Mechanical Systems and Signal Processing 52/53 (2015) 327–337 doi: 10.1016/j.ymssp.2014.07.011

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

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

[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

Bearings are considered to be the most frequent cause for failures in rotational machinery. Hence efficient means to anticipate the remaining useful life (RUL) on-line, by processing the available sensory records, is of substantial practical relevance. Many of the data-driven approaches rely on conjecture that evolution of condition monitoring (CM) indices are related with the aggravation of the condition and, indirectly, with the remaining useful life of a bearing. Problems with trending may be threefold: (i) most of the operational life show no significant trend until the time very close to failure; this is usually accompanied by rapidly growing values of CM indices which is not easy to forecast, (ii) the evolution of CM indices is not necessarily monotonous, (iii) variable and immeasurable fluctuations in operating may fool the trend. Motivated by these issues we propose an approach for bearing fault prognostics that employs Rényi entropy based features. It exploits the idea that progressing fault implicates raising dissimilarity in the distribution of energies across the vibrational spectral band sensitive to the bearing faults. The innovative way of predicting RUL relies on a posterior distribution following the Bayes' rule using Gaussian process (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

According to several surveys, bearing faults represent the most frequent cause for failure of mechanical drives [1, 2]. Therefore, suitable methods for fault detection and prognostics of bearing faults is of paramount practical importance. As a result, a plethora of methods for detection of bearing faults have been developed.

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

The most known methods rely on set of features, which is based on characteristic bearing fault frequencies associated to specific bearing surface faults [3]. Although effective for bearing fault detection, these features turn ineffective for estimating bearings' remaining useful life (RUL). The reason is that their values stay close to zero up to the moment when the failure itself occurs [4]. To cope with the problem we exploit the idea according to which the progressing fault implicates raising dissimilarity in the pattern of distribution of energies across particular vibrational spectral band, which is sensitive to the bearing faults. These patterns are characterised by entropy indices while the dissimilarity is expressed in terms of divergence. Hence the divergence appear as a feature indicating the condition aggravation level. Addressing the problem of bearing fault prognostics, in this paper we propose new Rényi entropy features associated with the statistical properties of the envelope of bearing's vibrations.

Another key idea of the paper relies on the conjecture that discrepancy between the distributional patterns is directly related with the bearing condition. As the worsening condition is monotone process it is related with the bearing's life span. Of course, it is impossible to determine this inter-relationship other than by using measurements from a set of run-to-failure experiments. In this paper we propose a fully data driven approach relying on Gaussian process (GP) models for calculating bearing's RUL.

The problems of bearing fault prognostics attracted a lot of attention in the past years. The majority of the proposed approaches try to describe the relationship between the defect growth and the time evolution of some condition monitoring indices (or features) calculated from vibrations like energy, peak-to-peak values, RMS, kurtosis, crest factor, changes in bearing natural frequency etc. [4–10]. In many cases additional signal properties are calculated using variations of wavelet transform [11, 12]. Using somewhat similar approach, Ocak et al. [13] model the evolution of the energy of particular wavelet packet nodes using hidden Markov models. 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.

In this paper we show that the evolution of properly selected (Jensen-)Rényi entropy based indices of the generated vibrations can be related to the bearing's RUL. In addition, the process of inference based on Jensen-Rényi divergence requires no prior information about the operating conditions. Under some conditions the prognostic scheme is able to operate even under incomplete prior knowledge about the physical characteristics of the monitored drive. This remarkable property has been demonstrated in the context of fault diagnosis by Boškoski and Juričić [15, 16].

Based on the values of the Jensen-Rényi divergence, the bearing's RUL is estimated using GP models. The reason why GP models are used is because it is non-parametric approach meaning that no prior assumptions about the candidate model structures is needed, which in a sense makes modelling simpler than in the case when parametric models are applied. Additional rationale is in the fact that the model output is not a vector of real numbers but joint probability density function of the outputs. Hence full information about the computed output, including uncertainties, is provided. The output of the GP models is a normal

distribution, expressed in terms of mean and variance. The mean value represents the most likely output and the variance can be interpreted as a measure of its confidence. 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 [17, 18], environmental systems [19], chemical engineering [20] and many others. Kocijan and Tanko [21] used GP models for modelling time series describing gear health and the prediction of the critical value of harmonic component feature that indicates the wear of gear.

