Bearing fault prognostics using Rényi entropy based features and Gaussian process models

Pavle Boškoski[a,*], Matej Gašperin[a,c], Dejan Petelin[a,b]

[a] *Jožef Stefan Institute, Department of Systems and Control, Jamova cesta 39, SI-1000 Ljubljana, Slovenia*
[b] *Jožef Stefan International Postgraduate School, Jamova cesta 39, SI-1000 Ljubljana, Slovenia*
[c] *University of West Bohemia, Faculty of Electrical Engineering/RICE, Plzeň, Czech Republic*

Abstract

Standard bearing fault detection features are shown to be ineffective for estimating bearings' remaining useful life (RUL). Addressing this issue, we propose an approach for bearing fault prognostics, which employs Rényi entropy based features to describe the statistical properties of the envelope of the generated vibrations and a set of Gaussian process (GP) models to relate the feature value to RUL. GP models are non-parametric black-box models which search for the relationships among measured data rather than trying to approximate the modelled system by fitting the parameters of the selected basis functions. Bearing's RUL is estimated as a posterior distribution following the Bayes' rule using GP models' output as likelihood distribution. The proposed approach was evaluated on the data set provided for the IEEE PHM 2012 Prognostic Data Challenge.

Keywords: Prognostics, Gaussian process models, wavelet packet transform, Rényi entropy, remaining useful life, Jensen-Rényi divergence

1. Introduction

Several surveys show that bearing faults represent the most common cause for failure of mechanical drives [1, 2]. Therefore, suitable methods for fault detection and prognostics of bearing faults is of significant practical merit. As a result, a plethora of the methods for detection of bearing mechanical faults have been developed. Most of the available methods rely on a well-established feature set, which is based on characteristic bearing fault frequencies linked to specific bearing surface faults [3]. Despite the effectiveness for bearing fault detection, these features are ineffective for estimating bearings' remaining useful life (RUL) since their values are almost constantly zero up until the moment when a surface fault occurs [4]. Addressing the problem of bearing fault prognostics, in this paper we propose a combination of new Rényi entropy features based on the describing the statistical properties of the envelope of bearing's vibrations and Gaussian process (GP) models for calculating bearing's RUL.

The problems of bearing fault prognostics attracted a lot of attention in the past years. Majority of the proposed approaches try to describe the relationship between the defect growth and the time evolution of some statistical characteristic of the generated vibrations like energy, peak-to-peak values, RMS, kurtosis, crest factor, etc. [4–8]. Usually, the selected statistical characteristics are calculated on a specific frequency bands of the generated vibrations and their ratios are used as features for estimating the bearing's RUL.

The effectiveness of these ratios as features for RUL estimation can be explained by the time evolution of the bearing's natural frequency [9–11]. By observing this time evolution, Qiu et al. [12] were able to specify a relation between bearing's natural frequency, the running time and bearing's RUL. Although the

*Corresponding author.
Email addresses: pavle.boskoski@ijs.si (Pavle Boškoski), matej.gasperini@ijs.si (Matej Gašperin), dejan.petelini@ijs.si (Dejan Petelin)

results look promising, the paper lacks clear description how the evolving bearing natural frequency was estimated. Under similar assumption, Ocak et al. [13] model the evolution of the energy of particular wavelet packet nodes using hidden Markov models. The changes in the nonlinear dynamics of the bearing enabled Janjarasjitt et al. [14] to estimate the bearing's RUL by tracking the increase of the dimensional exponents of the generated vibrations.

The time evolution of the natural frequency and the increase of the dimensional exponents indicate that the generated vibrations become more "complex" as the bearings' RUL decreases. Following this idea, in this paper we present a set of features that quantify the statistical complexity of the generated vibrations by employing computationally efficient approach based on wavelet packet transformation.

The concept of signals statistical complexity is readily applied for analysis of EEG signals [15–17]. The idea exploits the fact that the increase of the number of complex (pseudo-)random components present in the observed signal increases its statistical complexity. This paper employs a particular definition of statistical complexity that is a product of Jensen-Rényi divergence (statistical contrast function) and Rényi information entropy. In the context of bearing prognostics, any change in the bearing's surface can be treated as a source of additional signal components with complex dynamics, hence increasing the statistical complexity of the generated vibrations. In this paper we show that the evolution of the (Jensen-)Rényi entropy based indices of the generated vibrations can be related to the bearing's RUL. In addition, the process for calculating the statistical complexity requires no prior information about the operating conditions and no previous knowledge about the physical characteristics of the monitored drive [18, 19].

Based on the values of the Jensen-Rényi divergence and Rényi entropy, the bearing's RUL is estimated using GP models. The GP models are probabilistic, non-parametric models based on the principles of Bayesian probability. They differ from most of the other black-box identification approaches in that they search for relationships among the measured data rather than try to approximate the modelled system by fitting the parameters of the selected basis functions. The output of the GP models is a normal distribution, expressed in terms of the mean and the variance. The mean value represents the most likely output and the variance can be interpreted as a measure of its confidence. The obtained variance, which depends on the amount and the quality of the available identification data, is important information when it comes to distinguishing the GP models from other computational intelligence methods. Due to their properties, the GP models are especially suitable for modelling when data are unreliable, noisy or missing, and therefore have been used in various fields, for instance: biological systems [20, 21], environmental systems [22], chemical engineering [23] and many others. Kocijan and Tanko [24] used GP models for the modelling of time series describing gear health and the prediction of the critical value of harmonic component feature that indicates the wear of gear. In this paper the GP models are used for smoothing noisy features and estimating the RUL based on smoothed features.

The proposed approach for estimating bearing's RUL is depicted in Figure 1. It consists of four main steps. In first step Rényi entropy based features are extracted from input signals. The detailed definition of the selected features and their applicability for bearing prognostics are presented in Sections 2 and 3, and their numerical estimation is presented in Section 4. In the second step these features are smoothed using the GP models. Afterwards, in the third step, RUL is estimated with GP models. The definition and properties of GP models are presented in Section 5. In the last step, presented in Section 6, the RUL distribution $p(RUL)$ is obtained as a posterior distribution by using the output of the GP models as likelihood. The evaluation of the proposed approach, presented in Section 7, is evaluated on the data set provided for the IEEE PHM 2012 Prognostic Challenge [25].

