Weak Monotonicity with Trend Analysis for Unsupervised Feature Evaluation

Lei Lu, Ying Tan, *Senior Member, IEEE*, Denny Oetomo, *Senior Member, IEEE*, Iven Mareels, *Fellow, IEEE*, and David A. Clifton

***Abstract*—Performance in an engineering system tends to degrade over time due to a variety of wearing or ageing processes. In supervisory controlled processes there are typically many signals being monitored that may help to characterise performance degradation. It is preferred to select the least amount of information to obtain high quality of predictive analysis from a large amount of collected data, in which labelling the data is not always feasible. To this end a novel unsupervised feature selection method, robust with respect to significant measurement disturbances, is proposed using the notion of "weak monotonicity" (WM). The robustness of this notion makes it very attractive to identify the common trend in the presence of measurement noises and population variation from the collected data. Based on WM, a novel suitability indicator is proposed to evaluate the performance of each feature. This new indicator is then used to select the key features that contribute to the WM of a family of processes when noises and variations among processes exist. In order to evaluate the performance of the proposed framework of the WM and suitability, a comparative study with other nine state-of-the-arts unsupervised feature evaluation and selection methods is carried out on well-known benchmark datasets. The results show a promising performance of the proposed framework on unsupervised feature evaluation in the presence of measurement noises and population variations.**

***Index Terms*—Weak monotonicity, suitability, unsupervised feature evaluation, robust feature selection.**

I. Introduction

MANY engineering systems naturally have monotonic characteristics (or trends) when ageing cannot be ignored, such as various types of wear encountered in industry from surface fatigue to corrosion and cracking [1], [2]. The existence of trends plays an important role in predicting future behaviours or predictive analysis [3], [4]. However, the trends may be not always visible from corrupted measurements. This work focuses on identifying key features that contribute to the trends of the system when noises and uncertainties exists.

The research was partially supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); and the Hong Kong Centre for Cerebro-Cardiovascular Health Engineering Limited.

L. Lu is with the Faculty of Engineering and Information Technology, The University of Melbourne, Parkville, VIC 3010, Australia; and the Department of Engineering Science, University of Oxford, Oxford, UK (Corresponding author: L. Lu; email: lei.lu@eng.ox.ac.uk).

Y. Tan and D. Oetomo are with the Faculty of Engineering and Information Technology, The University of Melbourne, Parkville, VIC 3010, Australia (emails: yingt@unimelb.edu.au; doetomo@unimelb.edu.au).

I. Mareels is with IBM Australia and New Zealand, Southbank VIC 3006, Australia (email: imareels@au1.ibm.com).

D. A. Clifton is with the Department of Engineering Science, University of Oxford, Oxford OX1 2JD, UK; and Oxford Suzhou Centre for Advanced Research, Suzhou, China (email: david.clifton@eng.ox.ac.uk)

Many existing techniques have been proposed to select key features from data measurements. Feature selection mainly focuses on choosing a small number of representative features from the original feature set [5]. It plays an important role in reducing dimensionality, and leads to an improved efficiency in modelling [6]. Various supervised learning algorithms have been proposed if the available data set is labeled, such as the Fisher criterion [7], the Pearson correlation [8], and feature selection with the sequential forward and sequential backward method [9].

It is highlighted that feature selection is different from feature learning, which learns features from datasets with transformation or latent learning. There are many used feature learning methods include the convolutional neural network [10], long short-term memory [11], and manifold learning [12]. The extracted features from these methods are a combination or transformation of the datasets, making the results difficult to be interpreted. On the contrary, feature selection usually identifies features based on some cost functions, which is usually application-driven.

When the dataset is not able to be labeled, unsupervised learning algorithm is a suitable choice. Unsupervised feature selection is widely encountered [13], as labelling the data requires extra efforts, and it is even not feasible to label the data in some cases, i.e., exploratory data analysis for initial investigations. Existing unsupervised feature selection techniques usually evaluate features by defining some cost functions, such as data similarity [8], and local discriminative information [14]; Then, features are selected by measuring the cost with their corresponding ranking. As unsupervised learning schemes are data-driven, the quality of data will greatly affect the model performance, and sometimes lead to unaccepted performance as indicated in [13]. In terms of selecting representative features for prediction analysis, it is assumed that adding trend information into the feature selection process will help to improve model performance on selecting key features for the prediction.

This work proposes a novel concept of weak monotonicity to add robust trends to unsupervised feature selection techniques, which tries to keep the trend in the presence of noises and uncertainties. With the concept of WM, it is possible to characterise the trend using statistical properties derived from a family of processes coming from the same population. Once the trend of a process is captured by the WM, it can be served as some inherent "model" information. Such model information can be used to define a cost to evaluate the importance of different features, and unsupervised select representative

features that contribute to the trend. Because the WM uses statistics calculated from a family of processes to define the cost, the unsupervised feature selection model is expected to have a better generalisation ability in identifying key features, and robust to measurement noises and disturbances.

The concept of monotonicity used for trend analysis is usually mathematically modelled with strong assumptions as shown in [15], [16]. In engineering applications, these analysis techniques are not directly applicable due to difficulty in checking assumptions. Some research utilises the concept of strict monotonicity (SM) to describe processes with clear trends, and evaluates features for practical applications [4], [17]. When the trends are corrupted by noises or variations among the family of processes, the SM cannot always work well. There are initial attempts to employ the concept of WM in biomedical engineering applications [18], however, the method developed in [18] uses a conservative fixed bound as *a priori*. Therefore, more robust measures with respect to noises and variations need further investigations.

Similarity is known as an important index to characterise the common behaviour when a family of processes with similar trends are considered [8], [18]. With leveraging the WM and the similarity measures, similar to standard unsupervised feature selection technique, a new indicator in terms of suitability is introduced to evaluate features. Such a robust measure can be used to capture the common trend for a family of processes with noises and variations among population, and with applications in unsupervised feature selection.

The contributions of this paper are summarised as follows.

1) A new concept of WM is proposed to describe the trend of a process with consideration of measurement noises and variations.
2) Along with the WM, the similarity measure is used to capture the common trend of a family of processes.
3) A new algorithm is developed to systematically estimate the uncertainty or noise level in a family of processes.
4) A novel suitability indicator (cost) is developed to evaluate and identify representative features for common trends existing in a family of processes.
5) Extensive comparative studies with nine state-of-the-art unsupervised feature selection methods using the well-known datasets are presented to show the effectiveness of the proposed method.

The remainder of this paper is organised as follows. Section II presents the basic concepts of SM and WM. Section III discusses unsupervised feature evaluation with a new cost of suitability. The effectiveness of the proposed method is illustrated by comparisons with the existing methods on well-known datasets in Section IV, and the discussion is presented in Section V. Finally, conclusions are drawn in Section VI.

II. Monotonicity and Weak Monotonicity

This section introduces the concept of a monotonicity signal and its extension: a weak monotonicity (WM) to capture the trend of a signal.

A. Notations

The set of real numbers is denoted as $\mathbb{R}$, and the set of natural numbers is denoted as $\mathcal{N}$. Let $\{\Sigma^j\}_{j=1,2,\cdots,M}$ be a family of M processes. The notation $\mathbf{F} \in \mathbb{R}^{N_p \times m \times M}$ denotes the matrix coming from the feature set for the M processes with N_p rows and m columns. The vector $\mathbf{x}_i^j$ denotes the i^{th} feature vector derived from the j^{th} process. Each element of $\mathbf{x}_i^j$ contains the feature value at time step k of the j^{th} process, which is denoted as $x_{k,i}^j \in \mathbb{R}, k = 1, 2, \cdots, N_p$, with N_p denoting the duration of the process. The set $\mathcal{F} \subset \mathbb{R}^{N_p \times m}$ contains m features with time duration N_p.

B. Strict Monotonicity (SM)

Firstly, we introduce the concept of the SM for a sequence of measurements of a scalar valued signal (it is seen from the following text, but we may as well state it up front), z_k is a real scalar valued variable, measured at time t_k, k is an integer counting the events. Without abuse of notion, the pair $\{t_k, z_k\}_{k\in\mathcal{N}}$ is sometimes used to present this signal.

Definition 1: A signal $\{t_k, z_k\}_{k\in\mathcal{N}}$ is called strictly monotonically increasing if the following condition holds:

$$\begin{aligned} t_{k+1} &> t_k,\ t_k \xrightarrow{k\to\infty} \infty \\ z_{k+1} &> z_k, k = 1, 2, \cdots. \end{aligned} \tag{1}$$

where, k is the index of sampling instant. ∘

It is usually hard to find strictly monotonically increasing or decreasing signals in engineering applications. Most signals have a trend of monotonicity. In order to capture such a trend, the following measure Mo is proposed for a finite duration signals $\{t_k, z_k\}_{k=1,2,\cdots,N_p}$ [4].

$$Mo = \frac{n^+}{N_p - 1} - \frac{n^-}{N_p - 1}, \tag{2}$$

where, n^+ is the number of points with larger values than the previous instants along N_p points of the signal z_k, which can be calculated as,

$$n^+ = \sum_{k=1}^{N_p} Num_k, \tag{3}$$

and Num_k is the monotonically increasing point,

$$Num_{k+1} = \begin{cases} 1 & \text{if } z_{k+1} > z_k \\ 0 & else \end{cases} \quad k = 0, \cdots, N_p - 1. \tag{4}$$

Similar, we can obtain the number of monotonically decreasing points n^- as,

$$n^- = \sum_{k=1}^{N_p} \overline{Num}_k. \tag{5}$$

and the monotonically increasing point Num_k is defined as,

$$\overline{Num}_{k+1} = \begin{cases} 1 & \text{if } z_{k+1} < z_k \\ 0 & else \end{cases} \quad k = 0, \cdots, N_p - 1. \tag{6}$$

Remark 1: It is worthwhile to highlight that we can use the SM to describe a trend of a monotonically decreasing signal as well. It is noted that if the signal $\{z_k\}_{k=1,2,\cdots,N_p}$ is monotonically increasing (decreasing), then $Mo = 1$ ($Mo = -1$).

