure 5: Z-24 Bridge scheme (on the top) and picture (on the lower right) as well as damage scenarios of anchor head failure and tendon rupture (on the lower left and right, respectively).

Table 1: Structural damage scenarios introduced progressively (details in (FIGUEIREDO et al., 2014b)).

Date	Description
04-08-1998	Reference measurement I (before any damage scenario)
09-08-1998	After installation of the settlement system
10-08-1998	Pier settlement = 2 cm
12-08-1998	Pier settlement = 4 cm
17-08-1998	Pier settlement = 5 cm
18-08-1998	Pier settlement = 9.5 cm
19-08-1998	Foundation tilt
20-08-1998	Reference measurement II (after removal of the settlement system)
25-08-1998	Spalling of concrete (12 m^2)
26-08-1998	Spalling of concrete (24 m^2)
27-08-1998	Landslide at abutment
31-08-1998	Concrete hinge failure
02-09-1998	Anchor head failure I
03-09-1998	Anchor head failure II
07-09-1998	Tendon rupture I
08-09-1998	Tendon rupture II
09-09-1998	Tendon rupture III

unsupervised characteristic of the algorithm proposed, next the training and test matrix are defined. The training matrix permits the algorithm to learn the underlying distribution and outline environmental and operational variability. It is composed of the four natural frequencies and all undamaged observations, on the other hand test matrix is composed of the same four frequencies and all undamaged observations, however damaged feature vectors are included. This result in a training matrix of $\mathbf{X}^{3116\times4}$ (1-3116 observations) and a test matrix composed by all observations $\mathbf{Z}^{3932\times4}$ (1-3932 observations).

In order to verify the applicability of the proposed approach for long-term monitoring, daily monitoring data measured at 5 a.m. (because of the lower differential temperature on the bridge) from an array of accelerometers are used to extract damage-sensitive features, which yields a feature vector (observation) per day of operation. An automatic modal analysis procedure based on the frequency domain decomposition was developed to extract the natural frequencies. It was verified that the automatic procedure was only able to estimate the first three frequencies with high reliability, yielding a three-dimensional feature vector per day (FIGUEIREDO et al., 2014a). During the feature extraction process, it was observed that the first and the third natural frequencies are strongly correlated (with a correlation coefficient of 0.94), which permits one to perform dimension reduction of the extracted feature vectors from three to two. The first two natural frequencies, along with circles referring the observations below 0°C, are depicted in Figure 7.

The last 38 observations correspond to the damage progressive testing period, which is highlighted, especially in the second frequency, by a clear drop in the magnitude. Note that the damage scenarios are carried out in a sequential manner, which cause cumulative degradation of the bridge. Therefore, in this study, it is assumed that the bridge operates within its undamaged condition (baseline condition), even though under operational and environmental variability, from 11th of November 1997 to 3rd of August 1998 (1-197 observations). On the other hand, the bridge is assumed in its damaged condition from 4th of August to 10th of September 1998 (198-235 observations). The observed jumps in the natural frequencies are related to the asphalt layer in cold periods, which contributes significantly to the stiffness of the bridge. Actually, Peeters et al. (PEETERS J. MAECK, 2001) showed the existence of a bilinear behavior in the natural frequencies for below and above freezing temperature. Finally, one can observe, especially in the first natural frequency, a structural condition recovery at observation #214 related to the removal of the settlement system. Indeed, as reported by Peeters (PEETERS, 1999), during that moment was witnessed concrete cracks closing after removal of that system.

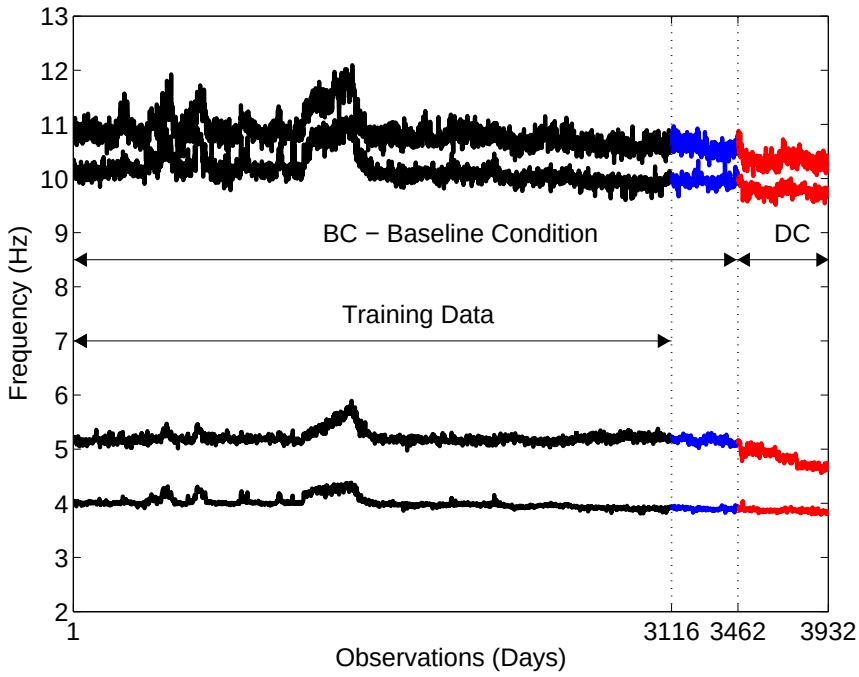


Figure 6: First four natural frequencies of Z-24 Bridge. The observations in the interval 1-3462 are the baseline/undamaged condition (BC) and observations 3463-3932 are related to damaged condition (DC) (FIGUEIREDO et al., 2014a).

In conclusion, the statistical modeling is carried out also taking into account only the first two frequencies and using all 235 observations, resulting in 197 observations from the undamaged condition (1-197 observations) and 38 observations from the damaged condition (198-235 observations). The corresponding training and test matrices are $\mathbf{X}^{197\times2}$ and $\mathbf{Z}^{235\times2}$, respectively. The heterogeneity among observations in a two dimensional space is evidenced in Figure 7, which suggests the existence of components that may be find through latent variables and clustering methods.

Since the training and test matrix are defined for both scenarios, the statistical modeling phase is started. In this step the algorithms GMM, MSD and NLPCA are compared to the GADBA from an unsupervised approach being trained using only undamaged observations. For estimate the number of neurons on the mapping, bottleneck and de-mapping layers of the NLPCA the strategy used in (KRAMER, 1991) is followed. A function expressing a threshold between the number of parameters to be adjusted and the fitting performance of the classifier is presented. The function adopted is the Akaike information criterion (AIC) (KRAMER, 1991) given by:

$$\mathbf{AIC} = \ln(\mathbf{e}) + 2\mathbf{N}_w/\mathbf{N}, \tag{3.1}$$

where $\mathbf{N}_w = (m + f + 1)(M_1 + M_2)$ is the number of network weights such as f is the number of factors retained in the bottleneck layer, M_1 and M_2 are the number of neurons in the mapping and de-mapping layers, respectively, $\mathbf{N} = nm$ the number of inputs of the matrix $\mathbf{X}$ and $\mathbf{e} = E/(2\mathbf{N})$ the mean square error. Applying AIC to the simplified

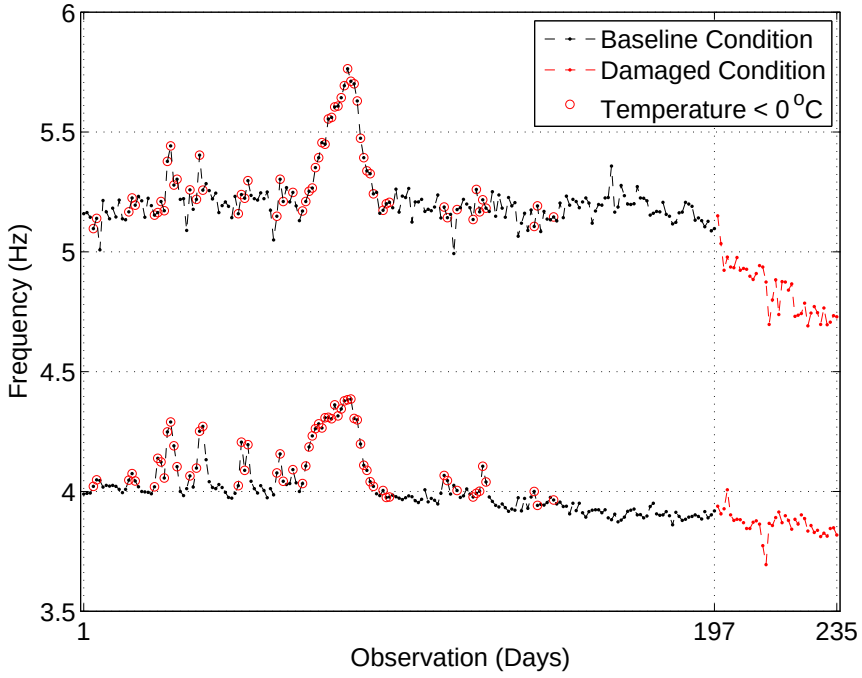


Figure 7: First two natural frequencies estimated from data collected daily at 5 a.m..