2. Signal complexity

The definitions of the statistical complexity of a signal vary depending of the context, such as data compression, computational algorithms and predictability. In the context of signals, one can define two extremes: periodic and purely random signals. Both cases belong to the class of low complexity signals: the former due to its repetitive pattern and the latter due to its compact statistical description [26, 27]. Consequently, the "complex" signals should be located somewhere in between. Therefore, typical candidates

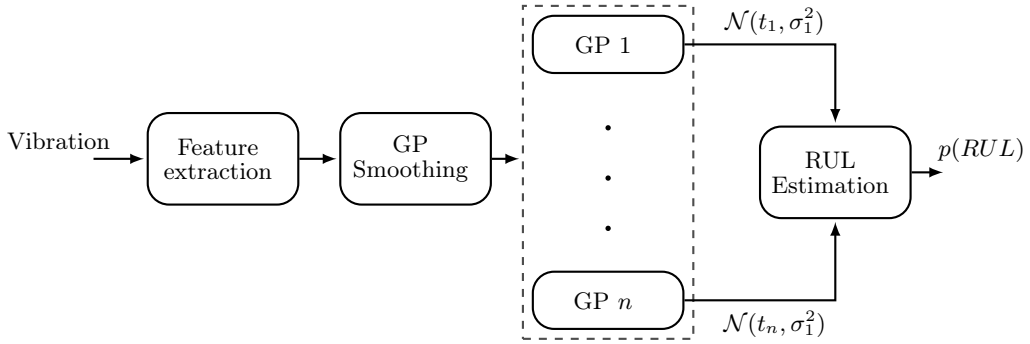


Figure 1: Schematic representation of the prognostics algorithm.

are signals generated by a system with chaotic behaviour. Despite the deterministic nature, such signals contain sufficiently complicated patterns, which are difficult to predict.

Let a random signal be generated by a random source described by its probability distribution $\mathcal{P}$. In this paper the statistical complexity $\mathscr{C}(\mathcal{P})$ is assessed through the information carried by the observed signal generated by the source [16, 28]. The statistical complexity provides a link between the entropy of the source $H(\mathcal{P})$ and the "distance" $D(\mathcal{P}, \mathcal{P}_e)$ between the probability distribution $\mathcal{P}$ and the uniform distribution $\mathcal{P}_e$. Before defining the statistical complexity $\mathscr{C}(\mathcal{P})$, one has to revisit the basic concepts of entropy and "distance" between two probability distributions.

2.1. Concepts of entropy

The concept of entropy serves to characterise the probability distribution functions (PDF). For a discrete probability distribution $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$, the simplest definition of entropy is the one according to Shannon:

$$H(\mathcal{P}) = -\sum_{p \in \mathcal{P}} p \ln(p). \tag{1}$$

For a discrete set with cardinality N Shannon entropy can acquire values between 0 and $N \ln N$. A problem with the Shannon entropy is that it is relatively insensitive to the changes in the tails of the distribution. In many cases, faults in the drives affect the tails. Consequently, we adopted an extension of the Shannon entropy in the form of Rényi entropy [29]:

$$H_\alpha(\mathcal{P}) = \frac{1}{1-\alpha} \ln \sum_{p \in \mathcal{P}} p^\alpha(x), \ \alpha \geq 0 \, \alpha \neq 1. \tag{2}$$

Rényi entropy introduces the parameter α, which can be employed in order to manage the sensitivity of the entropy towards particular segments of the probability distribution $\mathcal{P}$.

2.2. Jensen-Rényi divergence

Divergence is a concept which is helpful in expressing the dissimilarities (or "distance") between the distribution functions. The Jensen-Rényi divergence between two distribution functions $\mathcal{P}$ and $\mathcal{Q}$ defined on the same set is [30]:

$$D_\alpha^w(\mathcal{P}, \mathcal{Q}) = H_\alpha \left(w\mathcal{P} + (1-w)\mathcal{Q}\right) - \left\{wH_\alpha(\mathcal{P}) + (1-w)H_\alpha(\mathcal{Q})\right\}, \tag{3}$$

where $w \geq 0$. The values of the exponent α governs the sensitivity of these two quantifiers to particular segments of the PDF, i.e. it specifies the relative importance of small values versus large values of the probability mass [31].

2.3. Statistical complexity and its application for prognostics

The statistical complexity $\mathscr{C}(\mathcal{P})$ of a signal with distribution $\mathcal{P}$ based on (2) and (3) is defined as [16]:

$$\mathscr{C}(\mathcal{P}) = Q_0 D_\alpha^w(\mathcal{P}, \mathcal{P}_e) H_\alpha(\mathcal{P}), \tag{4}$$

where $\mathcal{P}_e$ is the uniform distribution and Q_0 is a normalisation constant so that $Q_0 D_\alpha^w(\mathcal{P}, \mathcal{P}_e) \in [0, 1]$. The product (4) is in accordance with the initial idea that signals with perfect order $H_\alpha(\mathcal{P}) = 0$ and maximal disorder $D_\alpha^w(\mathcal{P}, \mathcal{P}_e) = 0$ have the lowest complexity.

In the context of prognostics it is important to specify the time evolution of the statistical complexity (4). The concept of time is present in (4) indirectly through the entropy $H_\alpha(\mathcal{P})$ by using the fact that the system's entropy increases in time. Therefore, the statistical complexity $\mathscr{C}(\mathcal{P})$ is usually plotted versus the entropy $H_\alpha(\mathcal{P})$ [16]. This plot always covers a specific area depending on the number of bins used for the estimation of the probability $\mathcal{P}$, as shown in Figure 2. The pre-defined shape of the plot outlines the possible time evolution of the signal's complexity. The evolution of the statistical complexity is directly related to the nature of the observed system. Therefore, by trending the evolution of the statistical complexity within the pre-defined area one can perform the prognostics task. For the task of estimating the bearing's RUL, the first step is to analyse how bearing's condition affects the statistical complexity of the generated vibrations.

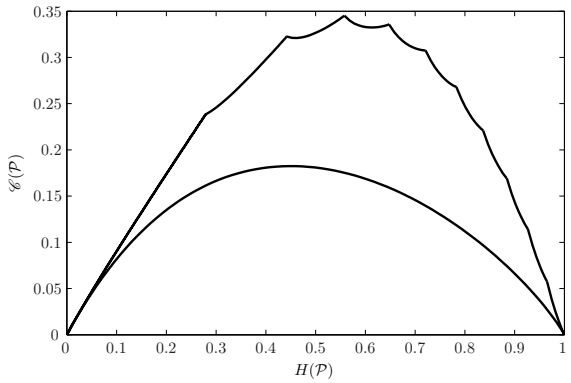


Figure 2: Signal's statistical complexity area.

3. Complexity of bearing vibrations

Healthy bearings produce negligible vibrations. However, in the case of surface damage, vibrations are generated by rolling elements passing across the damaged site on the surface. Each time this happens, impact between the passing ball and the damaged site triggers a system impulse response $s(t)$. The time of occurrence of these impulse responses as well as their amplitudes should be considered as purely random processes. Consequently, the vibrations generated by damaged bearings can be modelled as [32]:

$$y(t) = \sum_{i=-\infty}^{+\infty} A_i s(t - \nu_i) + n(t), \tag{5}$$

where A_i is the impulse of force that excites the entire structure and ν_i is the time of its occurrence. The final component $n(t)$ defines an additive random component that contains all non-modelled vibrations as well as environmental disturbances.