The paper is organised as follows. In section 2 the statistical model of vibrational signals is presented first. The detailed definition of the selected features and their numerical estimation is presented in Section 3. The properties of GP models are presented in Section 4. In Section 5, the RUL distribution $p(RUL)$ is obtained as a posterior distribution by using the output of the trained GP model as likelihood. The evaluation of the proposed approach, presented in Section 6, is done on the data set provided for the IEEE PHM 2012 Prognostic Challenge [22].

2. Statistical model 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 [23]:

$$y(t) = \sum_{i=-\infty}^{+\infty} A_i s(t - \nu_i) + n(t), \tag{1}$$

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.

For healthy bearing, the envelope of the generated vibrations will be without any visible structure due to the lack of impacts $s(t - \nu_i)$. The presence of bearing surface fault introduces additional components that influence the shape of the envelope hence altering the shape of its distribution. Therefore, the goal is to quantifying these changes in the shape of the envelope distribution for the purpose of RUL estimation.

3. Rényi entropy and related indices

Information about the evolving condition of the bearing is buried in the envelope of the vibration signal. So, the first step is the estimation of the probability distribution functions (PDF) of the envelope of the generated vibrations. Due to the link between the signal's envelope and its instantaneous power [24], in this

approach the underlying PDF is estimated through the energy distribution of the wavelet packet coefficients of vibration signal.

For the computation of the coefficients the wavelet packet transform (WPT) is used [25]. The structure of WPT is described by a binary tree structure, as shown in Figure 1. 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.

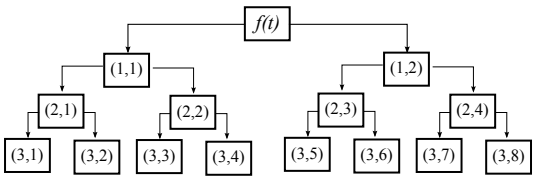


Figure 1: Example of a full WPT tree with depth $d_M=3$.

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 [26]. Using these coefficients, the portion of the signal's energy $E_{d,n}$ contained within one node (d,n) reads [27]:

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

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

$$E_{tot}=\sum_{\substack{t=0\\ d,n\in T}}^{N_d-1}\|W_{d,n,t}\|^2=\sum_{d,n\in T}E_{d,n}. \tag{3}$$

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):

$$\mathcal{P}^{d,n}=\left\{p_t^{d,n}=\frac{\|W_{d,n,t}\|^2}{E_{d,n}},\ t=0,\cdots,N_d-1\right\}. \tag{4}$$

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}:

$$\mathcal{P}^{T}=\left\{p_{d,n}=\frac{E_{d,n}}{E_{tot}},\ d,n\in T\right\}. \tag{5}$$

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 (7) and (8).

3.1. Entropy

The concept of entropy serves to characterise the 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:

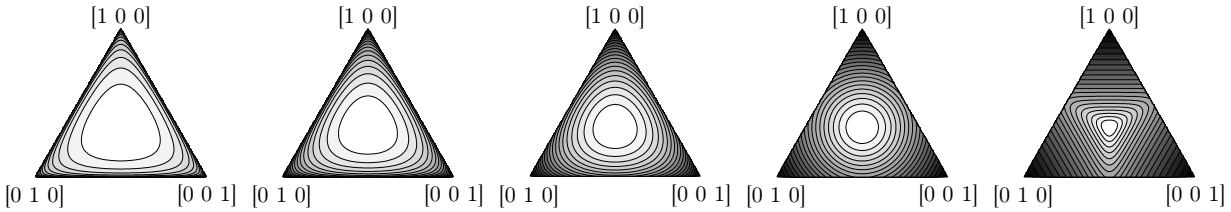
$$H(\mathcal{P}) = -\sum_{i=1}^{N} p_i \ln(p_i). \tag{6}$$

For a discrete set with cardinality N the 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 [28]:

$$H_\alpha(\mathcal{P}) = \frac{1}{1-\alpha} \ln \sum_{i=1}^{N} p_i^\alpha(x),\ \alpha \geq 0\, \alpha \neq 1. \tag{7}$$

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 $\mathcal{P}$.

The α exponent in the Rény entropy specifies the relative importance of small values versus large values of the probability mass. This effect can be visualised using the isoentropy plots of Xu and Erdogmuns [29] for all possible probability distributions over N bins. For the case where $N = 3$, the isoentropy plots for different values of α are shown in Figure 2.