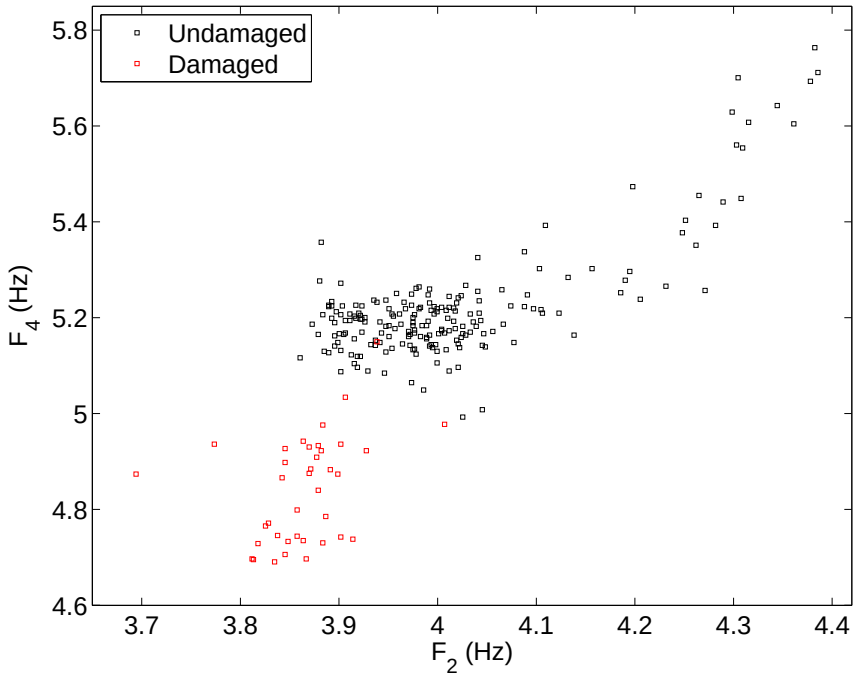


Figure 8: Feature distribution used as a function of the two most relevant natural frequencies.

training matrix $\mathbf{X}^{197\times2}$, a conclude from Figure 9 is that the ideal number of neurons in the mapping layer is four, being usually admitted the same value to de-mapping layer. On the other hand, to the bottleneck layer the criteria indicates the presence of only three factors. The algorithms GMM and MSD were implemented according to (FIGUEIREDO et al., 2014a) and its DIs stored into vectors of 3932 and 235 positions, respectively.

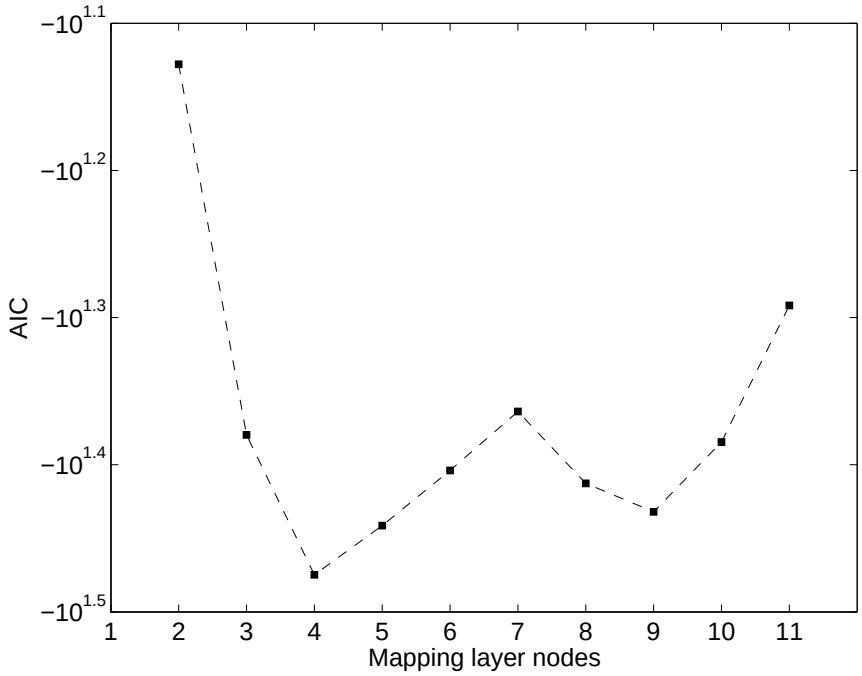


Figure 9: The AIC as a function of the number of nodes in the mapping (and de-mapping) layer for NLPCA using only the first and second natural frequencies extracted at 5 a.m. from Z-24 Bridge.

3.2 Tamar suspension Bridge

The Tamar suspension Bridge (Figure 10) was built in the mid of 1961 connecting the cities of Saltash to Cornwall through A38 road in the United Kingdom. Built on a rocky ground and having originally 335 m of a main span and two side spans of 643 m the bridge received an upgrade consisting in the addiction of a cantilever lane in each side, that provides more two lanes for traffic and a currently support for more than 50.000 cars per day. This changes accomplished in 2001 were made in order to meet a European Union Directive that bridges should be able to carry up more than 40 tonnes.

Aiming to measure environmental and operational variations there are currently three monitoring systems installed on the bridge tracking quantities of wind speed, mechanical post tension and structural temperature variations. The FUGRO monitoring system evaluated during five years the dynamic structural response identifying model parameters. This system worked collecting acceleration measures into intervals of 30 minutes from three biaxial accelerometers located at the main span being possible to analyse and study environmental and operational influences.

(a) (b)

(c)

Figure 10: The Tamar Suspension Bridge viewed from River Tamar margins (Figure 10a and Figure 10c) and cantilever (Figure 10b) perspectives.

3.2.1 Tamar Bridge data sets

The Figure 11 shows the first five natural frequencies obtained during feature extraction phase using covariance-driven stochastic subspace identification (PEETERS; ROECK, 1999) through data collected in the period from 1 of July 2007 to 24 of February 2009 (602 observations).

For the study carried out in the Tamar Bridge is applied a similar approach to described in previous section for Z-24 Bridge. The difference is that there is not damaged observations available, thereby only Type I errors may be generated. For the same reason ROC curves may not be constructed. From a total amount of 602 observations the first 301 are used for statistical modeling and the entire data base is used in the test phase. This results in a training matrix $\mathbf{X}^{301\times5}$ (1-301 observations) and a test matrix $\mathbf{Z}^{602\times5}$ (1-602 observations).

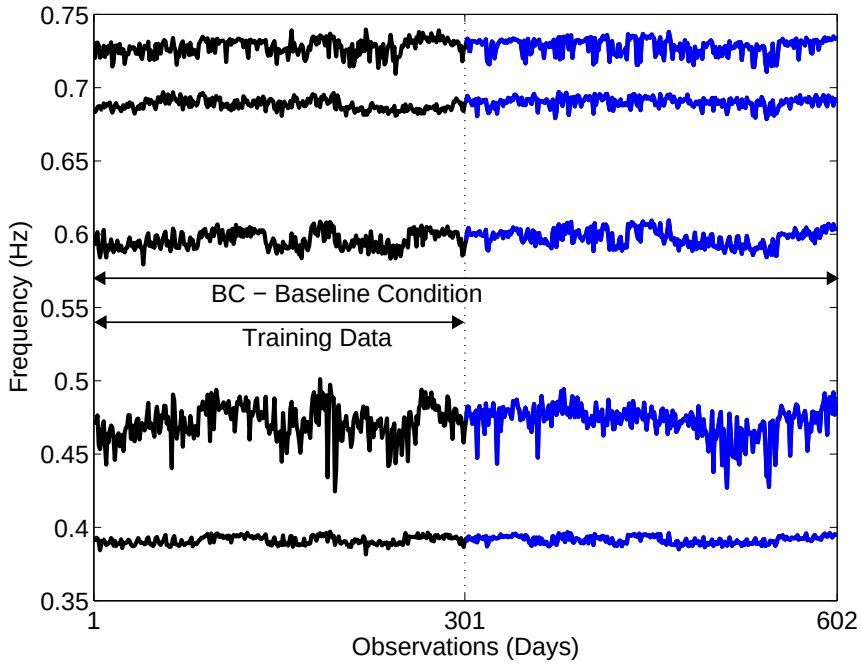


Figure 11: First five natural frequencies obtained in the Tamar Bridge. The observations in the interval 1-301 are used in the statistical modeling while observation 302-602 are used only in the test phase (FIGUEIREDO et al., 2012).