At this point, it should be noted that the impulse response $s(t)$ is influenced by the transmission path from the point of impact to the measurement point [33]. As the position of the damaged spot on the bearing surface rotates the transmission path changes in time. However, the main characteristic of $s(t)$, regardless of its true form, is that it usually resides in the high-frequency range. Since this is the only characteristic relevant for our analysis, we will adopt the model (5) as sufficiently accurate one.

3.1. Evolution of the statistical complexity of the generated bearing vibrations

The main diagnostic information regarding bearing faults are the time moments ν_i when the impulse responses $s(t)$ are excited. Therefore, the usual approach is to analyse the envelope of the generated vibrations. In our case, we look for any changes in the statistical characteristics of the envelope [18].

For healthy bearing, the envelope of the generated vibrations will be without any visible structure due to the lack of impacts $s(t-\nu_i)$. Therefore, the envelope will have low complexity but high entropy. In the context of the plot from Figure 2, such a signal would be positioned in the lower right corner.

The occurrence of a surface fault will introduce some "structure" in the envelope of the generated vibrations. Consequently, its statistical complexity will increase while in the same time the entropy will decrease. In the terminal phase, when the surface faults would became detectable even by the standard bearing fault detection features, the envelope will contain "repetitive" impulses with sufficiently high amplitude. From a statistical point of view, the presence of these impulses makes the signal statistically similar to a deterministic one. Thus, towards the end of the bearing's life, the signal complexity will sharply drop accompanied with a significant decrease in its entropy, hence the final position will be in the lower left corner of the pre-defined area shown in Figure 2. Therefore, the time evolution of the statistical complexity of the envelope of the generated vibrations from the lower right to the lower left point of the pre-defined area can be employed as a feature in the process of estimating RUL.

4. Wavelet based estimation of the statistical complexity of the signal envelope

According to (4), the first step in the computation of the statistical complexity is the estimation of the PDF of the envelope of the generated vibrations. Due to the link between the signal's envelope and its instantaneous power [34], in this approach the underlying PDF is estimated through the energy distribution of the wavelet packet coefficients.

For the computation of the coefficients the so-called wavelet packet transform (WPT) is used [35]. The structure of WPT is described by a binary tree structure, as shown in Figure 3. A wavelet packet tree with depth d_M and nodes (d,n), where $d=\{1,2,\ldots,d_M\}$ represents the depth of the tree and $n=\{1,2,\ldots,2^d\}$ stands for the number of the node at depth d. WPT allows arbitrary partition of the time-frequency plane. The wavelet coefficients in the set of terminal nodes contain all information regarding the analysed signal. The analysis of the envelope is performed by analysing the signal's energy within each terminal node.

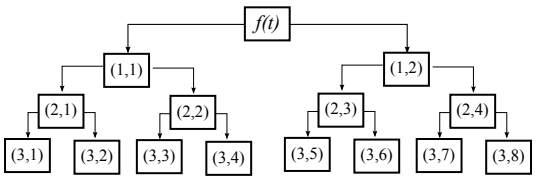


Figure 3: Example of a full WPT tree with depth $d_M=3$.

Each of the n nodes at level d contains N_d wavelet coefficients $W_{d,n,t}$ $t=0,\ldots,N_d-1$, $N_d=2^{-d}N_s$, N_s is the sample length of the signal [36]. Using these coefficients, the portion of the signal's energy $E_{d,n}$ contained within one node (d,n) reads [37]:

$$E_{d,n}=\sum_{t=0}^{N_d-1}\|W_{d,n,t}\|^2. \tag{6}$$

The total signal's energy can be obtained by summing the energy contained within the set of terminal nodes T:

$$E_{tot}=\sum_{\substack{t=0\\ d,n\in T}}^{N_d-1}\|W_{d,n,t}\|^2=\sum_{d,n\in T}E_{d,n}. \tag{7}$$

The set $\mathcal{P}^{d,n}$ expresses the contribution of each wavelet coefficient to the energy of the signal within the terminal node (d, n):

$$\mathcal{P}^{d,n} = \left\{ p_t^{d,n} = \frac{\|W_{d,n,t}\|^2}{E_{d,n}},\ t = 0, \cdots, N_d - 1 \right\}. \tag{8}$$

A similar set can be defined for the contribution of the energy of each terminal node $(d, n) \in T$ in the total energy of the signal E_{tot}:

$$\mathcal{P}^T = \left\{ p_{d,n} = \frac{E_{d,n}}{E_{tot}},\ d, n \in T \right\}. \tag{9}$$

The elements contained in both sets $\mathcal{P}^{d,n}$ and $\mathcal{P}^T$ can be treated as realisation of a random process. Based on these realisations one can estimate the corresponding probability distributions and calculate their entropies and statistical complexity according to relations (2)–(4).

4.1. Condition monitoring based on the statistical characteristics of the sets $\mathcal{P}^{d,n}$ and $\mathcal{P}^T$

The idea of monitoring the condition of a drive is illustrated in Figure 4. At the beginning of the monitoring process the reference condition should be defined by computing the values of $\mathscr{C}(\mathcal{P})$ and the corresponding Rényi entropy for both $\mathcal{P}^{d,n}$ and $\mathcal{P}^T$. In the course of time, the values are calculated on a short segment of signal. If the bearing's condition is normal, no significant difference between the two distributions should exist. A fault in the system can cause changes in the distribution of the particular node at hand, hence altering the corresponding values of $D_\alpha^w(\mathcal{P}, \mathcal{P}_e)$ and $H_\alpha(\mathcal{P})$. This can be used as means to detect and, in some cases, to isolate a fault.

It is important to emphasise that the window length is usually very short and the operating conditions within the node can therefore be assumed constant. If the speed actually varies, the spectral content will also move along the frequency axis. In spite of that, the distribution pattern associated with the WPT will not change much as the shifted harmonics are still within the specific frequency band associated to the particular node. However, if a change in the operating speed is severe enough, it might happen that the frequency content from one node moves to the adjacent node, thus fooling entirely the diagnostic reasoning. In the case of variations in the load, mild variations normally have no significant impact on the frequency distribution pattern. Furthermore, even in the case of significantly increased load, additional sideband components might occur but without any major impact on the energy distribution within a node.

4.2. Running example

To show the applicability of the concept, simulated signals reflecting fault-free run as well as run with bearing fault are shown in Figure 4. The signals are processed by WPT with depth $d_M = 3$. The histograms of $\mathcal{P}^{d,n}$ values for the node $(3, 4)$ are shown at the bottom.