If $Mo > 0$, the signal $\{t_k, z_k\}_{k=1,2,\cdots,N_p}$ has a trend of monotonically increasing, while $Mo < 0$ indicates a trend of monotonically decreasing in a signal. ○

Due to the existences of noises in sensor measurements, sometimes, the SM trend is not clear from the measured signal. In order to capture the existing trend with the consideration of noises, the notion of weak monotonicity is introduced next.

C. Weak Monotonicity (WM)

Definition 2: A signal $\{t_k, z_k\}_{k\in\mathcal{N}}$ is called weakly monotonically increasing if there exists a constant $\Delta > 0$ such that the following condition holds:

$$\begin{aligned} t_{k+1} &> t_k,\ t_k \xrightarrow{k\to\infty} \infty \\ z_{k+1} &\geq z_k - \delta_k, k = 1, 2, \cdots, \end{aligned} \tag{7}$$

where, δ_k is a random zero mean variable with either a positive or negative value bounded by $\max_{k\in\mathcal{N}} |\delta_k| \leq \Delta$. ○

Remark 2: The sequence $\{\delta_k\}_{k=1,2,\cdots,}$ represents the noise and uncertainty at each measurement, and it affects the calculation of the monotonicity measure. There also may have uncertainty in the measurement time instants. However, the time uncertainties are typically small comparing to δ_k, and of course do not affect the trend in the signal. Therefore, the current study only models the value measurement uncertainty.

The following measure WMo is used to check the WM (increasing) trend of a finite duration of this signal,

$$WMo = \frac{An^+}{N_p - 1} - \frac{An^-}{N_p - 1}, \tag{8}$$

where, An^+ is the number of increasing points in the sense of weak monotonicity,

$$An^+ = \sum_{k=1}^{N_p} ANum_k, \tag{9}$$

and the weak monotonically increasing point $ANum_k$ is defined as,

$$ANum_{k+1} = \begin{cases} 1 & \text{if } z_{k+1} \geq z_k - \delta_k \\ 0 & else \end{cases}, \tag{10}$$

Then, the number of weak monotonically decreasing points An^- can be calculated in the similar way,

$$An^- = \sum_{k=1}^{N_p} \overline{ANum}_k, \tag{11}$$

where, the weak monotonically decreasing point $\overline{ANum}_k$ is defined as,

$$\overline{ANum}_{k+1} = \begin{cases} 1 & \text{if } z_{k+1} < z_k - \delta_k. \\ 0 & else \end{cases} \tag{12}$$

Remark 3: Both Mo and WMo are in the range of $[-1, 1]$. When they are close to 1 or -1, they indicate strong monotonically increasing or decreasing trend of a signal. ○

Remark 4: In general, it is very hard to find the bound of noises for a single signal or process. Even if there exists such a bound, it is usually conservative. In practice, it is possible to estimate such variations or uncertainties using some statistic properties of the measurements. When a group of signals or processes are considered, Section III-B proposes a feasible solution to estimate variations among a family of processes, and effectiveness of this solution is validated by our testing on benchmark datasets. ○

As an illustrative example of how to compute the Mo and WMo of a signal, Figure 1 demonstrates three signals J^1, J^2, and J^3 with underlying trends of monotonicity, but affected by measurement noises. For simplicity, the variation of each signal is set as 10% of the current value of each data point.

It is noted that both the SM and the WM can capture the trend of a signal, while the SM is sensitive to measurement uncertainty. The WM introduces a bound for the signal and allows each data point to have a certain degree of variation. The calculation of WM can robustly represent the relationship between data points and capture the underlying trend of a signal. Take points A and B in the J^1 trajectory for example, the value of point A is smaller than its previous point, then it is regarded as a negative point (-) when computing Mo. On the other hand, if the value of point A is larger than the lower bound of its previous point, then it can be regarded as a positive point (+) in order to compute WMo. It can be seen that the value at point B in the J^1 signal is smaller than the lower bound of its previous point, then it is treated as a negative point (-) in computing both Mo and WMo.

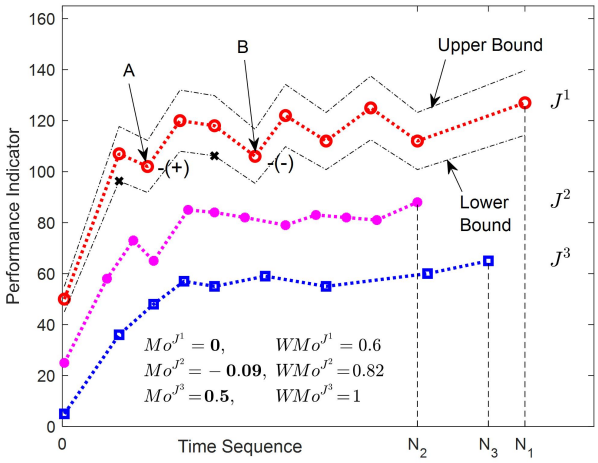

Fig. 1: Illustrations of signals with SM and WM characteristics

The three signals in Figure 1 have weak monotonically increasing patterns. We first calculate the monotonicity measure for each signal according to Eq. (2), and have $Mo^{J^1} = 0$, $Mo^{J^2} = -0.09$, and $Mo^{J^3} = 0.5$, it is clear that the $Mo^{J^1} = 0$ does not show the trend correctly. While by the concept of WM, we have $WMo^{J^1} = 0.6$, $WMo^{J^2} = 0.82$ and $WMo^{J^3} = 1$. Comparing with SM, the WM can better capture the trend of a process with variations. It is also robust to noises as shown in the experimental study in Section IV.

The concepts of SM and WM can be used in various applications to characterise the trend of a family of signals (processes). For example, it can be used to detect implicit degradation processes such as the early-stage mechanical wear in bearings. In order to demonstrate the effectiveness of these concepts, they are used in the challenging problem of

unsupervised feature evaluation and selection for a family of similar processes with some common trend. In other words, both SM and WM can be used to select significant features, to which contribute the trend or monotonicity for a family of processes. It is worthwhile to highlight that due to variations among the population (within the family of processes), the robust analysis will be more complicated than a single process.

III. Unsupervised Feature Evaluation

Let us consider a family of processes that share the common trend such as the flank wear of a set of cutters in a milling process, and the wear of bearings in machines. The focus of this work is to identify features that can clearly represent the trend of these processes using noisy measurements of sensors. Such features may play important roles to monitor the performance of the systems and detect faults of the processes or failures of the sensors.

As it is not easy to label a large amount of data for industrial processes, it is desirable to use unsupervised feature selection. This section discusses how to use the concept of WM to assist unsupervised feature selection and pick up the key features of a family of processes with a common trend.

A. Problem Formulation

Given a family of processes $\{\Sigma^j\}_{j=1,2,\cdots,M}$ having some common trends with a large set of signals $\Sigma^j : \left\{t_k, \mathbf{x}_k^j\right\}_{k\in\mathcal{N}}$ measured from these processes. Here $\mathbf{x}_k^j$ consists of elements $x_{k,i}^j$ representing the i^{th} feature of the sensor measurement for the j^{th} process at the k^{th} sampling instant.

These signals have the following properties:

$$\begin{aligned} t_{k+1} &> t_k,\ t_k \xrightarrow{k\to\infty} \infty, \\ x_{k+1,i}^j &\le \rho_i^j x_{k,i}^j + \delta_{k,i}^j, \quad k = 0,1,\cdots, i = 1,\cdots,m, \\ & \qquad j = 1,\cdots,M, \end{aligned} \tag{13}$$

for some positive constants $\rho_i^j \in (0,1)$ with $\max\limits_{k\in\mathcal{N},i=1,\cdots,m,j=1,\cdots,M} \left|\delta_{k,i}^j\right| \le \Delta$, where Δ is a positive constant to bound the size of noises or uncertainties.

Remark 5: It is noted that the damping ratio ρ_i^j is introduced in order to simplify the calculations when noises exist. The inequality (13) is a special case of the inequality (7). In our analysis, the information of ρ_i^j is not used. ∘

It can be seen from the inequality (13) that there is a common trend (strict monotonicity) in each feature when the uncertainty $\delta_{k,i}^j = 0$, and the speed of convergence ρ_i^j depends on the feature and the process. More precisely, each feature demonstrates strictly decreasing (see *Definition 1*) when there are no measurement noises, variations, and other unmeasured uncertainties. With the consideration of disturbances, the processes are weakly monotonic (see *Definition 2*).

If such disturbances $\delta_{k,i}^j$ are too large, the trend of the j^{th} process or the family of the processes will disappear. In general, there are many features obtained from sensors, i.e., $m >> 1$. The challenge is to pick up significant features from a large set of features that are less sensitive to noises, disturbances, and variations so that the common trend of a family processes $\{\Sigma_j\}_{j=1,\cdots,M}$ can still be captured.

Hence the objective of this paper is to identify key features that can capture a common trend of a family of processes $\{\Sigma_j\}_{j=1,\cdots,M}$ in the presence of noises, disturbances, and variations.

B. Estimation of Subject Variation

As discussed in Section II-C, it is difficult to accurately describe noises and uncertainties for each process due to complicated industry applications. Different methods can be used to quantify noise levels, e.g., the sample entropy [19]. The current study focuses on estimating uncertainty from a family of processes, and uses statistical indices to describe variations or uncertainties for each feature in the process.

Keeping in mind, the goal of the feature evaluation and selection is to pick up the features that are less sensitive to noises or variations. Therefore, the variation of the feature among the population (the family of processes) will be used to approximate the bound of noises, uncertainties, and variations.