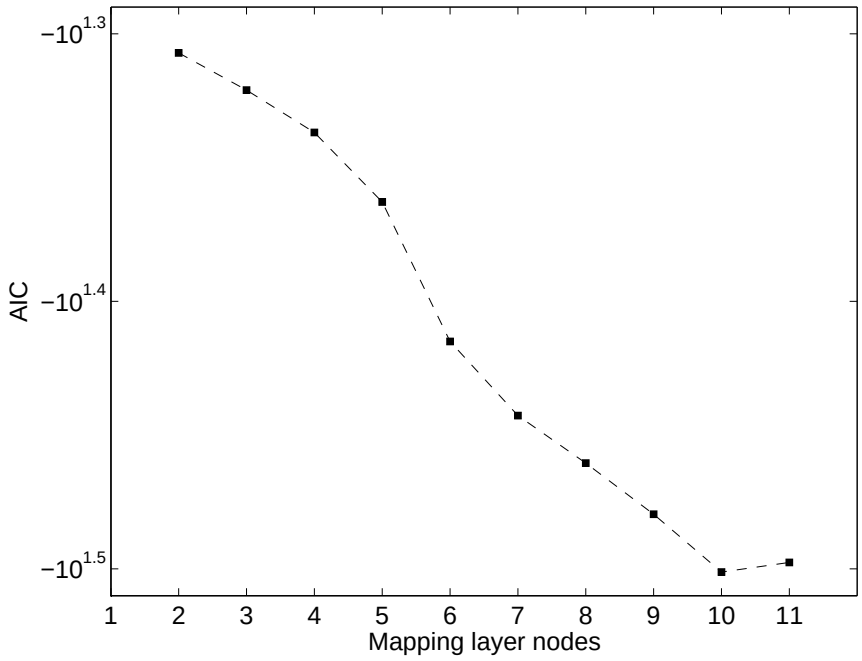


Figure 12: The AIC as a function of the number of nodes in the mapping (and de-mapping) layer for NLPCA for Tamar Bridge data sets.

The same classification approach used in the Z-24 Bridge is employed for the Tamar Bridge. As previous is necessary to compute the appropriate model for apply NLPCA in order to find the number of factors to be retained in the hidden layer. According to Figure 12 the AIC suggests that the number of nodes in mapping and de-mapping layers is between nine and eleven, hence as the models misfit becomes insignificantly is possible to use any one of three configurations. In this work, nine nodes are used in the mapping and de-mapping layers, and eight nodes in the hidden layer, as suggested in (CROSS et al., 2013).

4 Experimental results and analysis

4.1 Z-24 Bridge: full data sets results

The ROC curves are a comfortable and comprehensive manner to analyse the performance of classifiers providing a trade-off between the number of true alarms and false alarms. A perfect classification occurs when a technique hit 100% of observations, in this case it corresponding ROC curve must have to be positioned on the upper left of the graphic. The Figure 13 shows a plot of the ROC curves in linear scale for each algorithm. On the other hand, Figure 14 shows the same curves in log scale aiming to make it easy a posteriori analysis. In the Figure 13 one can verify that none of the algorithms reached a perfect classification with a linear threshold. The MSD had worst performance classifying damaged observations as undamaged (false-negative indication of damage).

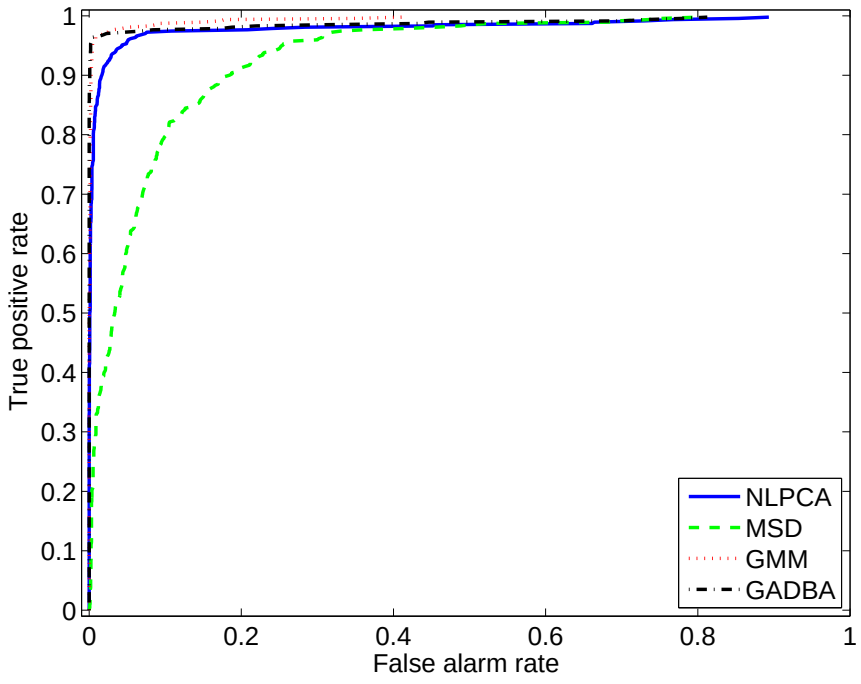


Figure 13: The ROC curves in linear scale for each algorithm.

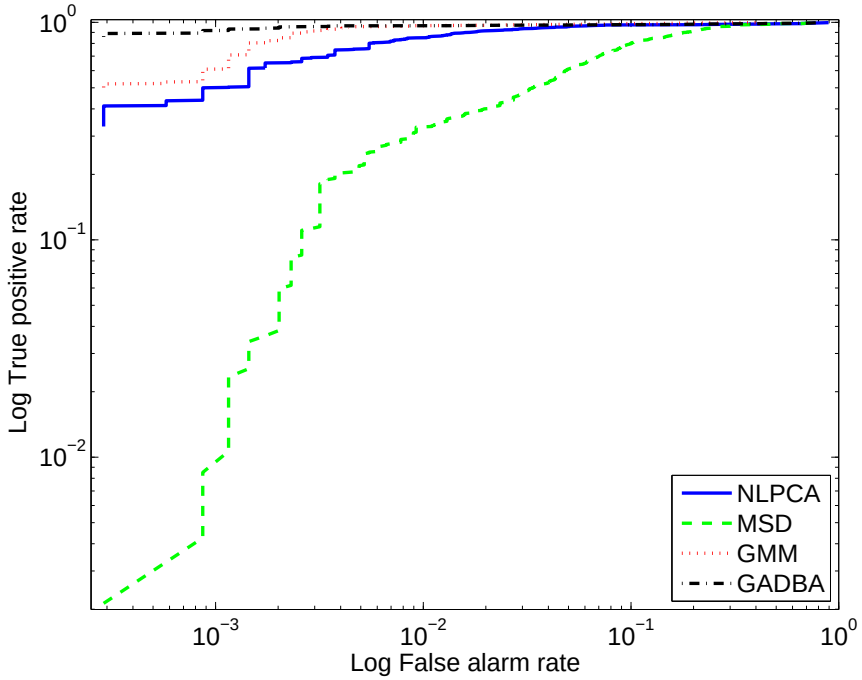


Figure 14: ROC curves in log scale to detect the differences between curves.

Additionally, it is possible to infer that the GADBA and GMM had a similar performance and provides better results when compared to NLPCA in terms of true positive detections. However, analysing Figure 14 is verified that for significance levels at around of 5% (commonly used in real applications) the GADBA demonstrate a better performance if compared to GMM in terms of false positive detections. For low probability rates all techniques demonstrate acceptable levels of errors for true positive rates. For this reason, GADBA demonstrate to be more effective than GMM and NLPCA in the most part of the ROC curves, specially in the task of minimize false indications of damage.

Aiming to quantify the performance of the algorithms using a limiar based on a 95% cut-off value over the training data the Figure 15 plots the DIs from all test matrix $\mathbf{Z}^{3932\times4}$. Only GADBA and GMM shows a monotonic relationship between the amplitude of the DIs and the level of damage in the feature vectors. Additionally, the MSD and NLPCA algorithms shows poor performance for filtering the nonlinear effects in the undamaged observations caused by structural freezing, as highlighted by the concentration of outliers around the middle range of the undamaged observations. Note that generally the temperature variability causes global linear changes in the structural response of bridges; however, in this case, the freezing temperatures cause stiffness changes in the structure, creating a bi-linear response as a function of temperature, which is considered a linear influence that cause a nonlinear effect.