The histogram of the fault free case shows that the majority of the wavelet packet coefficients $W_{3,4,t}$ have zero value. According to (8) we can claim that for the fault free case the node $(3, 4)$ is without any content. The small number of wavelet coefficients that have value greater than zero can be attributed to the simulated Gaussian noise $n(t)$ from (5).

Conversely, presence of bearing fault significantly alters the frequency content in this node. The histogram of $\mathcal{P}^{3,4}$ has completely different shape than the one of the fault free case. By quantifying the shape alterations of the histograms of particular WPT nodes, using the relations (2)–(4), one can estimate the bearing's RUL sufficiently accurate.

5. Gaussian process models

GP models are flexible, probabilistic, non-parametric models. Their modelling properties are reviewed in [38–41]. A Gaussian process is a collection of random variables, which have a joint multivariate Gaussian distribution. Assuming a relationship of the form $y = f(\mathbf{x})$ between input $\mathbf{x}$ and output y, we have $y_1, \ldots, y_N \sim \mathcal{N}(0, \mathbf{K})$, where $\mathbf{K}_{pq} = \mathrm{Cov}(y_p, y_q) = C(\mathbf{x}_p, \mathbf{x}_q)$ gives the covariance between output points

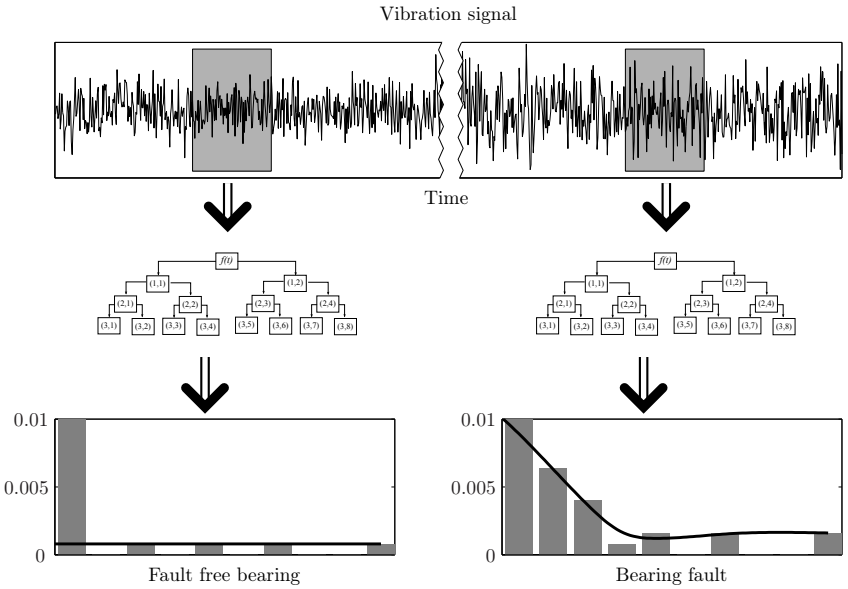


Figure 4: Calculation scheme.

corresponding to input points $\mathbf{x}_p$ and $\mathbf{x}_q$. Thus, the mean $\mu(\mathbf{x})$ and the covariance function $C(\mathbf{x}_p, \mathbf{x}_q)$ fully specify the Gaussian process.

The value of covariance function $C(\mathbf{x}_p, \mathbf{x}_q)$ expresses the correlation between the individual outputs $f(\mathbf{x}_p)$ and $f(\mathbf{x}_q)$ with respect to inputs $\mathbf{x}_p$ and $\mathbf{x}_q$. It should be noted that the covariance function $C(\,\cdot\,,\,\cdot\,)$ can be any function that generates a positive semi-definite covariance matrix. It is usually composed of two parts,

$$C(\mathbf{x}_p, \mathbf{x}_q) = C_{\mathrm{f}}(\mathbf{x}_p, \mathbf{x}_q) + C_{\mathrm{n}}(\mathbf{x}_p, \mathbf{x}_q), \tag{10}$$

where C_{f} represents the functional part and describes the unknown system we are modelling, and C_{n} represents the noise part and describes the model of noise.

Presuming white noise, the most commonly used is the constant covariance function. The choice of the covariance function for the functional part depends on the stationarity of the process. Assuming stationary data most commonly used covariance function is the square exponential covariance function. The composite covariance function is therefore

$$C(\mathbf{x}_p, \mathbf{x}_q) = v_1 \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d (x_{dp} - x_{dq})^2\right] + \delta_{pq} v_0, \tag{11}$$

where w_d are the automatic relevance determination hyperparameters, v_1 and v_0 are hyperparameters of the covariance function, D is the input dimension, and $\delta_{pq} = 1$ if $p = q$ and 0 otherwise. Hyperparameters can be written as a vector $\mathbf{\Theta} = [w_1, \ldots, w_D, v_1, v_0]^T$. The w_d indicate the importance of individual inputs. If w_d is zero or near zero, it means the inputs in dimension d contain little information and could possibly be discarded. Other forms and combinations of covariance functions suitable for various applications can be found in [38].

To accurately reflect the correlations presented in the training data, the hyperparameter values of the covariance function need to be optimized. Due to the probabilistic nature of the GP models, the common model optimization approach where model parameters and possibly also the model structure are optimized through the minimization of a cost function defined in terms of model error (e.g. mean square error), is not readily applicable. A probabilistic approach to the optimization of the model is more appropriate. Actually, instead of minimizing the model error, the probability of the model is maximized.

Consider a set of N D-dimensional input vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$ and a vector of output data $\mathbf{y} = [y_1, y_2, \ldots, y_N]$. Based on the data $(\mathbf{X}, \mathbf{y})$, and given a new input vector $\mathbf{x}^*$, we wish to find the predictive distribution of the corresponding output y^*. Based on training set $\mathbf{X}$, a covariance matrix $\mathbf{K}$ of size $N \times N$ is computed. The output of GP model is predictive distribution $p(y^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*)$ of the target y^*, given the training data $(\mathbf{y}, \mathbf{X})$ and an input $\mathbf{x}^*$. However, this distribution is conditioned on the hyperparameters $\mathbf{\Theta}$, which should be integrated out as:

$$p(y^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*) = \int p(y^*|\mathbf{\Theta}, \mathbf{y}, \mathbf{X}, \mathbf{x}^*) p(\mathbf{\Theta}|\mathbf{y}, \mathbf{X}) d\mathbf{\Theta}. \tag{12}$$

The computation of such integrals can be difficult due to the intractable nature of the non-linear functions. A solution to the problem of intractable integrals is to adopt numerical integration methods such as the Monte-Carlo approach. Unfortunately, significant computational efforts may be required to achieve a sufficiently accurate approximation.