(a) $\alpha = 0.2$ (b) $\alpha = 0.5$ (c) $\alpha = 1$ Shannon entropy (d) $\alpha = 2$ (e) $\alpha = 10$

Figure 2: Rény Isoentropy plots for probability distributions with three components as in [29]

3.2. 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 reads [30]:

$$D_\alpha^w(\mathcal{P}, \mathcal{Q}) = H_\alpha\left(w\mathcal{P} + (1-w)\mathcal{Q}\right) - \left\{wH_\alpha(\mathcal{P}) + (1-w)H_\alpha(\mathcal{Q})\right\}, \tag{8}$$

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]. In the presented approach the value of α was set to $\alpha = 0.5$.

The Jensen-Rényi entropy reflects the condition of the bearing. This fact will be utilised in the context of prognostics as well. In most of the applications vibrational data records are collected during the repetitive acquisition sessions at high enough sampling rate. Each session results in a vibrational record for which the divergence index is calculated from coefficients of the wavelet packet transform.

The idea is illustrated in Figure 3. At the beginning of the monitoring process the reference condition $\mathcal{P}_0$ should be defined by computing the corresponding Rényi entropy for both $\mathcal{P}^{d,n}$ and $\mathcal{P}^T$. 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 into $\mathcal{P}_t$, hence altering the corresponding values of (8). 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 too big, 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.

3.3. Statistical complexity

The statistical complexity $\mathscr{C}(\mathcal{P})$ of a signal with distribution $\mathcal{P}$ based on (7) and (8) is defined as [32]:

$$\mathscr{C}(\mathcal{P}) = Q_0 D_\alpha^w(\mathcal{P}, \mathcal{P}_e) H_\alpha(\mathcal{P}), \tag{9}$$

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 (9) 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.

4. Gaussian process models

The relationship between cause and consequence, or system input and system output, can be modelled in many ways. If the physics behind the relationship is known one can use first principle models. If this is not the case, models need to be derived out of the available input and output data. Those approaches are referred to as data-driven.

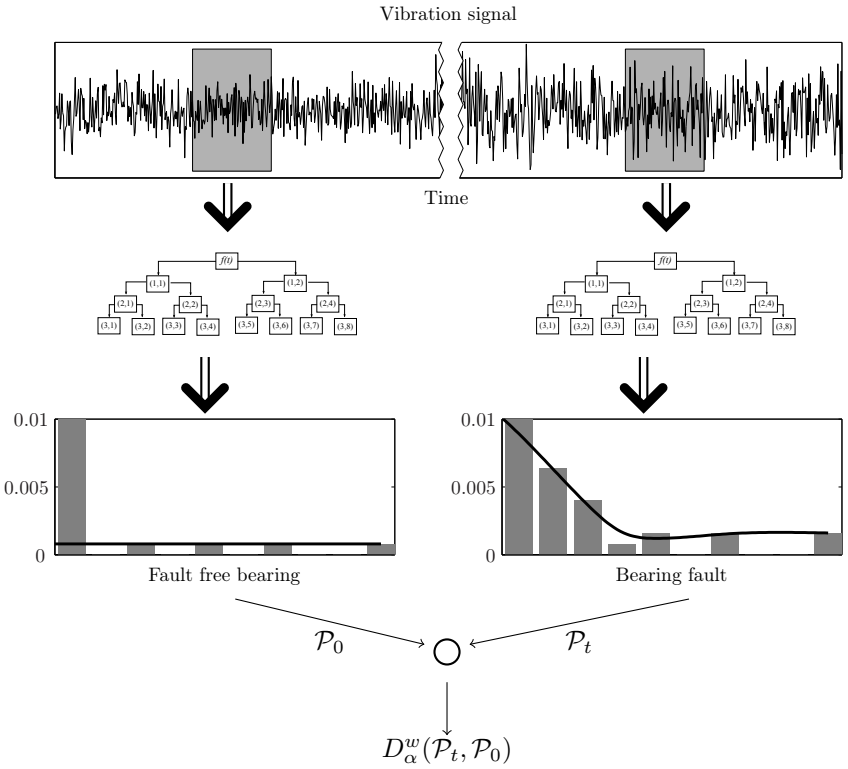


Figure 3: Monitoring the condition degradation by means of Jensen-Rényi divergence.