Type I (false-positive indication of damage) and Type II (false-negative indication of damage) errors are common metrics to analyse binary classifiers performance confirming the fact that a false positive indication of damage entails different consequences (for

example, a Type II error may imply in human victims, in the case of an undetected damaged state that can compromise the structure stability.) than a false negative indication of damage. The Table 2 summarizes the amount of Type I and Type II errors for all algorithms.

In general terms, the Table 2 ratifies the results indicated by ROC curves since GADBA performs a better performance trying to balance the number of Type I and Type II errors. The comparison between GADBA and NLPCA reveals a similar classification performance being GADBA relatively better than NLPCA in the task of minimize false-positive indications of damage (5.4% of the NLPCA against 5.57% of the GADBA), however the performance decrease when it tries to detect damage (3.82% of the NLPCA against 2.34% of the GADBA).

Table 2: Number and percentage of Type I and Type II errors for each algorithm.

Algorithm	Type I	Type II	Total
GADBA	193 (5.57%)	11 (2.34%)	204 (5.18%)
GMM	247 (7.13%)	8 (1.7%)	255 (6.5%)
MSD	162 (4.67%)	199 (42.34%)	361 (9.18%)
NLPCA	187 (5.4%)	18 (3.82%)	205 (5.21%)

The GMM obtained best performance as the true indications of damage (1.7%) being practically equivalent to GADBA (2.34%), although when both are compared evaluating the minimization of false indications of damage one can verify is that GADBA demonstrate a better performance than GMM (5.57% of the GADBA against 7.13% of the GMM). The MSD algorithm shows to perform better in terms of minimization of Type I errors (4.67%) and the GMM in terms of minimization of Type II errors (1.70%). However, even though the MSD is performing well at minimizing false indications of damage, it shows an unacceptable performance in terms of Type II errors (42.34%), indicating that it might not be appropriate when some sort of nonlinearities are present in the data sets. Besides that, it failures when trying to describe monotonic relationships between the DIs amplitude and the level of damage, which suggests that as the level of damage increases, it is not guaranteed that the MSD can attenuate the effects of environmental and operational variations. For this reason, is possible to verify nonlinear fluctuations along of the DIs amplitude in a similar manner as occurs in the NLPCA.

Another important observation is referred to the monotonic behaviour presented by GADBA compared to the other algorithms, mainly when compared to GMM that is another algorithm presenting the same behaviour. The Figure 15 highlights a DIs agglomeration when referenced to undamaged feature vectors and a gradual amplitude accent as the damage level increase. However, the DIs generated through GMM shows a greater dis-

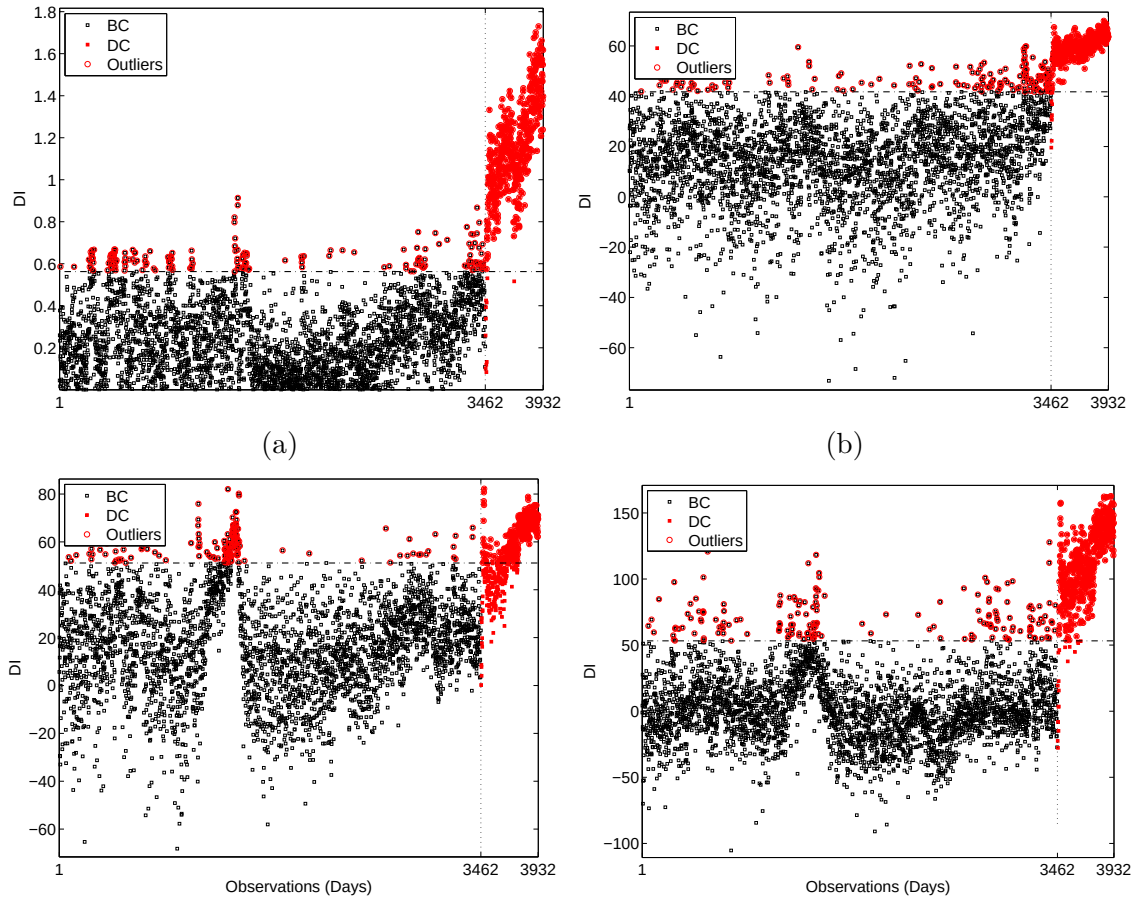


Figure 15: Damage indicators for outlier detection based on a cut off of 95% of confidence using 90% of the data in baseline condition: GADBA (upper left), GMM (upper right), MSD (lower left), and NLPCA (lower right), respectively.

persion of the DIs relative to undamaged feature vectors indicating sensibility to nonlinear perturbations caused by environmental and operational influences.

In summary, one conclude is that GADBA is the algorithm with better balance in terms of Type I and Type II errors and less total amount of errors (5.18%), when compared to the others algorithms. One can infer that not all techniques may be suitable for SHM applications it needs to fit nonlinear effects caused by environmental and operational conditions (FIGUEIREDO et al., 2011). Analysing Figure 15 one can infer is that not all techniques could deal with these effects, indeed only GADBA and GMM were able to lead with such influences caused mainly by structural freezing.

4.2 Z-24 Bridge: simplified data sets results

A second study carried out using only features observed during daily interval with greater thermal differential along the structure relating frequencies one and two is shown next. This differential may be easier explained by the change of Young's modulus (REYNDERS; WURSTEN; ROECK, 2013) caused by temperature effects in the asphalt surface which in contact with a nonlinear system results in nonlinear effects, even if influences are linear (in this case temperature).

The Figure 16 plots the ROC curves for all four algorithms in linear scale. Performing a qualitatively analysis it is possible to observe that for significance levels at around of 5% the algorithms GADBA, GMM, and NLPCA have a similar performance establishing an equivalence relation. On the other hand, the MSD struggles again to performance well for true positive rates. The same ROC curves in log scale in the Figure 17 confirm previously analysis.

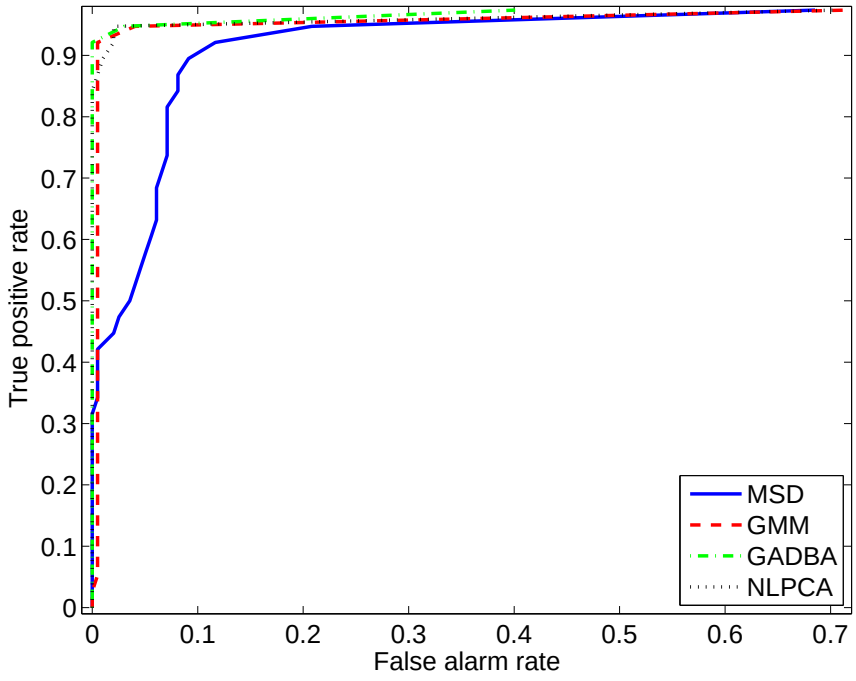


Figure 16: The ROC curves for each algorithm using a simplified dataset with only features one and two.