Another standard practice for determining the predictive distribution is by maximum-likelihood estimation of hyperparameter values. This is achieved by minimising the following negative log-likelihood function:

$$\mathcal{L}(\mathbf{\Theta}) = -\frac{1}{2} \log(\mid \mathbf{K} \mid) - \frac{1}{2} \mathbf{y}^T \mathbf{K}^{-1} \mathbf{y} - \frac{N}{2} \log(2\pi). \tag{13}$$

Since the covariance matrix $\mathbf{K}$ in (13) depends on $\mathbf{\Theta}$, the likelihood function is non-linear and multi-modal. Therefore efficient optimisation routines require gradient information. The computation of the derivative of $\mathcal{L}(\mathbf{\Theta})$ with respect to each of the parameters is as follows:

$$\frac{\partial \mathcal{L}(\mathbf{\Theta})}{\partial \theta_i} = -\frac{1}{2} \text{trace} \left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_i} \right) + \frac{1}{2} \mathbf{y}^T \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_i} \mathbf{K} - 1 \mathbf{y}. \tag{14}$$

GP models can be easily utilised for regression, where the goal is to find the distribution of the corresponding output y^* for some new input vector $\mathbf{x}^* = [x_1(N+1), x_2(N+1), \ldots, x_D(N+1)]$. For the collection of random variables $[y_1, \ldots, y_N, y^*]$ we can write:

$$p(\mathbf{y}, y^*|\mathbf{X}, \mathbf{x}^*) = \mathcal{N}(0, \mathbf{K}^*), \tag{15}$$

with the covariance matrix

$$\mathbf{K}^* = \left[\begin{array}{c|c} \mathbf{K} & \mathbf{k}(\mathbf{x}^*) \\ \hline \mathbf{k}^T(\mathbf{x}^*) & \kappa(\mathbf{x}^*) \end{array} \right], \tag{16}$$

where $\mathbf{y} = [y_1, \ldots, y_N]$ is an $1 \times N$ vector of training targets, $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}_1, \mathbf{x}^*), \ldots, C(\mathbf{x}_N, \mathbf{x}^*)]^T$ is the $N \times 1$ vector of covariances between the test and training cases, and $\kappa(x^*) = C(\mathbf{x}^*, \mathbf{x}^*)$ is the covariance between the test input itself. The predictive distribution of the output $p(y^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*)$ is obtained by marginalising (15) and has a normal PDF with mean and variance:

$$\mu(y^*) = \mathbf{k}(\mathbf{x}^*)^T \mathbf{K}^{-1} \mathbf{y}, \tag{17}$$

$$\sigma^2(y^*) = \kappa(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}^*). \tag{18}$$

As can be seen from (18), the GP model, in addition to mean value, also provides information about the confidence in prediction by the variance. Usually the confidence of the prediction is depicted with 2σ interval which corresponds to approximately 95%. This confidence region can be seen as a grey band in Figure 5. It highlights areas of the input space where the prediction quality is poor due to the lack of data or noisy data, by indicating a wider confidence band around the predicted mean.

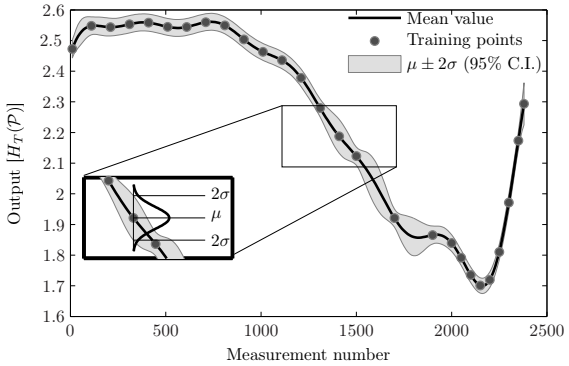


Figure 5: Modelling with GP models: in addition to mean value (prediction), we obtain a 95% confidence region for the underlying function f (shown in grey).

6. Procedure for the RUL estimation

The procedure for the RUL estimation from the complexity feature values consists of three main steps, namely the pre-processing of the feature values, RUL modelling with GPs and computation of the posterior PDF of the bearing's RUL.

6.1. Pre-processing of feature values

The computed features (2)–(4) include a relatively strong random component. Therefore, we introduced an intermediate step in which this random component is removed with a GP-based smoothing model. We assume, that the random component follows a normal distribution. The GP model for this task uses a composite covariance function (11).

The smoothing is then simply performed by estimating distribution of each training data time-series. The result of this process is a set of Gaussian distributions $\mathcal{N}(\mu_t, \sigma_t^2)$ estimated at each time moment t. This process is schematically shown in Figure 6. It should be noted that GP smoothing is performed without introducing any additional lag in the smoothed time-series, unlike other commonly used filtering methods such as moving average, exponential smoothing, etc.

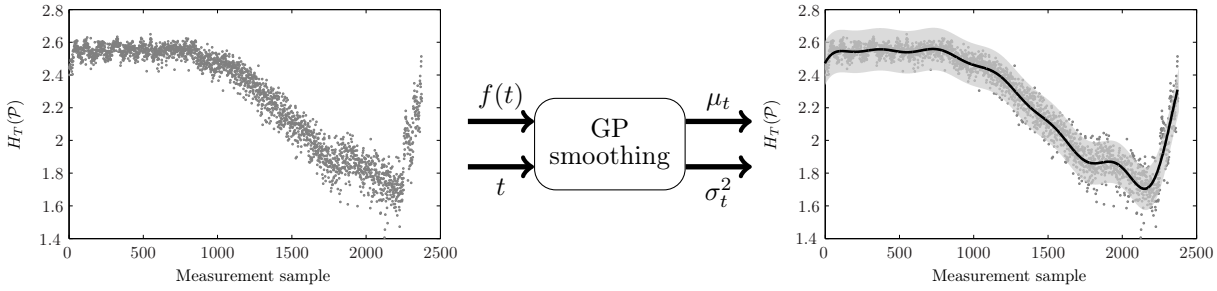


Figure 6: Smoothing process.

6.2. RUL modelling with GPs

The joint distribution of all smoothed time-series in the training dataset are used to construct a set of GP models that relate the feature value to the bearing's RUL. As the duration of the training datasets varies, the actual experiment time t was replaced by the life-cycle relative time index τ_i. The resulting training data are shown in Figure 7.

The result of the training process is a GP model which defines the evolution of the feature value for each $\tau_i \in [0,1]$. Given the input value of τ_i, the output of the GP model is a normal distribution describing the PDF of the feature value at the relative time τ_i. These points $\tau_i \in [0,1]$ are interpreted as a percentage of the used life. The training process can be described as:

$$GP : \mathcal{N}(\mu_i, \sigma_i^2 | \tau_i), \text{ where } \tau_i \in [0,1] \text{ and } i \in \mathbb{N}. \tag{19}$$

The mean values μ_i are shown with thick line in Figure 7.