The most frequent way to relate input $\mathbf{x} \in R^D$ and output $y \in R$ is to use regression model parameterized by a finite vector of model parameters $\theta \in R^D$. For the sake of simplicity we will consider static relationship

$$y(\mathbf{x}) = \mathbf{f}(\mathbf{x}; \theta) + \epsilon \tag{10}$$

where $f : R^D \to R$ is a known function (e.g. linear, polynomial, radial basis function, neural network etc.) known up to the vector θ, ϵ is noise term needed to describe model imperfection caused by random disturbances or modelling errors. Usually the noise term is described by a probability density function $\epsilon \sim p_\vartheta(\epsilon)$ parameterized by ϑ. Having the data records in terms of pairs $\{\mathbf{x}_i, y_i, i = 1, ..., N\}$ the model (10) can be identified from data by estimating the the unknown model parameters $\{\theta, \vartheta\}$.

The alternative to the parametric model (10) is to use structure-free non-parametric models. Such is the case with the GP model. It is also referred to as Bayesian kernel model.

The idea of GP models is rather simple. An outline from the intuitive point of view will be provided, however, more rigorous derivation can be found in [33, 34]. Assume we dispone of N D-dimensional inputs $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ and corresponding outputs $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$. GP model assumes that the output is realization of a GP with joint probability density function

$$\mathbf{y} \sim \mathcal{N}(\mathbf{m}, \mathbf{K}) \tag{11}$$

with mean and covariance being functions of the inputs $\mathbf{X}$. In the most general case we have

$$m_i = m(\mathbf{x}_i), \quad j = 1, \ldots, N \tag{12}$$

and

$$\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, \ldots, N \tag{13}$$

where the right side refers to as *covariance function* or *kernel*.

Presuming white noise and stationary data, the most commonly used is the composition of the squared exponential covariance function and the constant covariance function

$$k(\mathbf{x}_i, \mathbf{x}_j) = v_1 \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d (x_{di} - x_{dj})^2\right] + \delta_{ij} v_0, \tag{14}$$

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 $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. Hyperparameters can be written as a vector $\boldsymbol{\Theta} = [w_1, \ldots, w_D, v_1, v_0]^T$. 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 [34].

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.

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 $(\mathbf{X}, \mathbf{y})$ and an input $\mathbf{x}^*$. However, this distribution is conditioned on the hyperparameters $\boldsymbol{\Theta}$, which should be integrated out as:

$$p(y^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*) = \int p(y^*|\boldsymbol{\Theta}, \mathbf{y}, \mathbf{X}, \mathbf{x}^*) p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X}) d\boldsymbol{\Theta}. \tag{15}$$

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:

$$\mathcal{L}(\boldsymbol{\Theta}) = -\frac{1}{2}\log(\mid \mathbf{K} \mid) - \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi). \tag{16}$$

Since the covariance matrix $\mathbf{K}$ in (16) depends on $\boldsymbol{\Theta}$, the likelihood function is non-linear and multimodal. Therefore efficient optimisation routines require gradient information. The computation of the derivative of $\mathcal{L}(\boldsymbol{\Theta})$ with respect to each of the parameters is as follows:

$$\frac{\partial\mathcal{L}(\boldsymbol{\Theta})}{\partial\theta_i} = -\frac{1}{2}\text{trace}\left(\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta_i}\right) + \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta_i}\mathbf{K}{-}1\mathbf{y}. \tag{17}$$

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:

$$p(\mathbf{y}, y^*|\mathbf{X}, \mathbf{x}^*) = \mathcal{N}(0, \mathbf{K}^*), \tag{18}$$

with the covariance matrix

$$\mathbf{K}^* = \left[\begin{array}{c|c} \mathbf{K} & \mathbf{k}(\mathbf{x}^*) \\ \hline \mathbf{k}^T(\mathbf{x}^*) & \kappa(\mathbf{x}^*) \end{array}\right], \tag{19}$$

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 (18) and has a normal PDF with mean and variance:

$$\mu(y^*) = \mathbf{k}(\mathbf{x}^*)^T\mathbf{K}^{-1}\mathbf{y}, \tag{20}$$

$$\sigma^2(y^*) = \kappa(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^T\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}^*). \tag{21}$$

As can be seen from (21), the GP model, in addition to the mean value, also provides information about the confidence in prediction by the variance. Usually the confidence of the prediction is depicted with a 2σ interval which corresponds to approximately 95%. This confidence region can be seen as a grey band in Figure 4. 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.