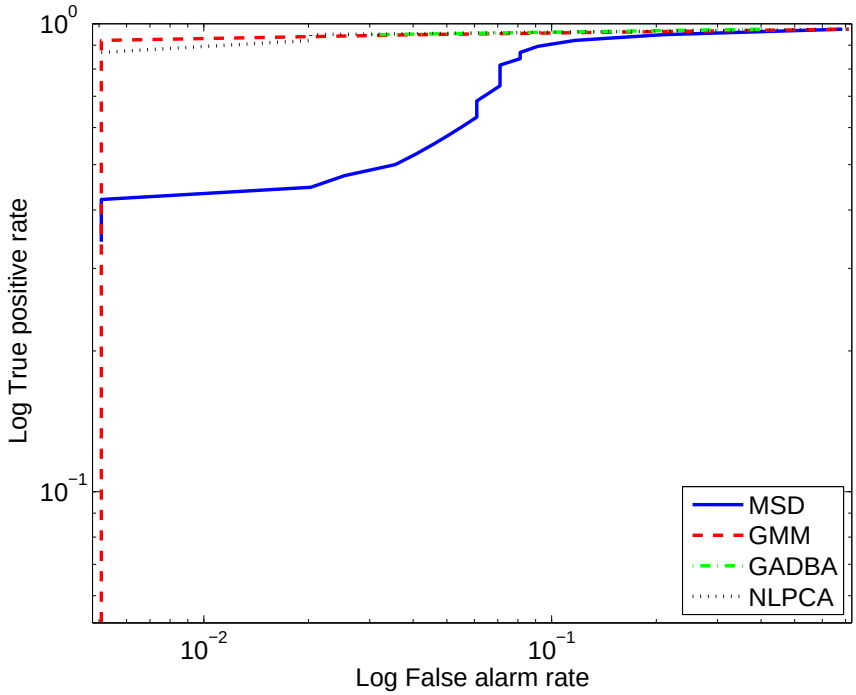


Figure 17: ROC curves in log scale.

Table 3: Number and percentage of Type I and Type II errors for each algorithm using only features two and four.

Algorithm	Type I	Type II	Total
GADBA	10 (5.07%)	1 (2.63%)	11 (4.68%)
GMM	10 (5.07%)	1 (2.63%)	11 (4.68%)
MSD	10 (5.07%)	15 (39.47%)	25 (10.63%)
NLPCA	10 (5.07%)	1 (2.63%)	11 (4.68%)

Additionally, the Figure 18 plots DIs from reduced test matrix $\mathbf{Z}^{235\times2}$ and again shows that only GADBA and GMM outputs a monotonic relation in the DIs amplitude related to damage level in the feature vectors. In the case of MSD and NLPCA, it is possible to note nonlinear distortions in the DIs amplitude caused by freezing effects. The Table 3 summarizes Type I and Type II errors and confirm the similar results expected from the ROC curves. The GADBA, GMM and NLPCA algorithms have the same classification performance, reaching a percentage of 5.97% and 2.63% of Type I and Type II errors, respectively, and a total amount of errors equal to 4.68% of all observations. The MSD obtained a similar result in relation to the amount of Type I errors, however its Type II errors reached more than 39% demonstrating its inefficiency when classifying abnormal conditions.

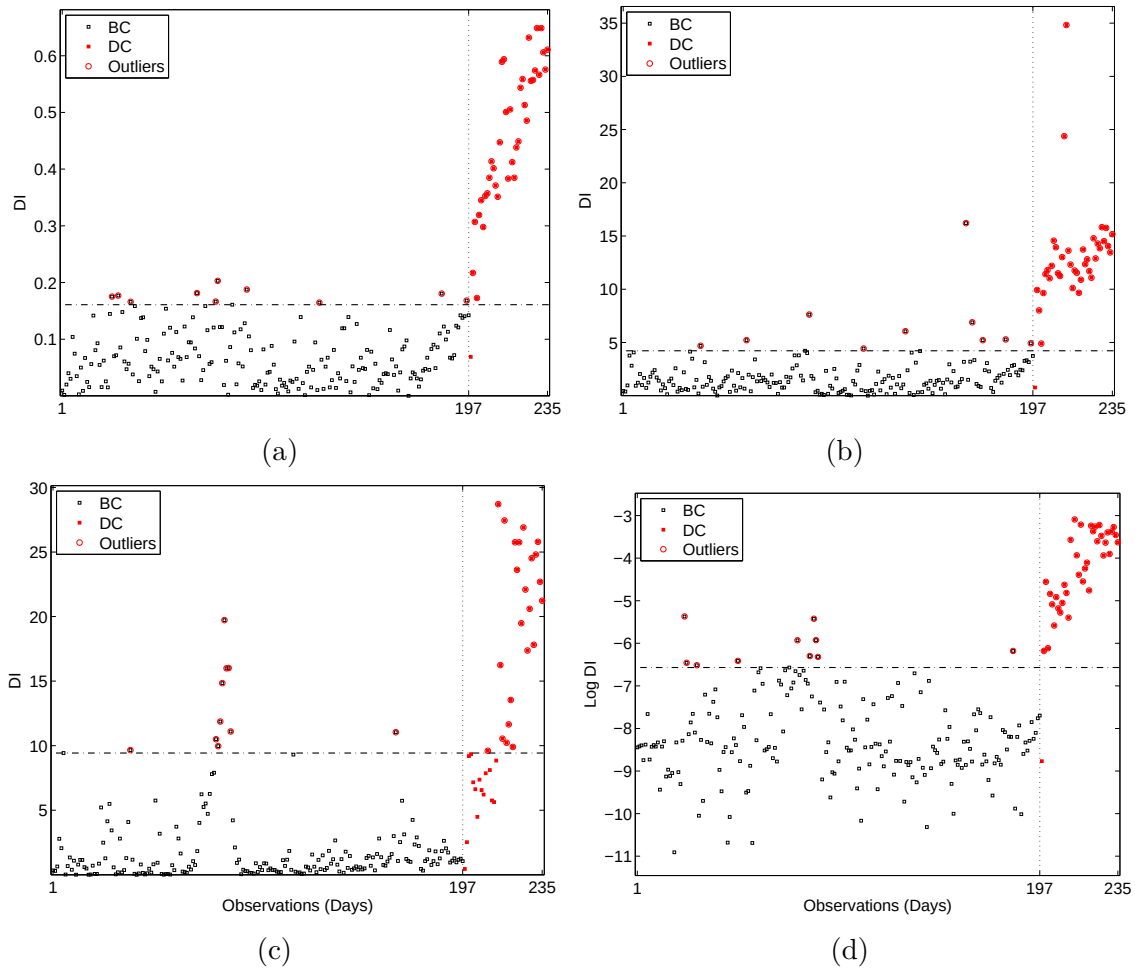


Figure 18: Damage indicators for outlier detection based on a cut off of 95% of confidence on undamaged data: GADBA (upper left), GMM (upper right), MSD (lower left), and NLPCA (lower right), respectively.

The challenge to simulate damage in high capital expenditure civil engineering structures is well-known, namely due to the one-of-a-kind structural type, the cost associated with the simulation of damage in such infrastructure, and due to the unfeasibility to cover all damage scenarios (FIGUEIREDO et al., 2014a; WESTGATE; BROWNJOHN, 2011). Therefore, the unsupervised approaches are often required as long as the existence of data from the undamaged condition is known a priori. Thus, and for real applications, the centroids defined by the CH algorithm are shown in Figure 19b, taking into account only feature vectors from the baseline condition. In this case, three clusters are positioned in close positions as indicated in Table 4. Comparing the results obtained from the CH and EM algorithms (FIGUEIREDO et al., 2014a), one can verify, once again, similarities in the cluster location. However, the CHM algorithm splits the observations under gradual freezing effects into two clusters.