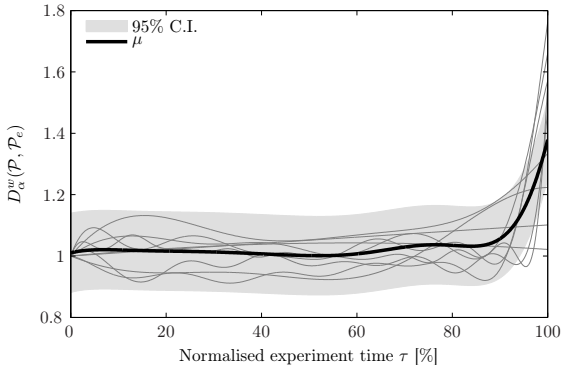


Figure 7: Time evolution of $D_\alpha^w(\mathcal{P}, \mathcal{P}_e)$ normed in the interval $[0,1]$.

6.3. RUL estimation

The bearing's RUL is estimated by computing the posterior distribution of the bearing life-cycle relative time index τ_i. As the training data points are normalised to lie on the interval $\tau_i \in [0,1]$, the RUL is simply $1 - \tau_i$. The posterior PDF of the distribution $p(\tau_i)$ is computed from the feature value $D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e)$ at the time instant t by following the Bayes' rule in the following form:

$$P(\tau_i | D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e)) \propto P(D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e) | \tau_i) P(\tau_i), \tag{20}$$

where the likelihood $P(D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e)|\tau_i)$ is given by the GP model (19) and the prior $P(\tau_i)$ in (20) can either include any additional knowledge related to the RUL distribution or can be set to an uninformative distribution.

If the informative prior is used in (20), the distribution $P(\tau_i)$ has to satisfy two main criteria. Firstly, it has to be conditioned on the current experiment duration t and secondly, it should be designed in a way that will give more weight to the prior at the beginning and more weight on the measurements, once they become significant. For this purpose, we propose the truncated normal distribution $T\mathcal{N}\left(\mu, \sigma^2\right)$ with PDF given as:

$$p(\tau) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{\exp\left(\dfrac{-(\tau-\mu)^2}{2\sigma^2}\right)}{\Phi\left(\dfrac{b-\tau}{\sigma}\right) - \Phi\left(\dfrac{a-\tau}{\sigma}\right)} I_{[a,b]}(\tau), \tag{21}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function and $I[a,b](\tau) = 1$ if $a \leq \tau \leq b$ and zero otherwise.

The posterior distribution is interpreted as a relative time of the experiment and therefore the prior should be limited to positive values of τ_i. To achieve this, the support of (21) is set to $a = 0, b = \infty$. Furthermore, the conditioning of the prior to current experiment time t is achieved by setting its mean value to $\mu = E(\tau) - t$, where $E(\tau)$ is the expected value of $1 - RUL$ (mean time to failure). Finally, the covariance is time dependent and set to $\sigma^2 = V_0 \cdot t$, where V_0 is the inflation constant. The result of inflation is that in

the initial stages of the bearing's life cycle, the prior will have a low covariance and will be the dominating part of (20). As the time progresses, the inflating covariance will effectively put more weight to the observed data and the GP model likelihood $P(D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e)|\tau_i)$ will dominate. Using the above definition, the proposed prior distribution $p(\tau)$ takes the form:

$$p(\tau) = \frac{1}{\sqrt{2\pi V_0 t}} \frac{\exp\left(\frac{-(\tau - (E(\tau) - t))}{2V_0 t}\right)}{1 - \Phi\left(\frac{-(E(\tau) - t)}{2V_0 t}\right)} I_{[0,\infty]}(\tau). \tag{22}$$

The important characteristic of this specific prior distribution is the truncation, which limits the prior only on positive values of time τ_i. From inspection of the (22), it can be seen that when the mean value $E(\tau) - t$ is far above 0 (e.g. more that 3σ), the truncation has practically no effect and the distribution is indistinguishable from a Gaussian one. However, when the mean value is approaching 0, the truncation limits the support to the selected interval and the denominator in (22) normalizes the function values. The resulting distribution thus has a mean value that is always greater than 0 and is slowly approaching it, which is an expected behavior of the distribution of the RUL.

The numerical estimation of the posterior (20) is schematically described in Figure 8. For a specific feature value $D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e)$, measured at time t, the likelihood $p(D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e)|\tau_i)$ is computed for each value of $\tau_i \in [0, 1]$. The likelihood is then multiplied by the prior (22), evaluated at the same values of τ_i and normalised. The result of the computation is the posterior PDF $p(\tau_i|D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e))$.

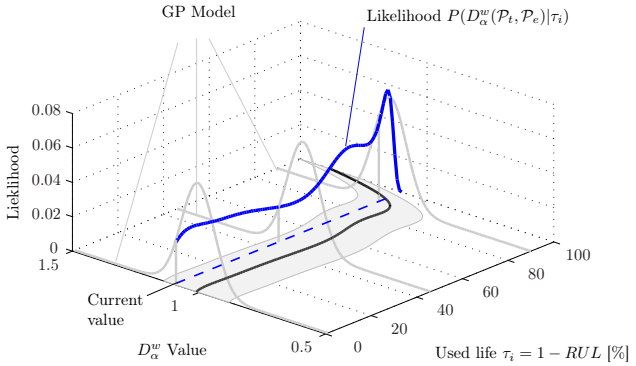


Figure 8: Calculating the probability for feature value $D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e) = 1.15$.

7. Prognostics results

7.1. Experimental setup

The proposed approach was evaluated on the data set for the IEEE PHM 2012 Data Challenge [25]. Provided data consist of three batches, each corresponding to different speed and load conditions. The generated vibrations were sampled with 22 kHz for duration of 100 ms, repeated every 5 minutes. The experiments were stopped when the RMS value of the generated vibrations surpassed 20 m/s^2.

Some of the experimental runs were rejected from the training process, since the time evolution of their features substantially differs from the majority. These rejections can be explained by two factors. Firstly, the tested bearings were subjected to loads several orders higher than the nominal ones. Secondly, the criterion for experiment end was selected as a hard threshold. Consequently, regardless of the initial high values of the vibration variance, some experiments lasted significantly longer. Therefore, as the majority of the experiments, 11 out of 17, show similar feature evolution, we assumed that the 6 rejected are not representative candidates, hence rejecting them from the training process.