Given a family of processes with finite duration, $\{\Sigma^j\}$, $j = 1,2,\cdots,M$, the processes can be represented by features derived from sensor measurements, $\left\{t_{k,i}^j, x_{k,i}^j\right\}_{k=1,2,\cdots,N_p}$, where $N_p \in \mathcal{N}$ is the length of each process , $i = 1,2,\cdots,m$ is the feature index. It is noted that in the problem formulation of Section III-A, the upper bound Δ is defined for the largest bound over all features, however, such a bound is always conservative. In order to be less conservative, it is assumed that

$$\max_{k\in\mathcal{N},j=1,\cdots,M} \left|\delta_{k,i}^j\right| \le \Delta_i \cdot \left|x_{k,i}^j\right|, \tag{14}$$

for some positive Δ_i. Next, we estimate the Δ_i for the i^{th} feature.

First, the standard deviation of the i^{th} feature across population can be calculated as,

$$\sigma_{k,i} = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left|x_{k,i}^j - \frac{1}{M}\sum_{j=1}^{M} x_{k,i}^j\right|^2}. \tag{15}$$

Then, one estimation of variation $\tilde{\Delta}_i$ for the i^{th} feature can be calculated as the mean value of standard deviation of the sequence as following,

$$\tilde{\Delta}_i = \frac{1}{N_p}\sum_{k=1}^{N_p}\sigma_{k,i}. \tag{16}$$

It should be noted that the estimation assumes the data size of processes are equal, or they have been truncated with the minimal length across the population, i.e., $N_p = N_p^{min} = \min\left\{N_p^{J_{s_i}}\right\}, J_{s_i} = 1,2,\cdots,M$. However, as demonstrated in Figure 1, the processes have different data lengths. With the minimal length truncation $N_p^{min} = N_2$ for J^1, J^2, and J^3, the information beyond the N_p^{min} point in the J^1 and J^3 signals will be lost.

In order to better utilise the data from sensors, an *one-leave* strategy is developed to reduce the waste of data due to

truncation. The strategy kicks off one process from the family each time, and calculate the approximate variation $\tilde{\Delta}_i$ with the standard truncated processes. As illustrated in Figure 1, three approximate variations can be calculated from $\mathcal{G}_1 = \{J^2, J^3\}, \mathcal{G}_2 = \{J^1, J^3\}$, and $\mathcal{G}_3 = \{J^1, J^2\}$, then the variation can be updated as the mean value of the three approximate variations. With this strategy, an updated bound estimation of noises, uncertainties, and variations becomes,

$$\begin{cases} \tilde{\Delta}_i^j = \dfrac{1}{M-1} \sum\limits_{s=1, s\neq j}^{M-1} \tilde{\Delta}_i^{j,s}, \\ \overline{\Delta}_i = \dfrac{1}{M} \sum\limits_{j=1}^{M} \tilde{\Delta}_i^j, \end{cases} \tag{17}$$

where, $\overline{\Delta}_i$ is the updated approximate variation, and $\tilde{\Delta}_i^{j,s}$ is calculated by Eq. (16) when the j^{th} process is deleted from the family. This leads to the following estimation,

$$\begin{cases} \hat{\Delta}_i = \beta_{1,\alpha_1} + \eta_i(\overline{\Delta}_i - \alpha_1(\tilde{\Delta}_i)), \\ \alpha_1(\tilde{\Delta}_i) = \min\limits_{j=1,\cdots,M} \tilde{\Delta}_i^j, \quad \alpha_2(\tilde{\Delta}_i) = \max\limits_{j=1,\cdots,M} \tilde{\Delta}_i^j, \\ \beta_{1,\alpha_1} = \min\{\lambda_{low}, \alpha_1(\tilde{\Delta}_i)\}, \\ \beta_{2,\alpha_2} = \min\{\lambda_{up}, \alpha_2(\tilde{\Delta}_i)/3\}, \\ \eta_i = \dfrac{\beta_{2,\alpha_2} - \beta_{1,\alpha_1}}{\alpha_2(\tilde{\Delta}_i) - \alpha_1(\tilde{\Delta}_i)}, \end{cases} \tag{18}$$

where, $\hat{\Delta}_i$ is the estimated uncertainty bound for the i^{th} feature, $\alpha_1(\tilde{\Delta}_i)$ and $\alpha_2(\tilde{\Delta}_i)$ are the calculated statistical indices, β_{1,α_1} and β_{2,α_2} are two bounds constrained by α and the regularisation parameter λ. The lower bound of the feature is regularised by λ_{low}, and it can be set as a small positive constant. While λ_{up} is used to regularise the upper bound, and can be tuned by trials-and-errors, our simulation results suggest that usually it can be set as 0.1 for signals with low noise level, and 0.15 for signals with high background noise. A detailed investigation of selection the two regularisation parameters are discussed in Section IV-C.

The pseudo code of the proposed estimation algorithm is presented in Algorithm 1. It is noted that when each process has the same length of measurements, such an algorithm can be simplified.

Algorithm 1: Estimation of the bound of noises, uncertainties, and variations for a feature.

Input : Feature matrix $\mathbf{F} \in \mathbb{R}^{N_p \times m \times M}$, $\lambda_{low}, \lambda_{up}$;
Output: Variation bound estimation $\Delta_i, i = 1, \cdots, m$.

for $i \leftarrow 1$ **to** m **do**
  for $j \leftarrow 1$ **to** M **do**
    Let $\tilde{\mathbf{F}} = \mathbf{F}$, delete the j^{th} process $\tilde{\mathbf{F}} \leftarrow \tilde{\mathbf{F}}^j = \emptyset$;
    for $j' \leftarrow 1$ **to** $M-1$ **do**
      Truncate $\tilde{\mathbf{F}}$ with the minimal length N_p^{min} ;
      Update matrix $\mathbf{F}' = \{x_{k,i}^{j'}\}, k = 1, 2, \cdots, N_p^{min}$;
      Standard deviation $\sigma_{k,i} = std(\mathbf{F}')$; $\tilde{\Delta}_i^{j,j'} = mean(\sigma_{k,i})$;
    end
    Calculate variation $\tilde{\Delta}_i^j = mean(\tilde{\Delta}_i^{j,j'})$;
  end
  Update variation $\overline{\Delta}_i = mean(\tilde{\Delta}_i^j)$;
  Compute $\alpha_1(\tilde{\Delta}_i) = min(\tilde{\Delta}_i^j)$, $\alpha_2(\tilde{\Delta}_i) = max(\tilde{\Delta}_i^j)$;
end
for $i \leftarrow 1$ **to** m **do**
  Regularize β_{1,α_1} with λ_{low}; and β_{2,α_2} with λ_{up};
  Compute the parameter η_i;
  Update variation $\hat{\Delta}_i = \beta_{1,\alpha_1} + \eta_i(\overline{\Delta}_i - \alpha_1(\tilde{\Delta}_i))$.
end

C. Similarity between Processes

As it is assumed that a family of processes sharing a similar trend, we use the similarity to characterise the common trend between processes. That is, for a family of processes $\Sigma^j : \left\{t_k, \mathbf{x}_k^j\right\}$ represented by a set of features or signals $x_{k,i}^j$, $k = 1, \cdots, N_p^{J_{s_i}}$ and $N_p^{J_{s_i}} \in \mathcal{N}$, the similarity between two processes is defined as following,

$$\begin{aligned} Tr_i^{J_{s_1}, J_{s_2}} &= Corr\left(x_{k,i}^{J_{s_1}}, x_{k,i}^{J_{s_2}}\right), \\ & k = 1, 2, \cdots, \min\left\{N_p^{J_{s_1}}, N_p^{J_{s_2}}\right\}, \end{aligned} \tag{19}$$

where, $Tr_i^{J_{s_1}, J_{s_2}}$ represents the similarity between processes $j = J_{s_1}$ and $j = J_{s_2}$, $s_1, s_2 = 1, 2, \cdots, N, s_1 \neq s_2$. $Corr(\phi_k, \psi_k)$ computes the correlation between sequences $\{\phi_k\}_{k=1,\cdots,N_1}$ and $\{\psi_k\}_{k=1,\cdots,N_1}$ for some $N_1 \in \mathcal{N}$.

The correlation $Corr(\phi_k, \psi_k)$ can be calculated with different methods. In the current study, we use the Pearson's correlation for the computation,

$$Corr(\phi_k, \psi_k) = \frac{\sum\limits_{k=1}^{N_1} (\phi_k - \bar{\phi})(\psi_k - \bar{\psi})}{\left(\sum\limits_{k=1}^{N_1} (\phi_k - \bar{\phi})^2\right)^{1/2} \left(\sum\limits_{k=1}^{N_1} (\psi_k - \bar{\psi})^2\right)^{1/2}}, \tag{20}$$

where, $\bar{\phi} = \frac{1}{N_1} \sum\limits_{k=1}^{N_1} \phi_k$ and $\bar{\psi} = \frac{1}{N_1} \sum\limits_{k=1}^{N_1} \psi_k$ are means of two truncated sequences respectively.

D. Suitability Evaluation

After calculating the WM and the similarity between processes represented by each feature, a new cost is proposed to evaluate the suitability of each feature in terms of characterising the common trend of the family of processes. The suitability of the i^{th} feature is defined as

$$\begin{aligned} S_{WMo_i} &= \frac{2}{M(M-1)} \sum_{s_i=1}^{M-1} \sum_{s_j=s_i}^{M} Am_i^{J_{s_i}, J_{s_j}}, \\ Am_i^{J_{s_i}, J_{s_j}} &= \left(\omega_i^{J_{s_i}} WMo_i^{J_{s_i}} + \omega_i^{J_{s_j}} WMo_i^{J_{s_j}}\right) Tr_i^{J_{s_i}, J_{s_j}}, \end{aligned} \tag{21}$$

where, $WMo_i^{J_{s_j}}$ is the calculated WM of the i^{th} feature for the J_{s_j} process. Here $\omega_i \in \{-1, 0, 1\}$ is the trend indicator,

it takes 1 for an increasing trend, -1 for a decreasing trend, and 0 indicates constant. $Tr_i^{J_{s_i},J_{s_j}}$ is the similarity between the J_{s_i} and J_{s_j} processes.