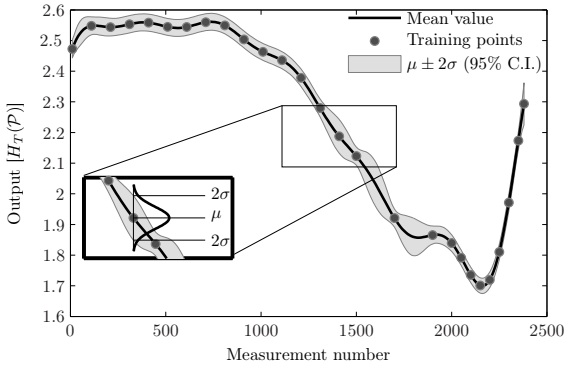


Figure 4: Modelling with GP models: in addition to mean value (prediction), we obtain a 95% confidence region for the underlying function f (shown in grey).

5. Procedure for RUL prediction

The procedure for estimation of RUL relies on the assumption that off line data records from run-to-failure test on similar bearings have been performed a priori. Each such data record contains time track of Jensen-Rényi divergences calculated at the corresponding instances of time. Hence one gets pairs $\{\tau, D_\alpha^w(\mathcal{P}_t, \mathcal{P}_0)\}$, where $\mathcal{P}_t$ denotes the distribution of energies in current operating window while $\mathcal{P}_0$ stands for the distribution at the beginning of the operating life, when bearing is assumed to be in the nominal condition.

Due to the statistical nature of the Jensen-Rényi divergence feature, the feature time series includes a relatively large random component. Therefore, each dataset is pre-processed by an individual GP model with a composite covariance function (14). The result of a GP modeling is a smooth time series described by a set of Gaussian distributions $\mathcal{N}(\mu_t, \sigma_t^2)$ for each dataset. The training feature used in the subsequent steps is the vector of GP model mean values μ_t.

In addition to acquisition of off-line data and its pre-processing, the procedure for estimation of RUL consists of two main steps, which will be presented here in more detail. First one is to infer an appropriate model of the feature value and the second one is to use the model and the current measurements to predict the RUL.

5.1. RUL modelling with GPs

The joint distribution of all the training time-series in the training dataset are used to infer a set of GP models that model the relation between the feature value and 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, where $0 \leq \tau \leq 1$. Value $\tau = 1$ means that the bearing has reached the end of life. Such rescaled training data are shown in Figure 5.

The result of the training process is a GP model which defines the evolution of the feature value for each $\tau \in [0,1]$. Given the input value of τ, the output of the GP model is a normal distribution describing the PDF of the feature value at the relative time τ:

$$p(D_\alpha^w(\mathcal{P}_t,\mathcal{P}_0)|\tau) \sim \mathcal{N}(\mu(\tau),\sigma^2(\tau)) \;\; \tau \in [0,1]. \tag{22}$$

The mean values μ_τ are shown with thick line in Figure 5.

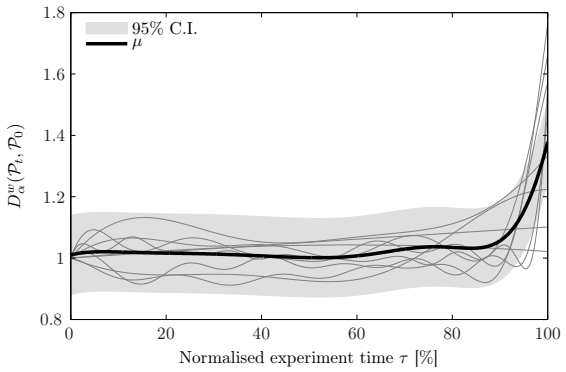


Figure 5: Time evolution of $D_\alpha^w(\mathcal{P}_t,\mathcal{P}_0)$ normed in the interval $[0,1]$.

5.2. RUL estimation

The bearing's RUL is estimated by computing the posterior distribution of the bearing relative time τ. As the training data points are normalised on the interval $\tau \in [0,1]$, the RUL is simply $1-\tau$. The posterior PDF of the distribution $p(\tau)$ is computed from the current feature value $D_\alpha^w(\mathcal{P}_t,\mathcal{P}_0)$ at the time instant t by following the Bayes' rule in the following form:

$$p(\tau|D_\alpha^w(\mathcal{P}_t,\mathcal{P}_0)) \propto p(D_\alpha^w(\mathcal{P}_t,\mathcal{P}_0)|\tau)p(\tau), \tag{23}$$

where the likelihood $p(D_\alpha^w(\mathcal{P}_t,\mathcal{P}_0)|\tau)$ is given by the GP model (22) and the prior $p(\tau)$ in (23) includes any additional knowledge related to the RUL distribution. If the knowledge is missing one set it to an uninformative distribution.