7.2. Results

The procedure for RUL estimation described in Section 6 is applied for each of the 16 WP nodes, which results into 16 GP models. Each GP model describes the evolution of the Rényi entropy based features for each WP node. The final prediction of RUL is performed by fusing the predictions of all 16 GP models.

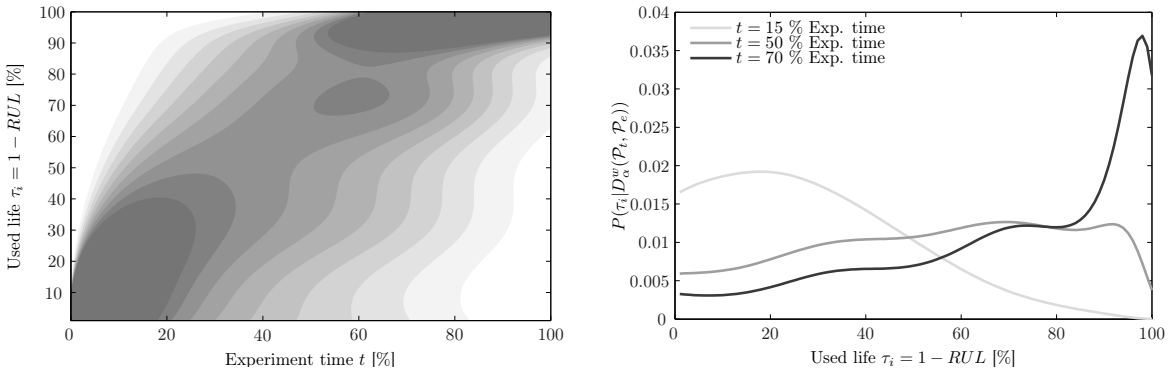


Figure 9: Evolution of $P(\tau_i|D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e))$ (bearings' used life) using the 3$^{\text{rd}}$ WP node.

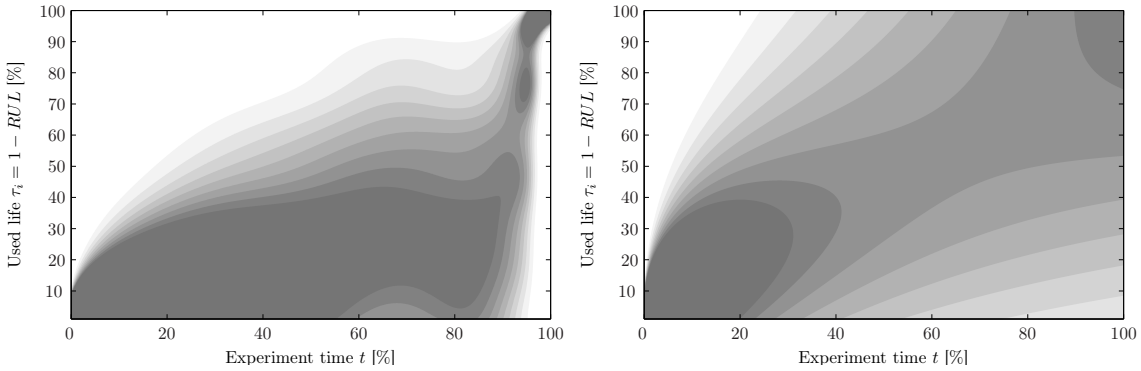
Using the Bayes' rule (20) with the truncated prior (22) bearing's RUL can be computed at any time moment. Such an evolution of RUL is shown in Figure 9. As experiment durations vary, the x-axis is normalised on the interval $[0, 1]$. The results exhibit almost linear relationship between the experiment time and the increase of the used life. At the very beginning, up to 20% of the experiment time, the variance of the posterior $P(\tau_i|D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e))$ (20) is small. In the middle of the experiment, the uncertainty of the estimates are somewhat higher. Towards the end of the experiment, when the measured feature values become sufficiently high, the estimates become more precise.

It should be noted that the RUL evolution differs depending on the WP node. In many cases, WP nodes spanning higher frequency bands exhibit early RUL decrease. On the other hand, the WP nodes spanning lower frequency bands become sensitive to RUL changes towards the end of the experiment. This effect, for WP nodes 4 and 14, is shown in Figure 10. It is clearly visible that the posterior distribution (20) for the 4$^{\text{th}}$ WP node has its mode around $\tau_i = 30\%$ for the majority of the experiment duration. At the same time, the posterior distribution (20) for the 14$^{\text{th}}$ WP node assigns sufficiently high likelihood values for $\tau_i > 70\%$ fairly early in the experiment. This effect can be employed as an early warning indicator of condition deterioration.

8. Conclusions

Monitoring the evolution of the Jensen-Rényi divergence and the Rényi entropy of vibrational signals using GP models leads to sufficiently accurate estimation of bearing's RUL. The proposed approach has two main advantages. Firstly, the calculation of the corresponding entropy based features requires no prior knowledge about the bearing's physical characteristics and no information about the operating conditions. Secondly, their wavelet based numerical estimation imposes no limits on the statistical characteristics of the analysed signals, which makes them suitable for monitoring bearings running under constant as well as variable operating conditions.

The benefits of applying GP models are twofold. As a pre-processing step, GP models performed smoothing of the feature values without inducing the lag in the time series. Based on this smoothed values, the training process defines the relation between the feature values and experiment time. The bearing's RUL is obtained as a posterior distribution using the Bayes' rule, where the likelihood is the relation specified by the GP model. The only element left is the prior distribution, which was in our case chosen to be a truncated Gaussian distribution.



(a) Node 4 (b) Node 14

Figure 10: $P(\tau_i|D_\alpha^w(\mathcal{P}_t, \mathcal{P}_e))$ estimates for different WP nodes.

The proposed approach was evaluated on limited data set. Increasing the set of available data should contribute to more precise definition of the prior distribution as well as the accuracy of the GP model mappings. However, regardless of the size and the quality of the available data set, the proposed approach is generally applicable for estimating bearing's RUL.

Acknowledgment

We like to acknowledge the support of the Slovenian Research Agency through the Research Programme P2-0001, the Research Project L2-4160. Additionally, we acknowledge the project EXLIZ – CZ.1.07/2.3.00/30.0013, which is co-financed by the European Social Fund and the state budget of the Czech Republic.