With calculating the suitability for each feature, the representative feature subset $\mathcal{I} \subset \mathbb{R}^{N_p \times l}, 1 \leq l \leq m$, can be identified with the following optimisation problem,

$$\begin{aligned} \min_{\mathcal{I} \subset \mathbb{R}^{N_p \times l}} & -\sum_{i \in \mathcal{I}} S_{WMo_i} \\ & = -\frac{2}{M(M-1)} \sum_{i \in \mathcal{I}} \sum_{s_i=1}^{M-1} \sum_{s_j=s_i}^{M} Am_i^{J_{s_i},J_{s_j}}, \\ s.t. \;\; & \mathcal{I} \subset \mathcal{F} \subset \mathbb{R}^{N_p \times m}, \; 1 \leq l \leq m. \end{aligned} \tag{22}$$

It is worth noting that the proposed feature evaluation method is unsupervised, as there is no label information involved in the computing procedure. It only uses properties of the feature to define the evaluation criterion. The pseudo code of the proposed unsupervised feature evaluation method is detailed in Algorithm 2.

Remark 6: The suitability in Eq. (21) is defined with the WM. It also can be defined with the SM. That is, for the i^{th} feature, it can be calculated as

$$\begin{aligned} S_{Mo_i} &= \frac{2}{M(M-1)} \sum_{s_i=1}^{M-1} \sum_{s_j=s_i}^{M} m_i^{J_{s_i},J_{s_j}}, \\ m_i^{J_{s_i},J_{s_j}} &= \left(\omega_i^{J_{s_i}} Mo_i^{J_{s_i}} + \omega_i^{J_{s_j}} Mo_i^{J_{s_j}}\right) Tr_i^{J_{s_i},J_{s_j}}, \end{aligned} \tag{23}$$

where, Mo_i^j is the monotonicity for the i^{th} feature of the j^{th} process (see *Definition 1*). And $\omega_i \in \{-1, 0, 1\}$ is the trend indicator. This leads to the similar optimisation problem,

$$\begin{aligned} \min_{\mathcal{I} \subset \mathbb{R}^{N_p \times l}} & -\sum_{i \in \mathcal{I}} S_{Mo_i} \\ & = -\frac{2}{M(M-1)} \sum_{i \in \mathcal{I}} \sum_{s_i=1}^{M-1} \sum_{s_j=s_i}^{M} m_i^{J_{s_i},J_{s_j}}, \\ s.t. \;\; & \mathcal{I} \subset \mathcal{F} \subset \mathbb{R}^{N_p \times m}, \; 1 \leq l \leq m. \end{aligned} \tag{24}$$

for the SM based feature evaluation and selection. This method is used in later performance comparison as well. ◦

IV. Experiments and Results

The proposed unsupervised feature evaluation method aims to identify representative features from a large datasets with evaluating their suitability for a family of processes with a common trend. In this section, appropriate datasets that have common trends are used. In particular, the NASA Ames Prognostics Data Repository is a collection of datasets that were extensively used for regression tasks [20], [21]. Most of these datasets contain time sequences from a nominal state to a failed state, and utilise several sensors, e.g., vibration and temperature, to measure the process with discrete time steps.

Algorithm 2: Unsupervised Feature Evaluation based on the WM.

Input : Feature matrix $\mathbf{F} \in \mathbb{R}^{N_p \times m \times M}, \omega_i, \Delta_i$;
Output: Representative subset $\mathcal{I} \subset \mathbb{R}^{N_p \times l}, 1 \leq l \leq m$.

Calculate Weak monotonicity WMo_i
for $j \leftarrow 1$ **to** M **do**
  Let $\mathbf{X}^j$ denotes the j^{th} matrix from $\mathbf{F}$;
  for $i \leftarrow 1$ **to** m **do**
    Estimate ω_i^j with linear fitting method;
    for $k \leftarrow 2$ **to** N_p **do**
      Calculate the estimation $\delta_{k,i}^j = \Delta_i \cdot |x_{k,i}^j|$;
      Compute An^+ and An^- with Eq. (12);
    end
    Calculate WMo_i^j with An^+ and An^-;
  end
end
Compute Similarity $Tr_i^{J_{s1},J_{s2}}$
for $i \leftarrow 1$ **to** m **do**
  Truncate $\mathbf{x}_i^{J_{s1}}$ and $\mathbf{x}_i^{J_{s2}}$;
  Calculate the correlation $Tr_i^{J_{s1},J_{s2}}$;
end
Evaluate features with suitability S_{WMo_i}:
for $i \leftarrow 1$ **to** m **do**
  for $J_{s1} \leftarrow 1$ **to** $M-1$ **do**
    for $J_{s2} \leftarrow J_{s1}+1$ **to** M **do**
      Calculate $Am_i^{J_{s1},J_{s2}}$ according to Eq. (21)
    end
  end
  Obtain S_{WMo_i} with the mean value of $Am_i^{J_{s1},J_{s2}}$;
end
Select features with top l suitability values for set $\mathcal{I}$.

A. Datasets and Signal Processing

The current study employs the widely used Milling Datasets (MDSCI) [20] and IMS Bearing Datasets (IMS) [21] to verify effectiveness of the proposed method. Detailed description of the two datasets can be found in the supplementary document. Figure 2 presents signals of the run-to-failure experiments in the two datasets. As shown in Figure 2, each experiment has some clear trend of monotonicity, corrupted by noises. The first experiment has fast performance degradation processes. The second experiment has slow degradation processes, in which the SM only appears clearly at the end of each process. Statistical features extracted from the sensor measurements are used as signals or processes for unsupervised feature evaluation and selection. In order to evaluate the performance of unsupervised feature evaluation and selection, the selected features are used to predict the trend of these processes.

Generally, various features can be extracted from the time domain, frequency domain, and time-frequency domain [9], [22]. Among them, time-frequency techniques have high resolutions and have been widely used for signal analysis [9]. The study uses the wavelet packet transform to decompose the signal with the 8^{th} Daubechies wavelet function into 4 levels, thus total 16 components can be obtained from each sample

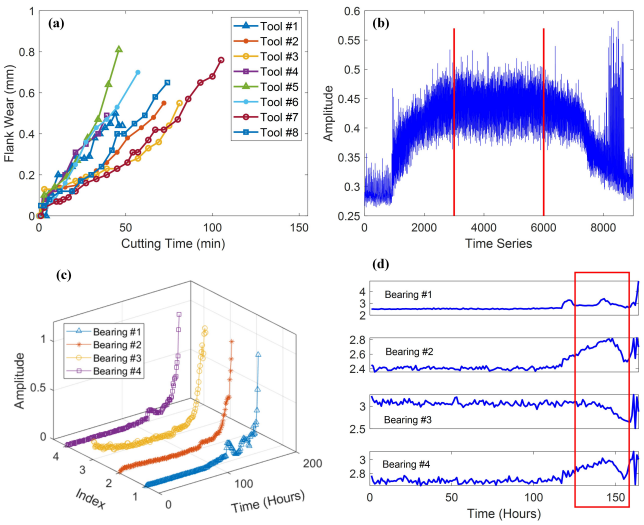

Fig. 2: Experiment information. (a) Flank wear values of cutting tools, (b) one data recording of vibration signal for a cutting tool, (c) bearing test, (d) kurtosis feature for four bearings.

[22]. Then, 10 widely used statistical features are extracted from each decomposed component, including the root mean square, standard deviation, kurtosis, peak-peak value, crest factor, clearance, impulse factor, shape factor, average energy, and the fifth central moment. More detailed descriptions of these features (processes) can be found in [23] and references therein. Therefore, total 160 features are extracted from each data recording for both datasets.

It is noted that both two datasets have some trends, when statistic features are used, these features may be not able to reflect the trend efficiently, as an illustration of the kurtosis feature as shown in Figure 2(d). Although kurtosis feature has been widely used to present many industrial processes, Figure 2(d) shows that the kurtosis feature of the second datasets is not able to capture the trend of the processes due to existence of large variations. Therefore, this feature cannot be selected to represent the common trend.

After the signal processing, the proposed unsupervised feature evaluation and selection can be applied.

B. Performance Evaluation Method

To verify the effectiveness of the selected representative features, support vector regression (SVR) from LibSVM is used for prediction analysis with the identified features [24]. The first 70% of the data is used for model training, and the rest 30% is used for model validation.

For the prediction analysis, it is assumed that the current health state of the system is a function of features at the current step, features of the previous step, and the output at the previous step [25]. The mapping between the model input and output can be described as,

$$\mathcal{HI}_k = f(\mathbf{X}_k, \mathbf{X}_{k-1}, y_{k-1}), \qquad k = 2, 3, \ldots, N_p, \tag{25}$$

where, $\mathbf{X}_k$ and $\mathbf{X}_{k-1}$ are matrices of identified key features at the k^{th} step and the $(k-1)^{th}$ step. The notion of $\mathcal{HI}_k$ is the current health state, y_{k-1} is the performance degradation value at the $(k-1)^{th}$ step, and they are detailed as follows.

For the MDSCI datasets, the flank wear value (VB) is measured to indicate the health condition of cutting tools [20], then the performance degradation value y_{k-1} is the flank wear value, and $\mathcal{HI}_k$ is the predicted flank wear at the k^{th} step.