The CHM identified four structural components (Figure 19a) when GADBA ran with all observations. As indicated in Table 5, the first component is centered on the point

(3.97, 5.18) and has around of 69% of all data assigned. In this case, is possible to relate this component to the baseline condition obtained under environmental and operational influences. To the second component centered on the point (4.17,5.28) is assigned around of 10% of the all observations. This component is related to the gradual decreasing of temperature in the asphalt layer, enough to change slightly the elastic characteristics of the structure. In a similar manner, the asphalt layer suffer a gradual wear due to the decreasing in the temperature. The third component centered on (4.31,5.59) attracts 6% of observations and may be related to the structural freezing, which introduces nonlinear effects. The fourth component centered on (3.86, 4.84) is positioned in the lower region of the feature space assuming around of 15% of the entire observations, however only damaged data is assigned to this component (damages inserted artificially according to Table 1).

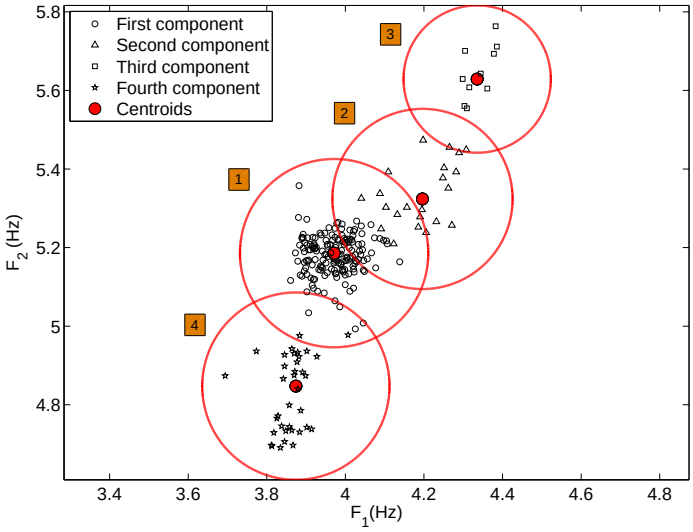
Table 4: Comparison of the parameter estimation using the CHM and EM algorithms on the baseline condition data (1-197) of the Z-24 Bridge (standard errors smaller than $10e-003$).

Algorithm	Description	Cluster 1	Cluster 2	Cluster 3
CHM	Weight (%)	81	12	7
	Mean (Hz)	(3.97, 5.19)	(4.17, 5.29)	(4.30, 5.60)
EM	Weight (%)	81	19	——
	Mean (Hz)	(3.97, 5.19)	(4.22, 5.39)	——

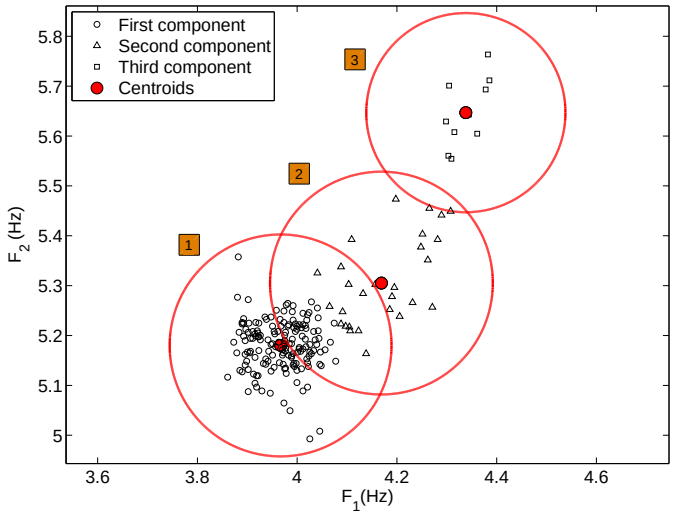
Table 5: Comparison of the parameters estimation using CHM and EM approaches using all data from the simplified dataset being standard errors smaller than $10e-003$.

Algorithm	Description	Cluster 1	Cluster 2	Cluster 3	Cluster4
CHM	Weight (%)	69	10	6	15
	Mean (Hz)	(3.97, 5.18)	(4.17, 5.28)	(4.31, 5.59)	(3.86, 4.84)
EM	Weight (%)	64	21	15	——
	Mean (Hz)	(3.97, 5.19)	(4.16, 5.32)	(3.86, 4.82)	——

The results suggest the possibility to correlate physical states of the structure with a finite and well defined number of main structural components previously unknown even under environmental and operational influences enabling to the logical correspondence between structural states and components identified. (FIGUEIREDO et al., 2014a) shows the existence of this phenomenon which is assigned to the natural grouping of similar observations in certain regions of the feature space. Comparing the results of CHM and EM approaches (FIGUEIREDO et al., 2014a) in the Table 5 one may verify that the similarity of the results obtained (a manner to verify this is looking for the centroids



(a)



(b)

Figure 19: Centroids along with the observations using the data sets from the Z-24 Bridge: (a) all the 235 observations; (b) 1-197 observations corresponding to the baseline condition.

position). However, EM agglutinates the CHM components two and three relating all gradual changes in the temperature to only one component.

4.3 Tamar Bridge data sets results

For overall analysis, the Table 6 shows Type I errors for all four algorithms. For a significance level at around of 95% over the training data, the GADBA offers the best model for indirectly filter environmental and operational influences and fit normal conditions reaching only 27 (8.97%) errors against 68 (22.6%), 33 (10.96%) and 46 (15.3%) Type I errors for GMM, MSD and NLPCA, respectively. The importance of this result derives from the fact of this scenario be close of conditions found in real monitoring scenarios in which there is not excitations or damages artificially introduced. The basic results difference consists in the structural response performed by the Z-24 and Tamar Bridges under different types of influences. It is well-known that in the structural classification process of linear and nonlinear structures is evaluated the structural response when submitted to the both types of influence factors, taking account that influences of high incidence may becomes a structural response be different of the habitual. In the case of the Z-24 Bridge is notorious the nonlinear response under temperature linear effects. On the other hand Tamar Bridge performs a less sensitivity to nonlinear responses when stimulated by linear factors.

The Figure 20 plots DIs for all observations in the test matrix $\mathbf{Z}^{602\times5}$ and confirms previous results of the Table 6 demonstrating a monotonic behaviour in certain uniform distribution, on the other hand DIs of the GMM and NLPCA assumes a significantly dispersed distribution. In addition, for MSD one can verify is that besides its good performance identifying correctly more than 89% of the undamaged observations it fails when filtering nonlinear effects caused by environmental and operational influences assuming two considerable nonlinear changes in DIs amplitude at around of 150 and 450 observations.

As well as verified to Z-24 Bridge the GADBA demonstrate be more effective for damage identification. Although, there is no damaged observations available it is possible to verify that for data not used in the training phase, the algorithm could not identify only 3% of the total amount. On the other hand, most part of misclassification performed by GMM, MSD and NLPCA models occurred in data not used for modeling purposes, reaching 17%, 5% and 10%, respectively. This allows to conclude that for applications wherein the task of minimize the Type II errors is critical then GADBA becomes the most appropriate algorithm.

Table 6: Number and percentage of Type I errors.

Algorithm	Type I
GADBA	27 (8.97%)
GMM	68 (22.6%)
MSD	33 (10.96%)
NLPCA	46 (15.3%)