If the informative prior is used in (23), the distribution $p(\tau)$ has to satisfy two main criteria. Firstly, it has to include information about 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:

$$p(\tau) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{\exp\left(\dfrac{-(\tau-\mu)^2}{2\sigma^2}\right)}{\Phi\left(\dfrac{b-\tau}{\sigma}\right) - \Phi\left(\dfrac{a-\tau}{\sigma}\right)} I_{[a,b]}(\tau), \tag{24}$$

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. To achieve this, the support of (24) is set to $a = 0, b = 1$. Furthermore, the conditioning of the prior to the current experiment time t is achieved by setting its mean value to $\mu_t = E(1-\tau) - t$, where $E(1-\tau)$ is the mean time to failure. Finally, the covariance is time dependent and set to $\sigma_t^2 = V_0 \cdot t$, 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 low covariance and will be the dominating part of (23). As the time progresses, the inflating covariance will put more weight to the observed data and the GP model likelihood $P(D_\alpha^w(\mathcal{P}_t, \mathcal{P}_0)|\tau)$ will dominate. Using the above definition, the proposed prior distribution $p(\tau)$ takes the form:

$$p(\tau) = \frac{1}{\sqrt{2\pi V_0 t}} \frac{\exp\left(\dfrac{-(\tau - (E(1-\tau) - t))}{2V_0 t}\right)}{\Phi\left(\dfrac{1 - (E(1-\tau) - t)}{2V_0 t}\right) - \Phi\left(\dfrac{-(E(1-\tau) - t)}{2V_0 t}\right)} I_{[0,\infty]}(\tau). \tag{25}$$

The important characteristic of this specific prior distribution is the truncation, which limits the prior only to the positive values of time τ. From (25), it can be seen that when the mean value is far above 0, the truncation has practically no effect and the distribution is indistinguishable from Gaussian one. However, when the mean value is approaching 0, the truncation limits the support to the selected interval and the denominator in (25) 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 (23) is schematically described in Figure 6. For a specific feature value $D_\alpha^w(\mathcal{P}_t, \mathcal{P}_0)$, measured at time t, the likelihood $p(D_\alpha^w(\mathcal{P}_t, \mathcal{P}_0)|\tau)$ is computed for each value of $\tau \in [0,1]$. The likelihood is then multiplied by the prior (25), evaluated at the same values of τ. The result of the computation is the posterior PDF $p(\tau|D_\alpha^w(\mathcal{P}_t, \mathcal{P}_0))$.

6. Experimental results

6.1. The experimental setup

The proposed approach was evaluated on data sets for the IEEE PHM 2012 Data Challenge [22]. 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 exceeded 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 justified in two ways. Firstly,

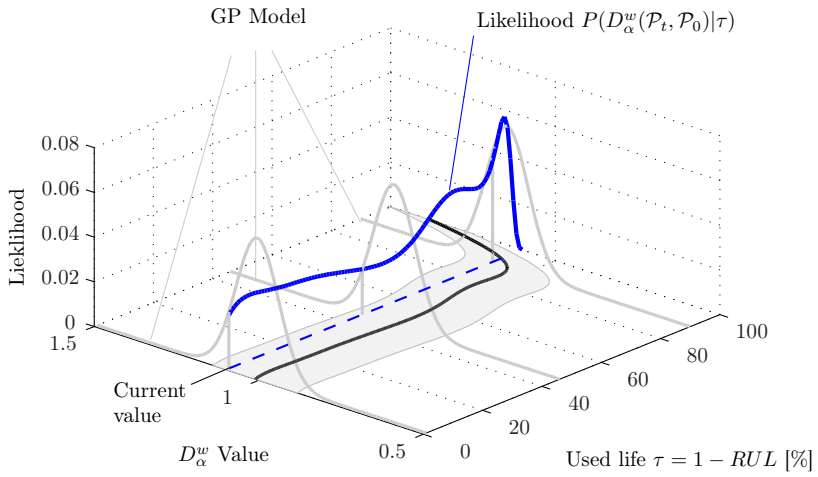
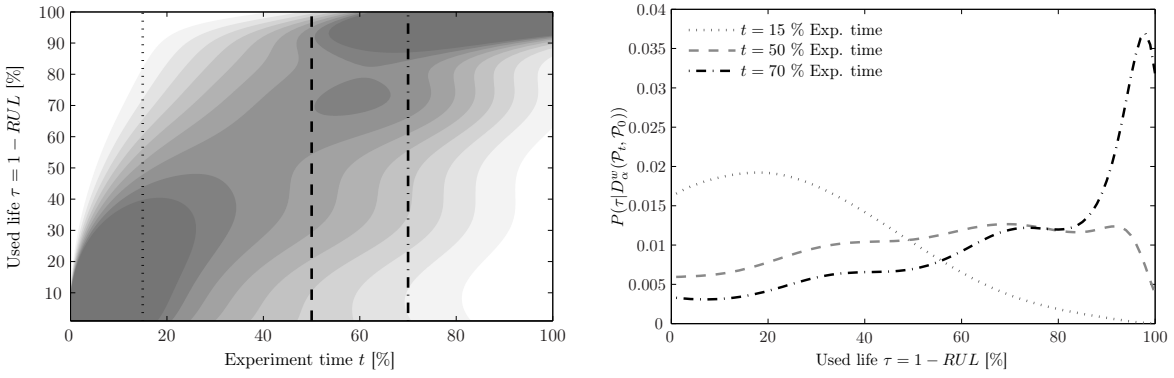