For the IMS datasets, the ground truth wearing value can not be measured for bearings with continues running, while as indicated in [26], the exponential fitting with the form of $a + be^{ck}$ is demonstrated to be suitable to describe the performance degradation process. Then, the performance degradation value y_{k-1} in Eq. (25) for the IMS datasets is coming from the fitted curve, and the $\mathcal{HI}_k$ is the remaining useful life (RUL) ranging from 0 to 1, with 1 indicating the beginning health condition and 0 indicating the wear out condition.

For the training of the standard SVR model, the radial basis function (RBF) is used as the kernel. To make the prediction fair, parameters of the SVR model are optimised in a large range $2^K, K = \{-8 : 8\}$ with greedy searching strategy [27]. Three widely used evaluation criteria are used to measure the prediction performance, including the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean square error (RMSE).

C. Tuning Parameters

As discussed in Section III-B, the estimation of uncertainty bound for each feature is depended on two parts: the variation calculated from a family of processes, and two parameters λ_{low} and λ_{up} to regularise the estimations. The role of λ_{low} is used to define a lower bound of the estimation, while λ_{up} is an estimation of the upper bound. These two bounds usually come from the knowledge of the processes (case-dependent). This work discusses how the performance of prediction errors in terms of MAE would be affected by the two tuning parameters.

Figure 3(a) shows the prediction performance on the MDSCI datasets as λ_{low} increases from 10^{-6} to 0.05 with different numbers of features selected, when λ_{up} is fixed at 0.1. Similar performances for the IMS datasets can be found in Figure 3(d). It can be seen from Figure 3(a) that the curves of prediction errors have very similar trends with increasing λ_{low} from 10^{-6} to 0.05 for the MDSCI datasets, except the prediction error has slightly large value with 10 features when the value of λ_{low} is 0.05. While it can be clearly seen from Figure 3(d) that the model has large prediction errors of 1.952 for the λ_{low} with 0.01 and 0.05.

The impact of the parameter λ_{up} on the feature selection performance is shown as Figure 3(b) and (e), respectively. The value of λ_{up} gradually changes from 0.01 to 0.2, when the λ_{low} is fixed as 10^{-4}. Figure 3(b) shows the prediction errors on the MDSCI datasets when changing the λ_{up}, and it indicates that the model has smaller prediction errors when λ_{up} is larger than 0.1. Figure 3(e) demonstrates the prediction performance on the IMS datasets, which shows that the model has worse prediction performance for λ_{up} taking values of 0.01 and 0.05. These simulations suggest that a larger λ_{up} is preferred to be larger than 0.1.

To have a better understanding of the impact of the parameters, Figures 3(c) and (f) provide the three dimension mapping

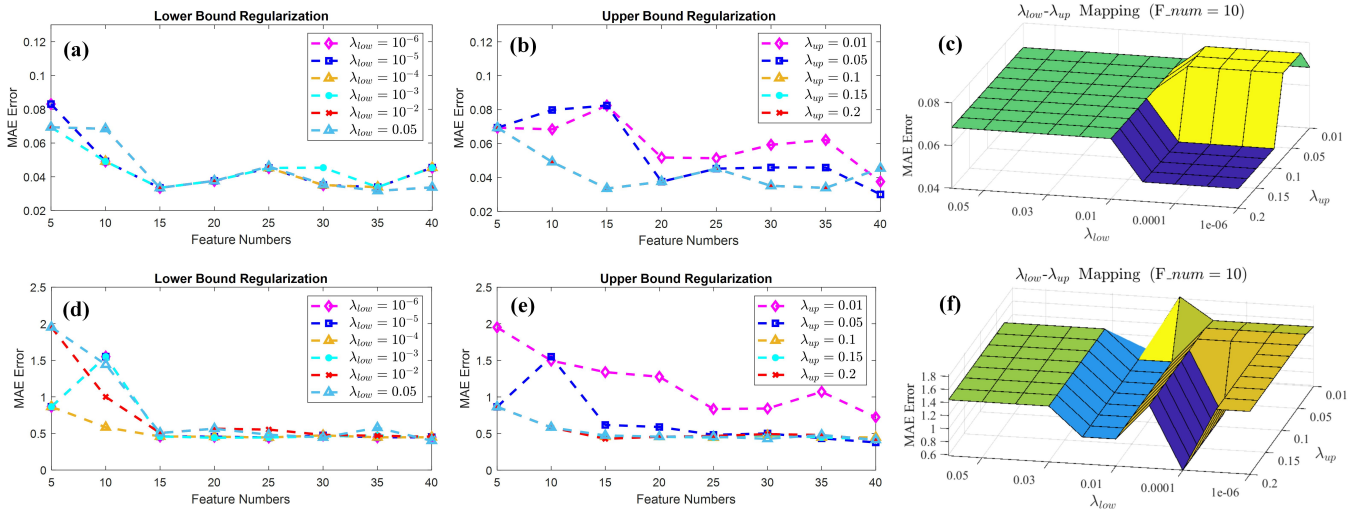

Fig. 3: Impact of regularisation parameters with respect to the MAE error. (a)-(c) Impact of λ_{low} and λ_{up} parameters, and the mapping of λ_{low}-λ_{up} on the MDSCI datasets, (d)-(f) Impact of λ_{low} and λ_{up} parameters, and the mapping of λ_{low}-λ_{up} on the IMS datasets.

of λ_{low} and λ_{up} when the feature number equals to 10. It can be seen from the figures that the model has high prediction errors with a large λ_{low} value. Therefore, the λ_{low} should take a low value to calculate the variations. It is similar for the parameter of λ_{up}, the model has large prediction errors with a small value of λ_{up} parameter.

D. Comparison Methods

Without extra label information, many unsupervised feature selection methods have been developed to select features by calculating similarities between features and select representative ones with larger similarity values [28]. In order to demonstrate the effectiveness of the proposed WM method on unsupervised feature selection, the WM method is compared with nine unsupervised feature selection techniques.

Next, brief introduction is provided to describe the methods used for the comparison study, and we also list the mentioned SM-based feature selection method.

1) Pearson's Correlation (PC) [29]: The Pearson's correlation is one of most popular techniques to measure the correlations. A larger correlation value indicates more suitability of the feature to represent different processes.

2) Kendall's Tau Coefficient (KTC) [30]: The Kendall's Tau coefficient is used to measure relationships hidden in datasets. It is designed to capture the mapping between different processes represented by features.

3) Spearman's Rho Correlation (SRC) [31]: The Spearman's Rho correlation is a non-parametric measurement, and it is used to rank features of the correlation by assessing the potential monotonic relationship between processes.

4) Distance Correlation (DC) [32]: The distance correlation measures the joint independence of random processes represented by features.

5) Mutual Information (MI) [33]: The mutual information calculates the mutual dependence between two variables, and it can be used to measure the nonlinear dependency between the same type of feature derived from different processes.

6) Maximal Information Correlation (MIC) [34]: The maximal information coefficient measures the linear or non-linear relationship between two processes represented by features. It is capable to characterize them according to properties such as monotonicity.

7) Laplacian Score (LS) [35]: The Laplacian score evaluates the importance of a feature by its locality preserving power, and selects features that respect a local graph structure.

8) Spectrum Graph (SG) [36]: The spectrum graph feature selection method constructs a Laplacian matrix for spectral decomposition, and uses spectral graph theory to measure feature relevance and importance.

9) Strict Monotonicity (SM) [4]: The strict monotonicity characterises the underlying trend of a signal, and evaluates features according to the ability of describing the trend of different processes.

E. Performance Results and Comparison

For the prediction analysis and comparison with key features identified by the feature selection methods, the MDSCI datasets and IMS datasets are used. In the analysis of the proposed WM-based feature evaluation and selection method, as indicated in Section IV-C, the regularisation parameters for variation estimation are selected as λ_{up} with 0.1 and λ_{low} with 10^{-4} for both datasets.

For the MDSCI datasets, the comparisons among different methods are presented in Figure 4 with the number of features increasing from 5 to 40. From Figure 4, it can be seen that the WM method has prediction errors decreasing from 0.069 (5 features) to 0.045 (10 features) for the MAE error. When 5 key features are selected, the proposed WM based feature selection method has the smallest prediction error among all the different methods. As the number of key features selected increases, the proposed WM based feature selection method still has almost the best performance among all different methods. When the number of the key features increases

over 35, other methods such as PC and MI have the similar prediction performance with the WM method. Figure 4 also shows that the SM based feature selection methods has the highest prediction error 0.158 with 5 features, then decreases quickly to 0.08 with 15 features. When the number of features goes to 40, the SM-based method has the similar performance as the proposed method.

For the IMS datasets, it can be seen from Figure 4(d)-(f) that the proposed WM method is outperformed compared with other methods with exceptions of 10 and 35 features. The WM method is also less sensitive as the number of features increases. It can been from these sub-figures that the proposed WM method can identify the most significant features among 160 features with the best performance among all the unsupervised feature selection methods.

It notes that three performance evaluation indices demonstrate some overlap of prediction performance in Figure 4. While the difference between the three indices on the prediction analysis can be seen in Figure 4 (d)-(f), and the differences can also be found in the results of robustness analysis of the methods in the following subsection.

We compared the computation cost for all the unsupervised feature selection methods. The computation procedures of the developed WM method can be found in Algorithm 1 (Alg #1) and Algorithm 2 (Alg #2). The results of computation costs for the two algorithms and other nine methods are presented in Table I. All the methods were randomly ran on the two datasets for 50 times, with Matlab R2020a, Intel(R) Core(TM) i7-4750HQ, 8GB RAM. We calculated the mean value and standard deviation of computation time for each method. The results show that the Algorithm 1 is efficient in estimating the variations, and Algorithm 2 has a low computation cost for the feature evaluation.

TABLE I: Comparison of computation cost for all the methods.