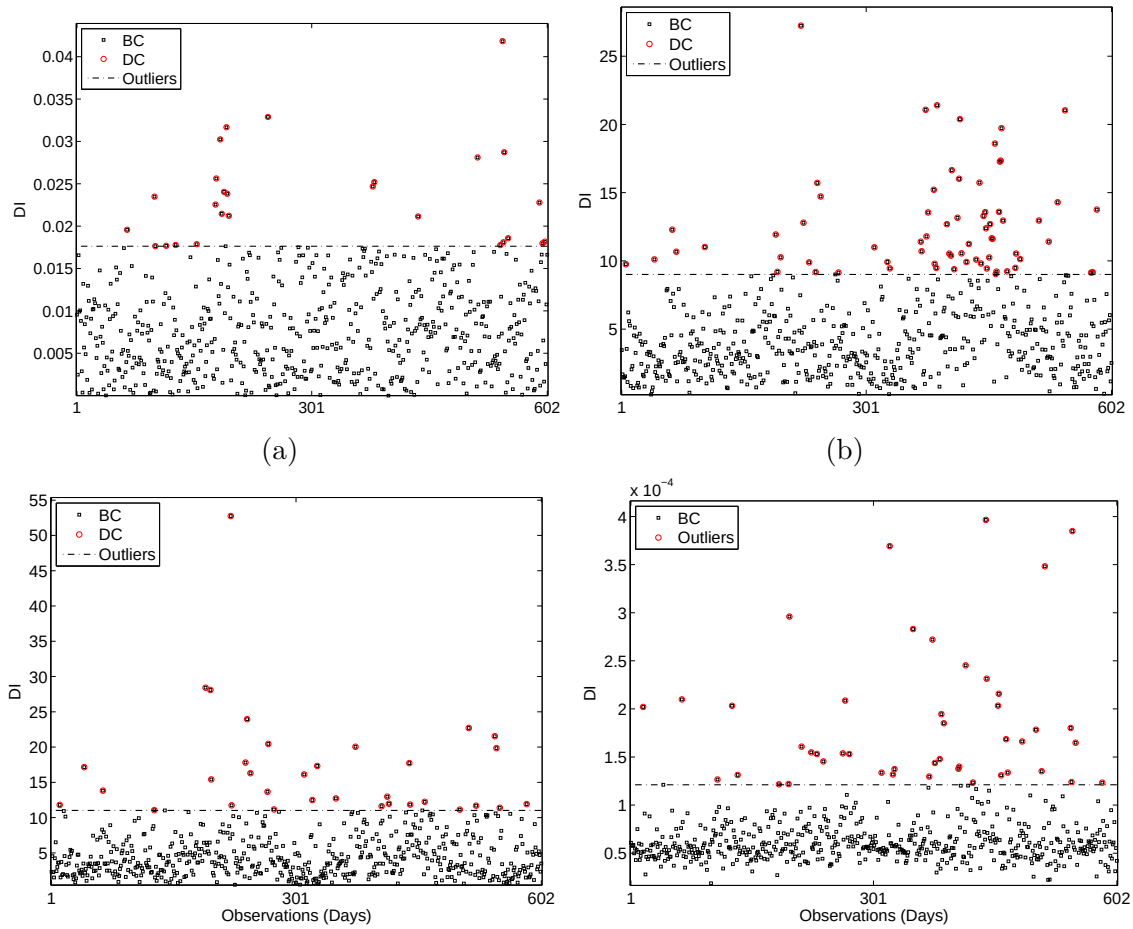


Figure 20: Damage indicators for outlier detection based on a cut off of 95% of confidence using 50% of the data in baseline condition: GADBA (upper left), GMM (upper right), MSD (lower left), and NLPCA (lower right), respectively.

The Figure 21 shows the centroids defined by CHM during model estimation and Table 7 compares the means and distribution weights inferred by CHM ($\mathbf{K} = 3$) and EM ($\mathbf{K} = 4$). The number of centroids is similar to both techniques, however due to the different means position one can infer it is the possible existence of redundant components indicated by EM making GMM results worse when compared to GADBA.

The Figure 21 confirms the results shown in Table 7 in which the main components are found in relative proximity, however for the second feature it is possible to distinguish

all components, suggesting that between the five frequencies the second is which allows the best distinction of structural components. Furthermore, one can figure out that three hyperspheres defined after the execution of linear inflation step has the expected behaviour stopping its inflation close to the boundary of each component. In this situation, it is verified that the boundary regions have at the same time lowest data concentration and most mixture of points belonging to different components. For this reason these areas are found on the intersection of components one and two and in the boundary of hypersphere one and three.

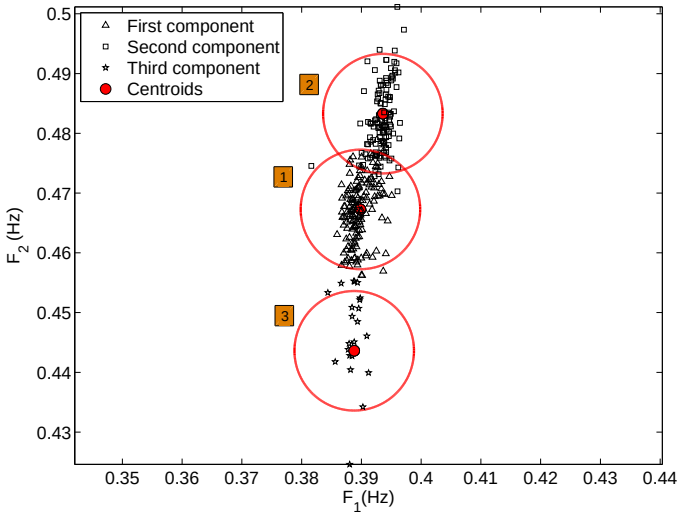


Figure 21: The three main clusters defined by the CH algorithm, with their centroids and corresponding final hyperspheres in the two-dimensional feature space using only the first two frequencies of the Tamar Bridge.

Table 7: Parameters estimation using CHM and EM approaches for Tamar Bridge. Approximation errors smaller than $10e-003$.

Algorithm	Description	Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4
CHM	Weight (%)	——	36	52	12	——
	Mean (Hz)	f_1	0.38	0.39	0.38	——
		f_2	0.46	0.48	0.44	——
		f_3	0.59	0.60	0.59	——
		f_4	0.68	0.68	0.68	——
		f_5	0.72	0.73	0.72	——
EM	Weight (%)	——	29	33	24	15
	Mean (Hz)	f_1	0.39	0.39	0.39	0.39
		f_2	0.47	0.48	0.47	0.47
		f_3	0.59	0.60	0.59	0.60
		f_4	0.69	0.69	0.69	0.69
		f_5	0.72	0.73	0.73	0.72

5 Summary and conclusions

This study presented the performance of an unsupervised and nonparametric clustering-based approach (GADBA) applied to detect damage in bridges, even in the presence of environmental and operational influences. This approach is supported by a novel method (CH) based on spacial geometry and sample density of each cluster, aiming to eliminate redundant clusters, also known as structural components.

The proposed approach was compared with other three techniques extensively studied in the literature (GMM, NLPCA and MSD), through their application on two conceptually different but real-world data sets, from the Z-24 and Tamar Bridges, located in Switzerland and United Kingdom, respectively. The structures were subjected to strong known environmental and operational influences, which cause structural changes due mainly to nonlinear effects of freezing and boundary conditions like thermal expansions and contractions.

In terms of result analysis, as verified on the test bed structures, the GADBA-based approach demonstrates to be: (i) as robust as the GMM-based one to detect the existence of damage; and (ii) potentially more effective to model the baseline condition and to remove the effects of the operational and environmental variability, as suggested by the minimization of false alarms on the data from the Tamar Bridge. In a global perspective is concluded that GADBA has the best classification performance in terms of minimization of Type I and Type II errors, besides that GADBA showed to be the most appropriated method when the main goal is attenuate false indications of damage. One can note is none techniques presented need of a direct measure of variability sources, but only the structural response in terms of temporal series acquired under environmental and operational influences.

In terms of theory formulation, the proposed approach assumes no particular underlying distribution. Additionally, its genetically guided characteristic increases the likelihood to obtain a solution close to the global optimal. On the other hand, the GMM assumes the existence of Gaussian mixture distributions and the EM converges toward a local optimum. Therefore, the GADBA-based approach is conceptually simpler to be deployed in real-world applications and embedded in hardware (e.g., sensor nodes), in situations where it is not possible to make any assumption about the data distribution. Moreover, the CH algorithm provides special capabilities (inflation and observation density analysis) to regularize the number of components and better define clusters, resulting in more accurate models to accomplish data normalization.

Finally, based on the data sets used in this study, both the GADBA- and GMM-based approaches fit the well-known theorem that *there is no free lunch* in which machine learning algorithms are classified in two classes: specialized methods for some category of problems and methods that maintain a reasonable performance in the solution of most part of problems. Thus, the GMM fits in the category of specialized methods that do not generate good results for all type of applications. On the other hand, the GADBA fits the category in which results are acceptable, i.e., it has a superiority in terms of generalization.

In the future it is intended to evaluate the GADBA performance coupled to kernel projection methods and thereby employ new density functions radially symmetric in the CHM (i.e., Epanechnikov kernel).

Bibliography

CHAMBERS, L. D. *The Practical Handbook of Genetic Algorithms: Applications, Second Edition*. 2. ed. [S.l.]: Chapman and HallCRC, 2000. ISBN 1584882409,9781584882404. Cited on page 4.

COWGILL, M.; HARVEY, R.; WATSON, L. A genetic algorithm approach to cluster analysis. *Computers & Mathematics with Applications*, 1999. v. 37, n. 7, p. 99 – 108, 1999. ISSN 0898-1221. Cited on page 5.