Figure 6: Calculating the probability for feature value $D_\alpha^w(\mathcal{P}_t,\mathcal{P}_0) = 1.15$.

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, therefore were omitted from the training process.

6.2. Results

Using the Bayes' rule (23) with the truncated prior (25) bearing's RUL can be computed at any time moment. Such an evolution of RUL is shown in Figure 7. 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.



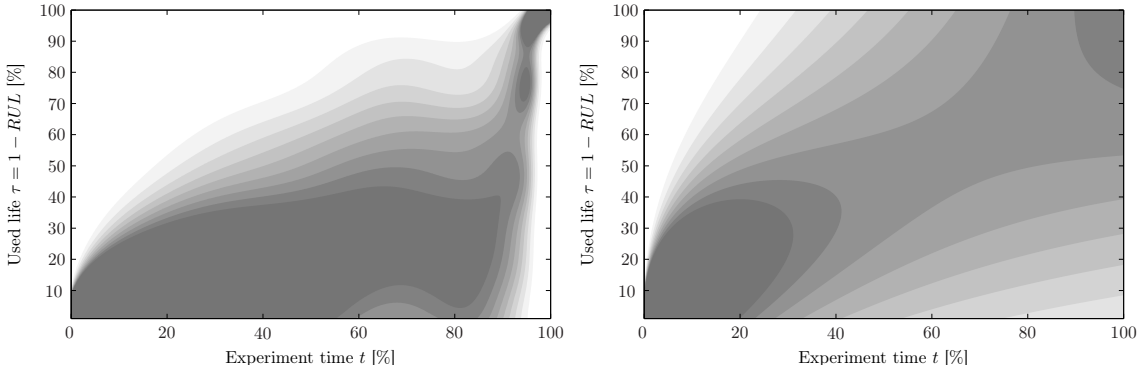
(a) Distribution of RUL throughout the experiment

(b) Distribution of RUL at certain time moments

Figure 7: Evolution of $P(\tau|D_\alpha^w(\mathcal{P}_t,\mathcal{P}_0))$ (bearings' used life) using the 3$^{\text{rd}}$ WP node.

The distribution of the RUL at particular time moments is shown in Figure 7(b). The distributions

represent vertical slices at particular time moments of the posteriors shown in Figure 7(a). At the very beginning, up to 15% of the experiment time, the prior (25) has sufficiently low variance and dominates the shape of the posterior (23). In the middle of the experiment, around 50% of the experiment time, the posterior is almost uniformly spread, which limits the capabilities for accurate prediction. However, towards the experiment end, around 70% of the experimental time, the posterior clearly shows that the bearing has used up almost all of its useful time. The posterior PDF is right skewed with sufficiently low variance and mode in the vicinity of 90%.



(a) Node 4 (b) Node 14

Figure 8: $P(\tau|D_\alpha^w(\mathcal{P}_t, \mathcal{P}_0))$ estimates for different WP nodes.

6.3. Overall RUL prediction

It should be noted that the RUL evolution differs depending on the wavelet packet (WP) node. In the current approach the employed WP tree has 16 terminal nodes. For each WP node one can calculate the corresponding posterior (23). The final RUL prediction can be achieved using different approaches, for instance: selecting the WP node that spans the most informative frequency band or perform fusion of posteriors from each WP nodes.