	Computation Time (s)	
	MDSCI datasets	IMS datasets
Alg #1	0.010 ± 0.011	0.010 ± 0.008
Alg #2	0.675 ± 0.163	0.193 ± 0.027
SM	0.564 ± 0.046	0.157 ± 0.021
PC	0.539 ± 0.028	0.167 ± 0.015
KTC	1.532 ± 0.073	0.687 ± 0.015
SRC	0.864 ± 0.052	0.288 ± 0.010
DC	2.165 ± 0.155	1.461 ± 0.057
MI	0.282 ± 0.037	0.139 ± 0.010
MIC	6.245 ± 0.342	16.202 ± 0.138
LS	0.033 ± 0.028	0.058 ± 0.011
SG	0.039 ± 0.031	0.070 ± 0.014

F. Dynamic Bound vs. Fixed Bound

When dealing with subject variations and uncertainties, it is common to place a fixed bound for the process, usually the bound value is selected according to experience. Our proposed WM based method estimates the bound for each feature according to the variations among a family of processes, and the bound value varies in the range $(0, \lambda_{up}]$, it is a dynamical process to estimate the bound for each feature.

In this subsection, we present the results of comparison study with both the fixed bound (denotes as WM_{Fix}) and the dynamic bound by our proposed WM method (denotes as WM_{Dy}). It should be noted that all the computing processes of the two methods are the same. The bound value is set as 0.1 for the WM_{Fix} method, that is, the feature is allowed 10% variation of its value when calculating the monotonicity increasing and decreasing points. We also set the same value of λ_{up} as 0.1 in the proposed WM_{Dy} bound estimation method for comparison.

The prediction errors for the WM_{Fix} and WM_{Dy} methods are shown in Figure 5. Figure 5(a) demonstrates the comparison results with the two feature selection methods on the MDSCI datasets, it can be seen that the WM_{Dy} based method has better prediction performance for all feature dimensions when comparing with the WM_{Fix} method. The similar comparison results can also be seen from the analysis of the IMS datasets in Figure 5(b). Prediction errors obtained from other nine feature selection methods are also kept to make intuitive comparisons. The comparison results demonstrate our proposed WM_{Dy} feature selection method has overall better performance than the WM_{Fix} feature selection method.

G. Robustness Analysis

To test the robustness of the proposed WM method, random noises are added into the original datasets to generate noisy signals. We measure the signal-to-noise ratios (SNR) of the noisy signals, then, the proposed feature evaluation method is used to identify key features from the noisy datasets. Table II presents prediction errors with the identified key features on the MDSCI datasets, and Table III presents prediction errors with identified key features on the IMS datasets. The noisy signals are randomly generated 50 times, then, 160 statistical features are extracted from each of the noisy datasets. The standard SVR model is used for prediction analysis (see Section IV-A for more details).

In the two tables, Set $\#1$ indicates that the number of features is 10. Set $\#2$ and Set $\#3$ indicate the numbers of features are 20 and 30, respectively. Statistical results are presented with calculating the prediction errors by the MAE, MAPE and RMSE measurements of the noisy datasets. As shown in the two tables, the mean value and standard deviation of the prediction errors are calculated from running on the 50 noisy signals of the two datasets. The lowest prediction error for each feature set is highlighted in bold. The two tables show that the proposed WM method outperforms other nine feature selection methods in terms of the MAE, MAPE and RMSE performance indices with only a few exceptions.

To verify the robustness of the proposed WM method, standard statistical tests were used to compare the prediction errors. Prior to the statistical analysis, the Shapiro-Wilk test was applied to check the data normality [37]. The significant value was set as 0.05. It was found that majority of the normality testing results have *p*-values larger then 0.05, which indicates normal distribution of most of the prediction errors.

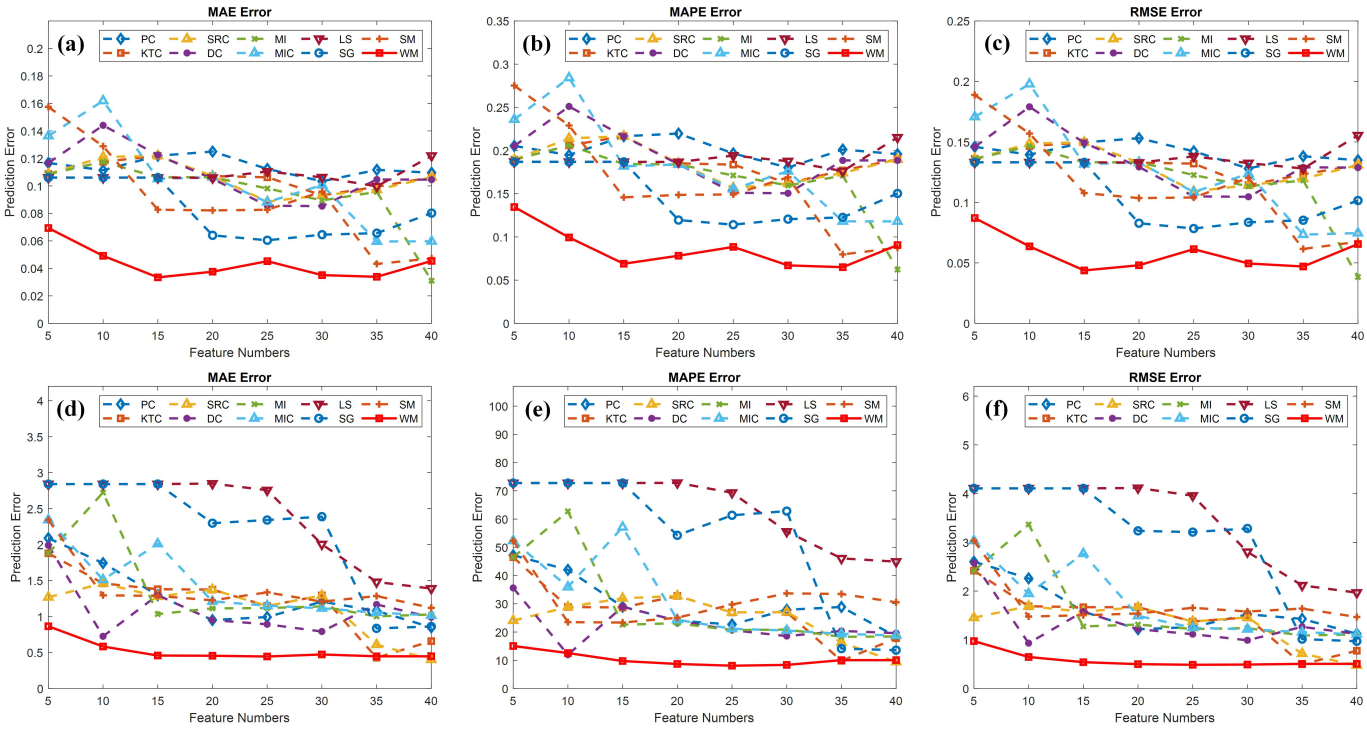

Fig. 4: Prediction error for the datasets. (a) MAE error for the MDSCI datasets, (b) MAPE error for the MDSCI datasets, (c) RMSE error for the MDSCI datasets, (d) MAE error for the IMS datasets, (e) MAPE error for the IMS datasets, (f) RMSE error for the IMS datasets.

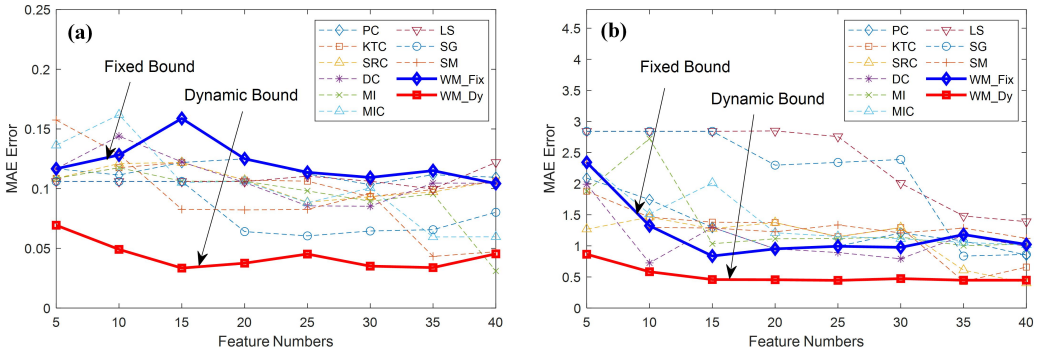

Fig. 5: Comparison study between the fixed bound and dynamic bound. (a) MAE error for the MDSCI datasets, (b) MAE error for the IMS datasets.

Then, depending on whether the data was normally distributed, the paired *t*-test or Mann-Whitney test was performed to compare the prediction errors between the WM method and each of the other nine methods, and the *p*-values were smaller than 0.05 with only a few exceptions, which indicated the differences between the developed WM method and other unsupervised feature selection methods.

V. Discussion

The monotonicity has been observed in many real-world applications, e.g., mechanical engineering, electrical engineering, and medical science. For example, many engineering systems would show performance deterioration over time, indicating a trend in the data measurements. This paper utilises this trend in a population to select the features that contribute to this trend, and the results demonstrated the robustness of the proposed method when the weak population trends exist.

Tremendous efforts have been made to tackle the challenging task of unsupervised feature selection [13], [38], [39], [40], [41], [42], [43]. Generally, these methods use projection [13] or representation [41] techniques to learn geometrical structures of datasets for unsupervised feature selection. With imposing some regularisation terms, i.e., $\ell_{2,1}$-norm [13], [38], [40], Laplacian regularisation [40], [42], these methods have shown excellent performance on minimising data redundancy and selecting discriminative features for clustering and classification tasks.