[1] P. F. Albrecht, J. C. Appiarius, D. K. Shrama, Assesment of the reliability of motors in utility applications, IEEE Transactions of Energy Conversion EC-1 (1986) 39–46.
[2] C. J. Crabtree, Survey of Commercially Available Condition Monitoring Systems for Wind Turbines, Tech. Rep., Durham University, School of Engineering and Computing Science, 2010.
[3] N. Tandon, A. Choudhury, A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings, Tribology International 32 (1999) 469–480.
[4] F. Camci, K. Medjaher, N. Zerhouni, P. Nectoux, Feature Evaluation for Effective Bearing Prognostics, Quality and Reliability Engineering International ISSN 1099-1638.
[5] Y. Li, T. Kurfess, S. Liang, Stochastic Prognostics for Rolling Element Bearings, Mechanical Systems and Signal Processing 14 (5) (2000) 747–762.
[6] N. Lybeck, S. Marble, B. Morton, Validating Prognostic Algorithms: A Case Study Using Comprehensive Bearing Fault Data, in: Aerospace Conference, 2007 IEEE, 1–9, 2007.
[7] M. N. Kotzalas, T. A. Harris, Fatigue Failure Progression in Ball Bearings, Transactions of ASME 123 (2001) 238–242.
[8] R. Li, P. Sopon, D. He, Fault features extraction for bearing prognostics, Journal of Intelligent Manufacturing 23 (2012) 313–321, ISSN 0956-5515.
[9] W. Wang, Bearing Condition Monitoring for Turbomachinery, COMADEM 2011 Keynote, 2011.
[10] W. Wang, Autoregressive model-based diagnostics for gears and bearings, Insight 50 (5) (2008) 1–5.
[11] R. B. Randall, The Challenge of Prognostics of Rolling Element Bearings, in: Wind Turbine Condition Monitoring Workshop, 2011.
[12] J. Qiu, B. B. Seth, S. Y. Liang, C. Zhang, Damage Mechanics Approach for Bearing Lifetime Prognostics, Mechanical Systems and Signal Processing 16 (5) (2002) 817–829.
[13] H. Ocak, K. A. Loparo, F. M. Discenzo, Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling: A method for bearing prognostics, Journal of Sound and Vibration 302 (4–5) (2007) 951–961.
[14] S. Janjarasjitt, H. Ocak, K. Loparo, Bearing condition diagnosis and prognosis using applied nonlinear dynamical analysis of machine vibration signal, Journal of Sound and Vibration 317 (1–2) (2008) 112–126.
[15] M. Martin, A. Plastino, O. Rosso, Generalized statistical complexity measures: Geometrical and analytical properties, Physica A 369 (2) (2006) 439–462.
[16] X. C. R. López-Ruiz, H.L. Mancini b, A statistical measure of complexity, Physics Letters A 209 (1995) 321–326.
[17] O. Rosso, M. Martin, A. Figliola, K. Keller, A. Plastino, EEG analysis using wavelet-based information tools, Journal of Neuroscience Methods 153 (2006) 163–182.

[18] P. Boškoski, Đ. Juričić, Fault detection of mechanical drives under variable operating conditions based on wavelet packet Rényi entropy signatures, Mechanical Systems and Signal Processing 31 (2012) 369—381, ISSN 0888-3270.
[19] P. Boškoski, Đ. Juričić, Rényi Entropy Based Statistical Complexity Analysis for Gear Fault Prognostics under Variable Load, in: T. Fakhfakh, W. Bartelmus, F. Chaari, R. Zimroz, M. Haddar (Eds.), Condition Monitoring of Machinery in Non-Stationary Operations, Springer Berlin Heidelberg, ISBN 978-3-642-28767-1, 25–32, 2012.
[20] K. Ažman, J. Kocijan, Application of Gaussian processes for black-box modelling of biosystems, ISA Transactions 46 (4) (2007) 443–457.
[21] Ž. Južnič-Zonta, J. Kocijan, X. Flotats, D. Vrečko, Multi-criteria analyses of wastewater treatment bio-processes under an uncertainty and a multiplicity of steady states, Water Research 46 (18) (2012) 6121–6131.
[22] B. Grašič, P. Mlakar, M. Z. Božnar, Ozone prediction based on neural networks and Gaussian processes, Nuovo Cimento C 29 (2006) 651–661.
[23] B. Likar, J. Kocijan, Predictive control of a gas-liquid separation plant based on a Gaussian process model, Computers & chemical engineering 31 (3) (2007) 142–152.
[24] J. Kocijan, V. Tanko, Prognosis of gear health using Gaussian process model, in: Proceedings of EUROCON 2011, International Conference on Computer as a Tool, Lisbon, Portugal, 1–4, 2011.
[25] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Morello, N. Zerhouni, C. Varnier, PRONOSTIA: An Experimental Platform for Bearings Accelerated Life Test, in: IEEE International Conference on Prognostics and Health Management, Denver, CO, USA, 2012.
[26] J. P. Crutchfield, K. Young, Inferring statistical complexity, Phys. Rev. Lett. 63 (2) (1989) 105–108.
[27] C. Adami, What is complexity?, BioEssays 24 (12) (2002) 1085–1094.
[28] A. M. Kowalski, M. T. Martin, A. Plastino, O. A. Rosso, M. Casas, Distances in Probability Space and the Statistical Complexity Setup, Entropy 13 (6) (2011) 1055–1075.
[29] A. Rényi, On measures of information and entropy, in: 4th Berkeley Symposium on Mathematics, Statistics and Probability, 1960.
[30] M. Basseville, Divergence measures for statistical data processing, Tech. Rep., IRISA, 2010.
[31] A. O. Hero, B. Ma, O. Michel, J. Gorman, Alpha-divergence for classification, indexing and retrieval, Tech. Rep. CSPL-328, Communications and Signal Processing Laboratory, The University of Michigan, 2002.
[32] R. B. Randall, J. Antoni, S. Chobsaard, The relationship between spectral correlation and envelope analysis in the diagnostics of bearing faults and other cyclostationary machine signals, Mechanical Systems and Signal Processing 15 (2001) 945 – 962.
[33] J. Antoni, R. B. Randall, A stochastic model for simulation and diagnostics of rolling element bearings with localized faults, Journal of Vibration and Acoustics 125 (3) (2003) 282–289.
[34] J. Antoni, Cyclostationarity by examples, Mechanical Systems and Signal Processing 23 (2009) 987–1036.
[35] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, Burlington, MA, 3rd edn., 2008.
[36] D. B. Percival, A. T. Walden, Wavelet Methods for Time Series Analysis, Cambridge University Press, Cambridge, 2000.
[37] S. Blanco, A. Figliola, R. Q. Quiroga, O. A. Rosso, E. Serrano, Time-frequency analysis of electroencephalogram series. III. Wavelet packets and information cost function, Phys. Rev. E 57 (1) (1998) 932–940.
[38] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.
[39] C. E. Rasmussen, Advances in Gaussian processes, Advances in Neural Information Processing Systems .
[40] M. Seeger, Gaussian processes for machine learning, International Journal of Neural Systems 14 (2) (2004) 69–106.
[41] D. J. C. MacKay, Introduction to Gaussian processes, NATO ASI Series 168 (1998) 133–166.