When performing fusion of posteriors from different WP nodes, it should be noted that 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 8. It is clearly visible that the posterior distribution (23) for the 4^{th} WP node has its mode around $\tau = 30\%$ for the majority of the experiment duration. At the same time, the posterior distribution (23) for the 14^{th} WP node assigns sufficiently high likelihood values for $\tau > 70\%$ fairly early in the experiment. This effect can be employed as an early warning indicator of condition deterioration.

Since the bearings used for generating the data sets were from the same type, it is rather straightforward to determine the frequency band in which bearing faults are most visible. Therefore, we selected the information from the 4^{th} WP node as the most representative one for RUL prediction.

6.4. Comments on the results

The result of the proposed method is the posterior distribution (23) describing the bearing's used life. Generally, the posterior distribution is defined only on positive semi-axis and therefore is not Gaussian. Consequently when analysing the results, besides the mean and the variance of the the posterior (23), additional statistical properties should be considered, such as skewness, kurtosis etc. Therefore, the complete posterior distribution of the bearing's used life can be considered as more informative.

Despite the apparent complexity of the approach, the only design parameter for calculating the posterior distribution (23), is the prior distribution $P(\tau)$. In our case, the prior describing the expected bearing's behaviour was specified by a truncated Gaussian distribution (25). The parameters of the prior can be determined by taking into consideration typical information regarding the bearing quality, for instance the L_{10} coefficient. Therefore, the proposed method inherently allows integration of prior knowledge either in a form of producer's information or experience from historical data.

7. Conclusions

In the paper we show that monitoring the evolution of the Jensen-Rényi divergence of vibrational signals using GP models leads to the sufficiently accurate prediction 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 numerical using wavelet packet transform 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 bearing's remaining useful life is estimated as a posterior distribution using the Bayes' rule. The likelihood distribution, describing the evolution of the Jensen-Rényi divergence, is estimated by using GP models. This paper proposes an informative prior in a form of truncated Gaussian distribution, whose parameters can be selected based on some typical information about the bearing quality. As a result, the proposed procedure becomes broadly applicable.

The evaluation was performed on a relatively limited data set. Increasing the set of available data should contribute to more precise definition of the prior distribution as well as, to the accuracy of the estimated likelihoods. 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 and the Research Projects L2-4160 and Z2-5477. 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] P. Borghesani, P. Pennacchi, S. Chatterton, The relationship between kurtosis- and envelope-based indexes for the diagnostic of rolling element bearings, Mechanical Systems and Signal Processing 43 (1–2) (2014) 25 – 43, ISSN 0888-3270.

[10] 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.

[11] Y. Pan, J. Chen, L. Guo, Robust bearing performance degradation assessment method based on improved wavelet packet–support vector data description, Mechanical Systems and Signal Processing 23 (3) (2009) 669 – 681, ISSN 0888-3270.

[12] D. Wang, Q. Miao, R. Kang, Robust health evaluation of gearbox subject to tooth failure with wavelet decomposition, Journal of Sound and Vibration 324 (3-5) (2009) 1141–1157.

[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] 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.

[16] 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.

[17] K. Ažman, J. Kocijan, Application of Gaussian processes for black-box modelling of biosystems, ISA Transactions 46 (4) (2007) 443–457.

[18] 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.

[19] B. Grašič, P. Mlakar, M. Z. Božnar, Ozone prediction based on neural networks and Gaussian processes, Nuovo Cimento C 29 (2006) 651–661.

[20] 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.

[21] 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.

[22] P. Nectoux, P. 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.

[23] 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.

[24] J. Antoni, Cyclostationarity by examples, Mechanical Systems and Signal Processing 23 (2009) 987–1036.

[25] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, Burlington, MA, 3rd edn., 2008.

[26] D. B. Percival, A. T. Walden, Wavelet Methods for Time Series Analysis, Cambridge University Press, Cambridge, 2000.

[27] 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.

[28] A. Rényi, On measures of information and entropy, in: 4th Berkeley Symposium on Mathematics, Statistics and Probability, 1960.

[29] D. Xu, D. Erdogmuns, Information Theoretic Learning, chap. Rényi's Entropy, Divergence and Their Nonparametric Estimators, Springer, 47–102, 2010.

[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] X. C. R. López-Ruiz, H.L. Mancini b, A statistical measure of complexity, Physics Letters A 209 (1995) 321–326.

[33] D. J. C. MacKay, Introduction to Gaussian processes, NATO ASI Series 168 (1998) 133–166.

[34] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.