The developed WM method defines a new cost, which is

TABLE II: Prediction errors for the MDSCI datasets (3dB)

	MAE Error			MAPE Error			RMSE Error		
Name	Set #1	Set #2	Set #3	Set #1	Set #2	Set #3	Set #1	Set #2	Set #3
PC	0.089 ± 0.014	0.075 ± 0.008	0.070 ± 0.011	0.156 ± 0.024	0.132 ± 0.014	0.124 ± 0.017	0.116 ± 0.018	0.098 ± 0.010	0.090 ± 0.013
KTC	0.073 ± 0.011	0.068 ± 0.009	0.064 ± 0.008	0.128 ± 0.020	0.120 ± 0.014	0.115 ± 0.013	0.095 ± 0.014	0.088 ± 0.011	0.082 ± 0.010
SRC	0.071 ± 0.012	0.067 ± 0.010	0.065 ± 0.008	0.125 ± 0.020	0.119 ± 0.017	0.117 ± 0.012	0.092 ± 0.015	0.085 ± 0.013	0.083 ± 0.010
DC	0.067 ± 0.011	0.065 ± 0.009	0.062 ± 0.008	0.119 ± 0.018	0.116 ± 0.014	0.113 ± 0.013	0.085 ± 0.013	0.082 ± 0.011	0.079 ± 0.010
MI	0.077 ± 0.008	0.073 ± 0.009	0.068 ± 0.009	0.136 ± 0.015	0.129 ± 0.015	0.122 ± 0.014	0.101 ± 0.012	0.094 ± 0.012	0.088 ± 0.011
MIC	0.069 ± 0.011	0.065 ± 0.010	0.061 ± 0.010	0.123 ± 0.018	0.117 ± 0.016	0.111 ± 0.017	0.088 ± 0.013	0.083 ± 0.012	0.078 ± 0.012
LS	0.090 ± 0.002	0.096 ± 0.016	0.093 ± 0.014	0.153 ± 0.005	0.163 ± 0.028	0.158 ± 0.026	0.118 ± 0.003	0.126 ± 0.021	0.122 ± 0.018
SG	0.067 ± 0.005	0.055 ± 0.013	0.053 ± 0.013	0.114 ± 0.008	0.098 ± 0.022	0.096 ± 0.021	0.088 ± 0.006	0.071 ± 0.017	0.069 ± 0.016
SM	0.074 ± 0.007	0.069 ± 0.010	0.065 ± 0.009	0.131 ± 0.013	0.123 ± 0.016	0.116 ± 0.015	0.097 ± 0.010	0.089 ± 0.012	0.083 ± 0.011
WM	$\mathbf{0.060 \pm 0.009}$	$\mathbf{0.051 \pm 0.010}$	$\mathbf{0.049 \pm 0.009}$	$\mathbf{0.109 \pm 0.016}$	$\mathbf{0.094 \pm 0.016}$	$\mathbf{0.091 \pm 0.015}$	$\mathbf{0.076 \pm 0.012}$	$\mathbf{0.064 \pm 0.012}$	$\mathbf{0.061 \pm 0.011}$

TABLE III: Prediction errors for the IMS datasets (3dB)

	MAE Error			MAPE Error			RMSE Error		
Name	Set #1	Set #2	Set #3	Set #1	Set #2	Set #3	Set #1	Set #2	Set #3
PC	1.783 ± 0.209	1.264 ± 0.285	0.931 ± 0.262	43.668 ± 6.726	29.920 ± 9.145	21.029 ± 7.038	2.366 ± 0.314	1.664 ± 0.434	1.154 ± 0.357
KTC	1.285 ± 0.373	0.908 ± 0.194	$\mathbf{0.790 \pm 0.178}$	27.215 ± 10.320	19.065 ± 5.069	17.774 ± 5.491	1.582 ± 0.541	1.099 ± 0.247	0.991 ± 0.249
SRC	1.160 ± 0.277	0.921 ± 0.209	0.820 ± 0.190	23.039 ± 6.963	19.330 ± 5.274	18.130 ± 5.377	1.377 ± 0.367	1.113 ± 0.261	1.020 ± 0.249
DC	1.805 ± 0.821	0.930 ± 0.187	0.791 ± 0.155	36.235 ± 19.684	19.831 ± 4.996	17.588 ± 4.687	2.109 ± 1.034	1.126 ± 0.249	$\mathbf{0.988 \pm 0.217}$
MI	2.264 ± 0.312	1.609 ± 0.366	0.959 ± 0.186	54.589 ± 7.682	36.606 ± 10.704	20.510 ± 5.077	2.967 ± 0.404	2.020 ± 0.523	1.154 ± 0.239
MIC	1.924 ± 0.284	1.722 ± 0.298	1.227 ± 0.325	47.488 ± 6.553	43.133 ± 7.433	29.590 ± 9.635	2.508 ± 0.347	2.280 ± 0.377	1.559 ± 0.457
LS	3.646 ± 0.040	3.581 ± 0.046	2.788 ± 0.182	92.945 ± 1.028	91.121 ± 1.158	69.229 ± 5.572	5.079 ± 0.059	4.983 ± 0.066	3.852 ± 0.280
SG	3.646 ± 0.040	2.944 ± 0.092	3.129 ± 0.074	92.945 ± 1.028	74.542 ± 2.529	75.770 ± 1.575	5.079 ± 0.059	4.066 ± 0.139	4.250 ± 0.092
SM	2.441 ± 0.469	1.400 ± 0.306	1.124 ± 0.362	56.027 ± 11.375	33.684 ± 8.780	26.962 ± 9.840	3.077 ± 0.604	1.805 ± 0.436	1.440 ± 0.489
WM	$\mathbf{0.890 \pm 0.215}$	$\mathbf{0.795 \pm 0.212}$	0.821 ± 0.245	$\mathbf{16.805 \pm 4.487}$	$\mathbf{15.826 \pm 4.588}$	$\mathbf{16.735 \pm 4.834}$	$\mathbf{1.036 \pm 0.244}$	$\mathbf{0.959 \pm 0.237}$	1.003 ± 0.269

different from the costs that were used in these research [13], [38], [39], [40], [41], [42], [43]. The WM method aims to identify representative features that can capture underlying trend for regression analysis, rather than select discriminative features for classification or clustering as seen in these research. The method developed in this paper uses the concept of WM to learn local property between adjacent points in a data sequence, which is used as prior knowledge for the modelling. Then, the method uses similarity between different processes to represent the trendability of a population. With integrating the two properties, the developed method was shown superiority in identifying important features for regression analysis of sequential data when noises and uncertainties exist.

It is noted that deep learning methods have also been used for trend analysis with excellent performance [10], [11], while deep learning methods generally require a large number of data samples to train the model and guarantee the performance. The training process thus requires high computation cost and needs knowledge for hyperparameters tuning. On the contrary, the developed WM method in this paper is demonstrated with an efficient and transparent computation procedure.

The developed WM method has limitations. The method assumes the existence of a common trend in a family of processes. Therefore, the method is suitable to investigate processes with underlying trends, and may be not generalised to other applications when such trends do not exist. Our future work will generalise this concept to cases when trends exist in sub-groups of the population.

VI. Conclusion

This paper introduced the concept of weak monotonicity (WM) and developed a new index to characterise a large family of processes shared with common trends in the presence of measurement noises and population variations. Such a concept can be used to identify key features that contribute to the common trend of a family of processes. With the help of similarity measure among populations, a novel suitability indicator was proposed as a cost function for unsupervised feature evaluation and selection. The proposed method was compared with other nine widely used unsupervised feature selection methods on well-known datasets. The statistical results verified the effectiveness and robustness of the proposed method. Future work will focus on theoretical analysis of the proposed method with a more rigorous estimation of the upper bound of noises and uncertainties, and applications in more general conditions.

References

[1] G. Morales-Espejel, P. Rycerz, and A. Kadiric, "Prediction of micropitting damage in gear teeth contacts considering the concurrent effects of surface fatigue and mild wear," *Wear*, vol. 398, pp. 99–115, 2018.

[2] R. M. Nejad, M. Shariati, and K. Farhangdoost, "Prediction of fatigue crack propagation and fractography of rail steel," *Theor. Appl. Fract. Mech.*, vol. 101, pp. 320–331, 2019.

[3] Z.-H. Pang, G.-P. Liu, and D. Zhou, "Design and performance analysis of incremental networked predictive control systems," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1400–1410, 2015.