CROSS, E. et al. Long-term monitoring and data analysis of the tamar bridge. *Mechanical Systems and Signal Processing*, 2013. v. 35, n. 1–2, p. 16 – 34, 2013. ISSN 0888-3270. Cited on page 30.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 2002. v. 6, n. 2, p. 182–197, Apr 2002. ISSN 1089-778X. Cited 2 times on pages 4 and 8.

ESTES, A.; FRANGOPOL, D. Updating Bridge Reliability Based on Bridge Management Systems Visual Inspection Results. *Journal of Bridge Engineering*, 2003. v. 8, n. 6, p. 374–382, 2003. Cited on page 1.

FARRAR, C. R.; DOEBLING, S. W.; NIX, D. A. Vibration-based structural damage identification. *Philosophical Transactions of the Royal Society: Mathematical, Physical & Engineering Sciences*, 2001. v. 359, n. 1778, p. 131–149, 2001. Cited on page 1.

FARRAR, C. R.; WORDEN, K. An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society: Mathematical, Physical & Engineering Sciences*, 2007. v. 365, n. 1851, p. 303–315, 2007. Cited on page 1.

FARRAR, C. R.; WORDEN, K. *Structural Health Monitoring: A Machine Learning Perspective*. Hoboken NJ, United States: John Wiley & Sons, Inc., 2013. Cited on page 3.

FIGUEIREDO, E.; CROSS, E. Linear approaches to modeling nonlinearities in long-term monitoring of bridges. *Journal of Civil Structural Health Monitoring*, 2013. v. 3, n. 3, p. 187–194, 2013. Cited 3 times on pages 2, 3, and 4.

FIGUEIREDO, E.; MOLDOVAN, I.; MARQUES, M. B. *Condition Assessment of Bridges: Past, Present, and Future - A Complementary Approach*. Portugal: Universidade Católica Editora, 2013. Cited on page 1.

FIGUEIREDO, E. et al. Machine learning algorithms for damage detection under operational and environmental variability. *Structural Health Monitoring*, 2011. v. 10, n. 6, p. 559–572, 2011. Cited 2 times on pages 3 and 34.

FIGUEIREDO, E. et al. Applicability of a markov-chain monte carlo method for damage detection on data from the z-24 and tamar suspension bridges. *Proceedings of the 6th*

European Workshop on Structural Health Monitoring, 2012. p. 747–754, 2012. Cited 2 times on pages VI and 29.

FIGUEIREDO, E. et al. A Bayesian approach based on a Markov-chain Monte Carlo method for damage detection under unknown sources of variability. *Engineering Structures*, 2014. v. 80, n. 0, p. 1–10, 2014. Cited 9 times on pages VI, 2, 4, 19, 23, 24, 26, 37, and 38.

FIGUEIREDO, E. et al. A Bayesian approach based on a Markov-chain Monte Carlo method for damage detection under unknown sources of variability. *Engineering Structures*, 2014. v. 80, n. 0, p. 1–10, 2014. Cited 2 times on pages VIII and 23.

GATTULLI, V.; CHIARAMONTE, L. Condition Assessment by Visual Inspection for a Bridge Management System. *Computer-Aided Civil and Infrastructure Engineering*, 2005. v. 20, n. 2, p. 95–107, 2005. Cited on page 1.

GOLDBERG, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*. 1st. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0201157675. Cited on page 4.

HAKIM, S.; RAZAK, H. A. Modal parameters based structural damage detection using artificial neural networks - a review. *Smart Structures and Systems*, 2014. v. 14, n. 2, p. 159–189, 2014. Cited on page 3.

HALL, L.; OZYURT, I.; BEZDEK, J. Clustering with a genetically optimized approach. *Evolutionary Computation, IEEE Transactions on*, 1999. v. 3, n. 2, p. 103–112, Jul 1999. ISSN 1089-778X. Cited on page 5.

HRUSCHKA, E. et al. A survey of evolutionary algorithms for clustering. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 2009. v. 39, n. 2, p. 133–155, March 2009. ISSN 1094-6977. Cited on page 8.

HSU, T.-Y.; LOH, C.-H. Damage detection accommodating nonlinear environmental effects by nonlinear principal component analysis. *Structural Control and Health Monitoring*, 2010. v. 17, n. 3, p. 338–354, 2010. Cited on page 3.

KRAMER, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 1991. v. 37, n. 2, p. 233–243, 1991. Cited 2 times on pages 3 and 25.

LEE, J. et al. Improving the reliability of a Bridge Management System (BMS) using an ANN-based Backward Prediction Model (BPM). *Automation in Construction*, 2008. v. 17, n. 6, p. 758–772, 2008. Cited on page 1.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.]: University of California Press, 1967. v. 1, p. 281–297. Cited 2 times on pages 5 and 9.

MAULIK, U.; BANDYOPADHYAY, S. Genetic algorithm-based clustering technique. *Pattern Recognition*, 2000. v. 33, n. 9, p. 1455 – 1465, 2000. ISSN 0031-3203. Cited on page 5.

MITCHELL, M. *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1998. ISBN 0262631857. Cited on page 8.

MIYAMOTO, A.; KAWAMURA, K.; NAKAMURA, H. Development of a bridge management system for existing bridges. *Advances in Engineering Software*, 2001. v. 32, n. 10–11, p. 821–833, 2001. Cited on page 1.

NGUYEN, T.; CHAN, T. H.; THAMBIRATNAM, D. P. Controlled Monte Carlo data generation for statistical damage identification employing Mahalanobis squared distance. *Structural Health Monitoring*, 2014. v. 13, n. 4, p. 461–472, 2014. Cited on page 3.

PEETERS, B. *System Identification and Damage Detection in Civil Engineering*. Tese (Doutorado) — Katholieke Universiteit Leuven, 1999. Cited on page 24.

PEETERS, B.; ROECK, G. D. Reference-based stochastic subspace identification for output-only modal analysis. *Mechanical Systems and Signal Processing*, 1999. v. 13, n. 6, p. 855 – 878, 1999. ISSN 0888-3270. Cited 2 times on pages 21 and 28.

PEETERS J. MAECK, G. d. R. B. Vibration-based damage detection in civil engineering: excitation sources and temperature effects. *Smart Materials and Structures*, 2001. v. 10, p. 518–27, 2001. Cited on page 24.

REYNDERS, E.; WURSTEN, G.; ROECK, G. D. Output-only structural health monitoring in changing environmental conditions by means of nonlinear system identification. *Structural Health Monitoring*, 2013. 2013. Cited on page 35.

RUNKLER, T.; KELLER, J. Fuzzy approaches to hard c-means clustering. In: *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*. [S.l.: s.n.], 2012. p. 1–7. ISSN 1098-7584. Cited on page 5.

SOHN, H. Effects of environmental and operational variability on structural health monitoring. *Philosophical Transactions of the Royal Society: Mathematical, Physical & Engineering Sciences*, 2007. v. 365, n. 1851, p. 539–560, 2007. Cited 2 times on pages 1 and 2.

SOHN, H.; FARRAR, C. R. Damage diagnosis using time series analysis of vibration signals. *Smart Materials and Structures*, 2001. v. 10, n. 3, p. 446–451, 2001. Cited on page 1.

SOHN, H.; WORDEN, K.; FARRAR, C. R. Statistical Damage Classification Under Changing Environmental and Operational Conditions. *Journal of Intelligent Material Systems and Structures*, 2002. v. 13, n. 9, p. 561–574, 2002. Cited on page 3.

WEN, Q.; CELEBI, M. Hard versus fuzzy c-means clustering for color quantization. *EURASIP Journal on Advances in Signal Processing*, 2011. v. 2011, n. 1, p. 118, 2011. ISSN 1687-6180. Cited on page 5.

WENZEL, H. *Health Monitoring of Bridges*. United States: John Wiley & Sons, Inc., 2009. Cited on page 1.

WESTGATE, K. Y. K. R. J. .; BROWNJOHN, J. M. W. Environmental effects on a suspension bridge's dynamic response. In: . Leuven, Belgium: [s.n.], 2011. Cited on page 37.

WORDEN, K. et al. The fundamental axioms of structural health monitoring. *Philosophical Transactions of the Royal Society: Mathematical, Physical & Engineering Sciences*, 2007. v. 463, n. 2082, p. 1639–1664, 2007. Cited 2 times on pages 1 and 3.

WORDEN, K.; MANSON, G. The application of machine learning to structural health monitoring. *Philosophical Transactions of the Royal Society: Mathematical, Physical & Engineering Sciences*, 2007. v. 365, n. 1851, p. 515–537, 2007. Cited 2 times on pages 1 and 3.

XIA, Y. et al. Temperature effect on vibration properties of civil structures: a literature review and case studies. *Journal of Civil Structural Health Monitoring*, 2012. v. 2, n. 1, p. 29–46, 2012. Cited on page 1.