[4] P. Baraldi, G. Bonfanti, and E. Zio, "Differential evolution-based multi-objective optimization for the definition of a health indicator for fault diagnostics and prognostics," *Mech. Syst. Sig. Process.*, vol. 102, pp. 382–400, 2018.
[5] M. Komeili, W. Louis, N. Armanfard, and D. Hatzinakos, "Feature selection for nonstationary data: Application to human recognition using medical biometrics," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1446–1459, 2018.
[6] P. Kundu and S. Mitra, "Feature selection through message passing.," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4356–4366, 2017.
[7] Q. Cheng, H. Zhou, and J. Cheng, "The Fisher-Markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1217–1233, 2010.
[8] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.
[9] L. Lu, J. Yan, and C. W. de Silva, "Dominant feature selection for the fault diagnosis of rotary machines using modified genetic algorithm and empirical mode decomposition," *J. Sound Vib.*, vol. 344, pp. 464–483, 2015.
[10] E. Q. Wu, P. Xiong, Z.-R. Tang, G.-J. Li, A. Song, and L.-M. Zhu, "Detecting dynamic behavior of brain fatigue through 3-d-cnn-lstm," *IEEE Trans. Syst. Man Cybern. Syst.*, pp. 1–11, 2021.
[11] S. Z. Tajalli, A. Kavousi-Fard, M. Mardaneh, A. Khosravi, and R. Razavi-Far, "Uncertainty-aware management of smart grids using cloud-based lstm-prediction interval," *IEEE Trans. Cybern.*, pp. 1–14, 2021.
[12] Y. Liu, K. Liu, J. Yang, and Y. Yao, "Spatial-neighborhood manifold learning for nondestructive testing of defects in polymer composites," *IEEE Trans. Ind. Inform.*, vol. 16, no. 7, pp. 4639–4649, 2020.
[13] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.
[14] J. Li and H. Liu, "Challenges of feature selection for big data analytics," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 9–15, 2017.
[15] A. Mironchenko, I. Karafyllis, and M. Krstic, "Monotonicity methods for input-to-state stability of nonlinear parabolic PDEs with boundary disturbances," *SIAM Journal on Control and Optimization*, vol. 57, no. 1, pp. 510–532, 2019.
[16] E. N. Sadjadi, "On the monotonicity of smooth fuzzy systems," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 12, pp. 3947–3952, 2021.
[17] J. Alexander Jr, R. A. Edwards, M. Brodsky, L. Manca, R. Grugni, A. Savoldelli, G. Bonfanti, B. Emir, E. Whalen, S. Watt, *et al.*, "Using time series analysis approaches for improved prediction of pain outcomes in subgroups of patients with painful diabetic peripheral neuropathy," *PLoS One*, vol. 13, no. 12, p. e0207120, 2018.
[18] L. Lu, Y. Tan, M. Klaic, M. P. Galea, F. Khan, A. Oliver, I. Mareels, D. Oetomo, and E. Zhao, "Evaluating rehabilitation progress using motion features identified by machine learning," *IEEE. Trans. Biomed. Eng.*, vol. 68, no. 4, pp. 1417–1428, 2021.
[19] A. Delgado-Bonal and A. Marshak, "Approximate entropy and sample entropy: A comprehensive tutorial," *Entropy*, vol. 21, no. 6, p. 541, 2019.
[20] A. Agogino and K. Goebel, "Mill data set. BEST lab, UC Berkeley. NASA Ames Prognostics Data Repository," 2007.
[21] J. Lee, H. Qiu, G. Yu, and J. Lin, "Bearing data set, NASA Ames Prognostics Data Repository," 2007.
[22] Z.-R. Feng, Q. Zhou, J. Zhang, P. Jiang, and X.-W. Yang, "A target guided subband filter for acoustic event detection in noisy environments using wavelet packets," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 2, pp. 361–372, 2014.
[23] Y. Lei, M. J. Zuo, Z. He, and Y. Zi, "A multidimensional hybrid intelligent method for gear fault diagnosis," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 1419–1430, 2010.
[24] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *J. Mach. Learn. Res.*, vol. 6, no. Dec, pp. 1889–1918, 2005.
[25] Y. I. Lee and B. Kouvaritakis, "Robust receding horizon predictive control for systems with uncertain dynamics and input saturation," *Automatica*, vol. 36, no. 10, pp. 1497–1504, 2000.
[26] R. K. Singleton, E. G. Strangas, and S. Aviyente, "Extended Kalman filtering for remaining-useful-life estimation of bearings," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1781–1790, 2014.
[27] X. Chen, G. Yuan, W. Wang, F. Nie, X. Chang, and J. Z. Huang, "Local adaptive projection framework for feature selection of labeled and unlabeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6362–6373, 2018.
[28] R. Zhang, F. Nie, Y. Wang, and X. Li, "Unsupervised feature selection via adaptive multimeasure fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, 2019.
[29] L. Bravi, V. Piccialli, and M. Sciandrone, "An optimization-based method for feature ranking in nonlinear regression problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 4, pp. 1005–1010, 2016.
[30] J. Sinsomboonthong, "Robust estimators for the correlation measure to resist outliers in data," *J. Math. Fundam. Sci.*, vol. 48, no. 3, pp. 263–275, 2016.
[31] M. Pedersen, A. Omidvarnia, A. Zalesky, and G. D. Jackson, "On the relationship between instantaneous phase synchrony and correlation-based sliding windows for time-resolved fMRI connectivity analysis," *Neuroimage*, vol. 181, pp. 85–94, 2018.
[32] D. Edelmann, K. Fokianos, and M. Pitsillou, "An updated literature review of distance correlation and its applications to time series," *Int. Stat. Rev.*, vol. 87, no. 2, pp. 237–262, 2019.
[33] M. Han, W. Ren, M. Xu, and T. Qiu, "Nonuniform state space reconstruction for multivariate chaotic time series.," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1885–1895, 2019.
[34] Z. Li and A. G. Bors, "Selection of robust and relevant features for 3-D steganalysis," *IEEE Trans. Cybern.*, 2018.
[35] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *Advances in neural information processing systems*, vol. 18, 2005.
[36] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th international conference on Machine learning*, pp. 1151–1157, 2007.
[37] L. Lu, M. Robinson, Y. Tan, K. Goonewardena, X. Guo, I. Mareels, and D. Oetomo, "Effective assessments of a short-duration poor posture on upper limb muscle fatigue before physical exercise," *Frontiers in Physiology*, vol. 11, p. 1201, 2020.
[38] X. Zhu, S. Zhang, R. Hu, Y. Zhu, and j. song, "Local and global structure preservation for robust unsupervised spectral feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 3, pp. 517–529, 2018.
[39] X. Li, H. Zhang, R. Zhang, Y. Liu, and F. Nie, "Generalized uncorrelated regression with adaptive graph for unsupervised feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1587–1595, 2018.
[40] C. Tang, X. Zheng, X. Liu, W. Zhang, J. Zhang, J. Xiong, and L. Wang, "Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2021.
[41] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognition*, vol. 48, no. 2, pp. 438–446, 2015.
[42] C. Tang, X. Liu, X. Zhu, J. Xiong, M. Li, J. Xia, X. Wang, and L. Wang, "Feature selective projection with low-rank embedding and dual laplacian regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1747–1760, 2020.
[43] C. Tang, M. Bian, X. Liu, M. Li, H. Zhou, P. Wang, and H. Yin, "Unsupervised feature selection via latent representation learning and manifold regularization," *Neural Networks*, vol. 117, pp. 163–178, 2019.

Lei Lu received his Ph.D. degree in Mechatronics Engineering from Harbin Institute of Technology, in 2016. He had research experience at the Department of Mechanical Engineering, The University of British Columbia, Canada, from 2013 to 2015; and the Faculty of Engineering and Information Technology, The University of Melbourne, Australia, from 2017 to 2020. He is currently an EPSRC supported postdoctoral researcher at the Department of Engineering Science, University of Oxford, UK.

His research interests include signal processing, statistical machine learning, and deep learning with applications in decision-making for engineering systems and intelligent health condition monitoring. He received some important scientific awards including the IET J. A. Lodge Award in healthcare technologies, in 2021; and he was a PI in several research projects including the National Natural Science Foundation of China (NSFC).

Ying Tan is a Professor in the Department of Mechanical Engineering at The University of Melbourne, Australia. She received her Bachelor's degree from Tianjin University, China, in 1995, and her PhD from the National University of Singapore in 2002. She joined McMaster University in 2002 as a postdoctoral fellow in the Department of Chemical Engineering. Since 2004, she has been with the University of Melbourne. She was awarded an Australian Postdoctoral Fellow (2006-2008) and a Future Fellow (2009-2013) by the Australian Research Council. Her research interests are in intelligent systems, nonlinear systems, real-time optimization, sampled-data systems, rehabilitation robotic systems, human motor learning, and model-guided machine learning.

Denny Oetomo received the B.Eng. degree (Hons.) from the Australian National University, Canberra, ACT, Australia, in 1997, and the Ph.D. from the National University of Singapore, Singapore, in 2004.

In 2008, he joined the Department of Mechanical Engineering, The University of Melbourne, Melbourne, VIC, Australia, where he is currently a Professor. His research interests are in the area of robot dynamics and manipulation with a recent emphasis on the interaction dynamics between human and robots, as well as on the clinical applications of these capabilities, such as in the areas of rehabilitation robotics, assistive robotics, and neuroprosthetics.

Iven Mareels is the Director of the Centre for Applied Research, IBM A/NZ and honorary Professor of Electrical and Electronic Engineering at the University of Melbourne. He is a leading expert in the area of large scale systems, adaptive control and extremum seeking (forms of AI). He is a Fellow of the IEEE (USA); Fellow of IFAC (Austria); Fellow and Vice-President of the Australian Academy of Technology and Engineering (Australia) and a (Foreign) Fellow of the Flemish Royal Belgian Academy of Sciences and Humanities (Belgium). He has co-authored 5 books, and in excess of 150 journal papers and book chapters, and more than 300 conference papers. He has co-invented a suite of international patents addressing the management of large scale, gravity fed, irrigation systems.

David A. Clifton is a Professor of Clinical Machine Learning in the Department of Engineering Science of the University of Oxford, and OCC Fellow in AI & Machine Learning at Reuben College, Oxford. He is a Fellow of the Alan Turing Institute, Research Fellow of the Royal Academy of Engineering, Visiting Chair in AI for Healthcare at the University of Manchester, and a Fellow of Fudan University, China.

His research focuses on the development of machine learning for tracking the health of complex systems. His previous research resulted in patented systems for jet-engine health monitoring, used with the engines of the Airbus A380, the Boeing 787 "Dreamliner", and the Eurofighter Typhoon. Since 2008, he has focused mostly on the development of AI-based methods for healthcare. Patents arising from this collaborative research have been commercialised via university spin-out companies OBS Medical, Oxehealth, and Sensyne Health, in addition to collaboration with multinational industrial bodies. He holds a Grand Challenge award from the UK Engineering and Physical Sciences Research Council, which is an EPSRC Fellowship that provides long-term strategic support for "future leaders in healthcare". His research has been awarded over 35 academic prizes; in 2018, he was joint winner of the inaugural "Vice-Chancellor's Innovation Prize", which identifies the best interdisciplinary research across the entirety of the University of Oxford.